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 - 25
Sampling for Proportions and Percentages
Basic Concepts

Hello friends, welcome to the course Essentials of Data Science with R Software to where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. In this lecture, we are going to continue with the module on Sampling Theory with R Software and we are going to begin with the new chapter on the Sampling Theory for Proportions and Percentage.

So, from this lecture, we are going to consider the simple random sampling for a qualitative variable. When we studied the simple random sampling earlier, then we had assumed that the variable under study is quantitative. Quantitative means, the observations can be obtained in a numerical way. Numerical way means, for example, if I say suppose I have a variable height; so height can be say 152 centimeter, 155 centimeter, 155.5 centimeter and so on.

But, now there can be many situations in real life, where the variables are qualitative, they are dichotomous. The variables are defined in such a way, such that they takes only two possible values and those values are actually the values of an indicator variable. For example, if I conduct a survey to judge the taste of a coffee.

So, the person will sip the coffee and the response will be something like good or bad. This cannot be quantified; good is good and bad is bad. Similarly, if I say the gender of the students, that can be male or female. So, under this type of situations the first problem is this that, how to conduct the simple random sampling, how to obtain a representative sample. And after that, the observation are not actually a numerical value; but these are some indicator variables like good, bad, yes, no, etc.

So, how to convert them into a numerical way, such that statistics can be used over that? One limitation in statistics is that, the statistical tool work only on the numbers,

quantitative variables. So, first we have to understand how to draw a simple random sampling with and without replacement.

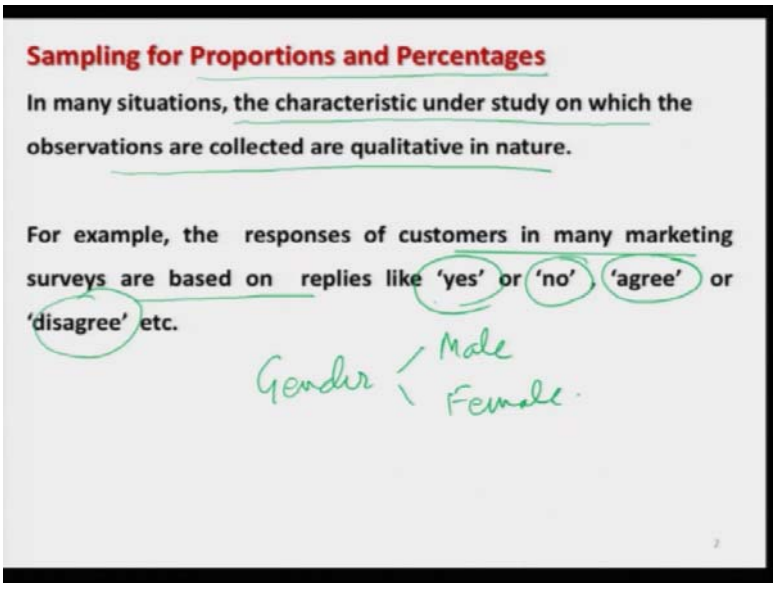
Then once you obtain the value of a dichotomous variable, which are essentially indicator variable, you need to convert them into some numericals and then you would like to apply the statistical tools over them.

So, in this lecture, I am going to assume that you know completely about the simple random sampling because I will be simply extending the concept of simple random sampling for quantitative variable to qualitative variable. I have two options; either I try to teach it right from the beginning, from the scratch as I did in the case of simple random sampling.

Or more intelligent option is that, since we already have done the simple random sampling, so why not to make a one to one connection between the simple random sampling of qualitative variable and quantitative variable. Once I am able to establish the connection between the two, then I can use the tools of simple random sampling that we already have developed directly for the cases when the x is dichotomous variable or an indicator variable or a qualitative variable.

So, let us begin this lecture with these concepts, ok.

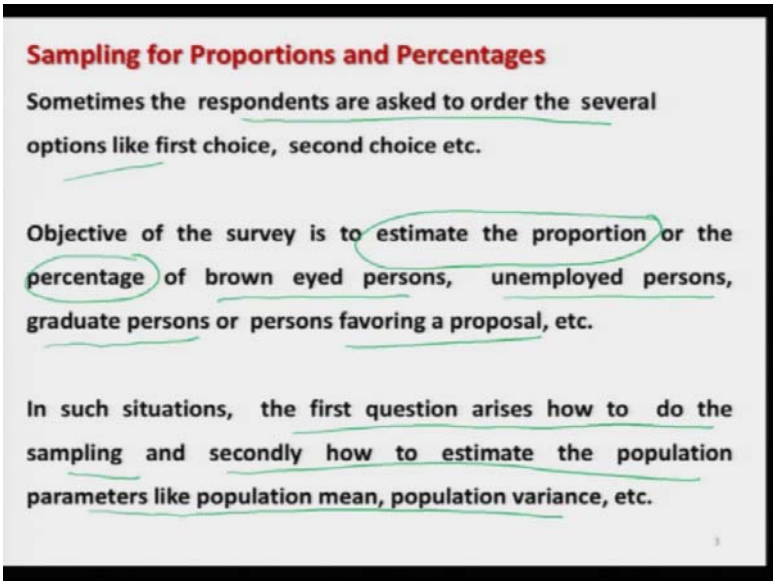
(Refer Slide Time: 04:44)



So, now we are going to consider the sampling for proportion and percentages. What does this mean? I will try to explain you in the forthcoming slides. In many situations, the characteristic under study on which the observations are collect are qualitative in nature.

For example, the responses of a customers in many marketing surveys can be like yes or no or they can be like agree or disagree; simply for a gender there can be two categories, male and female.

(Refer Slide Time: 05:25)



So, the question is this, what to do in this situations? The next scenario can be where the respondents are asked to order the several options; if there are three choices of a drink say coffee, tea or some juice. So, if you ask someone, then usually the person will like all the things; but then we ask them can you please order it.

What is your first choice, what is your second choice, and what is your third choice? So, the answer can be from one respondent like as tea, coffee and juice; whereas another respondent may say juice, coffee and tea and so on. So, this is what I mean when I say that the respondents are asked to order the options.

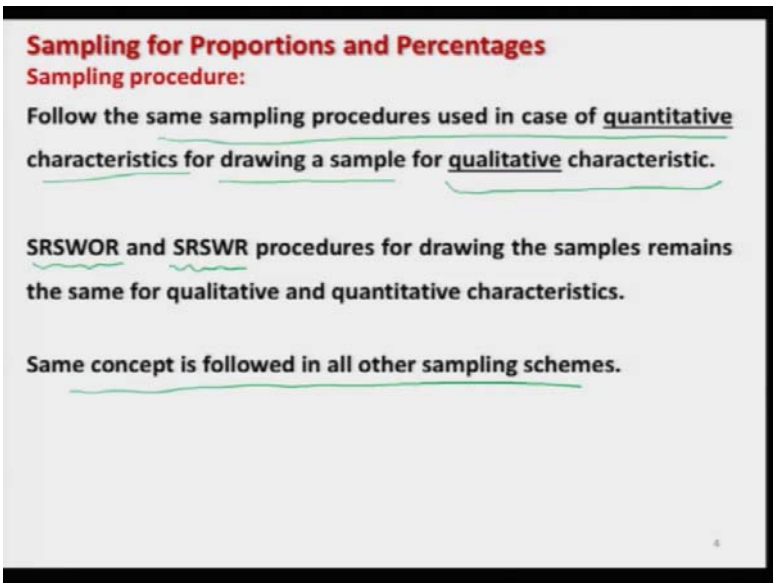
Similarly, there can be many situations that we want to have a survey of people who have got say brown eyes; the survey on the status of the person that the person is employed or say unemployed; the education of the person that the persons can be divided

into say these two groups, whether the person is graduate or a non-graduate. Or if there is some proposal, there can be two groups; one group where the people are favoring a proposal, and in second group people who are not favoring a proposal.

So, in such situation we are more interested in estimating the proportion or the percentage of people or the observation in one group and in the second group. So, in such condition, in such situation, the first question arises is how to do the sampling and the next condition is this, how to estimate the population parameters.

Population parameters of interest are like population mean and population variance that we discussed in the case of simple random sampling for qualitative variable.

(Refer Slide Time: 07:45)

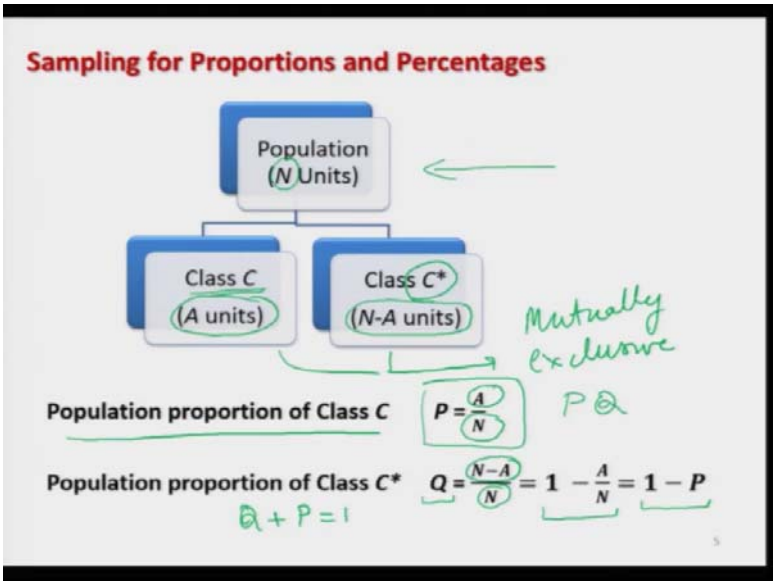


So, under this situation, the answer of the first question that; how to do the sampling is very straight forward. We are going to use the same sampling procedure what we have used in the case of quantitative characteristics for drawing a sample. That will be the same process that we will follow for the qualitative variable also, qualitative characteristic also.

So, up to now we have done two sampling procedures; simple random sampling without replacement, and simple random sampling with replacement for drawing the sample. So, the procedure that how are we going to draw the samples, this will remain exactly the same in case of qualitative characteristic, just like the in the case of quantitative

characteristic. And whatever concept we have developed in case of SRSWOR, SRSWR they will be followed exactly in the same way with some modification, ok.

(Refer Slide Time: 08:54)



Now, let us try to first understand how is this population and how are we going to draw the sample and then how are we going to do the mathematical manipulations? Mathematical manipulations means, how are we going to handle the observations, so that they can be used over a statistical tool like as mean, variance etc.

So, suppose we have here a population and this population has got N of units and this population is divided into two groups or we call it that population is divided into two classes; those classes are denoted by C and C*. For example, the population can be of the people who want to give their opinion on a proposal. So, there will be two classes; one class which says that they are in the favor of the proposal and the second case where the persons are not in the favor of the proposal.

So, we assume that out of N, suppose there are A number of units which belong to class C; that means A persons are there who are favoring the proposal and there are, obviously then the remaining units which are N - A they will belong to the second class.

So, we are assuming that these two classes are mutually exclusive; this means there is no unit which can belong to both the classes. So, the condition is this, unit can belong only to one of the two classes, either C or C*. So, in such cases, if you try to see here if you

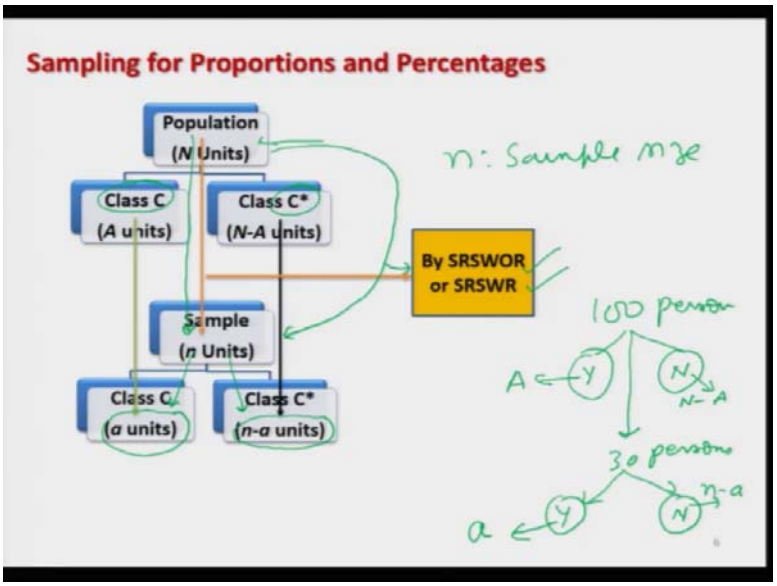
want to estimate the population proportion which is our basic objective whenever we are trying to deal with the qualitative characteristic.

So, the population proportion in this class C that is going to be the number of units in the class C, which is here A divided by the total number of units in the population that is N. So, the population proportion is defined and indicated here by P. So, P is equal to A/ N.

Then obviously, the proportion of the people or the proportion of the units in the class C*, which is the complementary class of C, this will become the remaining units which is N - A divided by total number of units in the class C*. And this proportion is denoted by Q Q. So, for the population proportion, we are using the symbols letters P and letters Q.

So, now this Q can be expressed as 1 - A/N which is same as 1 - P. So, you can see here that, now we have a relationship that Q + P it should be equal to 1. Now, what we try to do? We have to draw a sample.

(Refer Slide Time: 12:09)



So, now you can see here; we have a population here, which is divided into two classes class C and C* and from there we try to draw a sample. So, suppose we want to draw a sample of size n. So, n is my here sample size. So, what I try to draw?

I try to draw a n number of units from this population. Now, when I have drawn a n number of units from this population; so obviously in the sample, there will be two types of units, one set of unit which belong to class C and say another set of unit which belongs to class C*.

So, suppose we assume that the number of sampling units in the sample which belong to class C are a. So, there are a number of units in the sample, which are belonging to class C. And obviously, the remaining unit which are n number - a number; this number of units will belong to the complementary class C*.

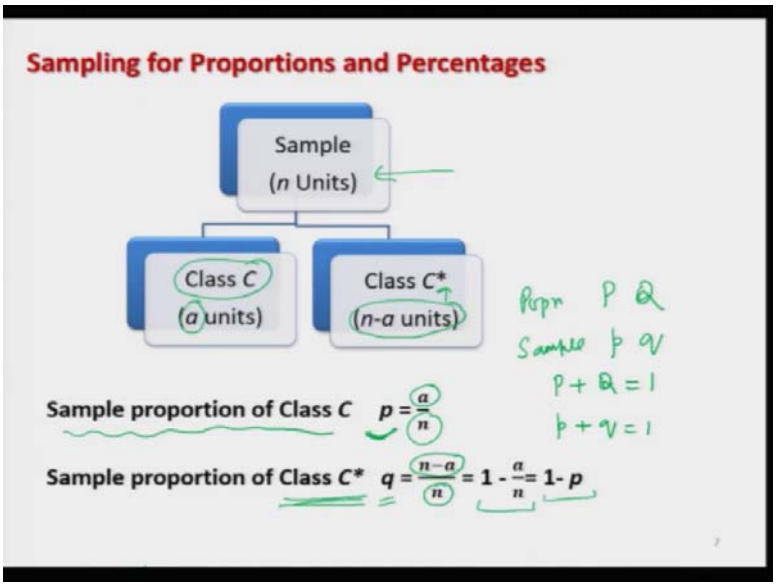
So, what is really happening? Suppose we have a population of people who are going to opine on some proposal. So, in the population, suppose there are suppose 100 persons; obviously in this group, there will be two persons who are saying yes to the proposal and who are saying no to the proposal.

Now, suppose I try to draw here a sample of size 30 persons. Now, in this sample also there will be two types of person, who are saying yes and who are saying no. So, the number of persons who are saying yes in the population this is here A and the persons and the number of persons who are saying no in the population, this is N - A. The number of person who are saying yes in the sample this is a, and the number of person who are saying no in the sample this is n - a.

So, this is my setup. Now, when I am trying to draw the sample from the population of size N and suppose we draw a sample of size n; then the sample can be selected by simple random sampling. And that will be your choice that will be the condition or the experiment which will help you in making call that, whether you want to follow a simple random sampling with replacement or simple random sampling without replacement. So, both these things are possible.

So, now as the procedure of selection of a unit is concerned, this is exactly the same as we discussed in the case of simple random sampling. So, I am not going to repeat it; but you have to recall or else if you cannot recall, please try to go through with the earlier lectures, ok.

(Refer Slide Time: 15:33)



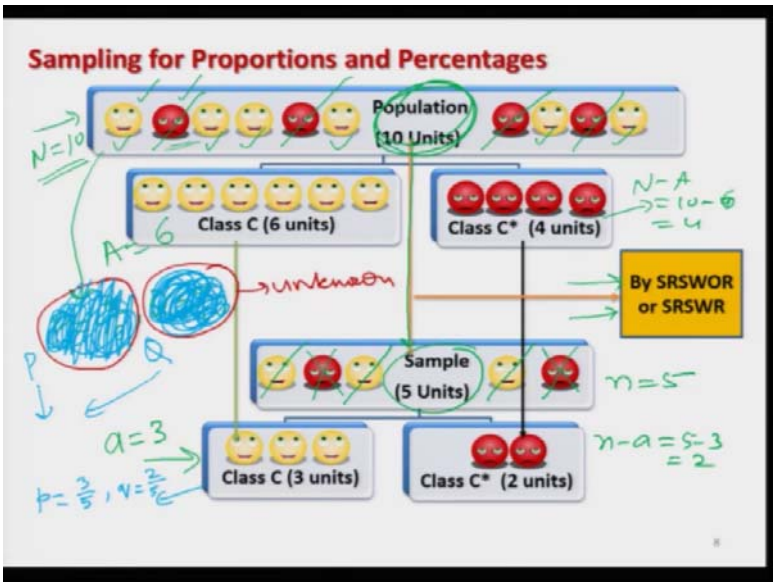
Now, I have got here a sample of n number of units, in which there are two types of samples; one sample consisting of a number of units from the class C and the remaining units n - a, which belongs to class C*.

So, now, exactly on the same way as we have is found the population proportion, we can also find the sample proportion; that means the number of persons belonging to class C and number of persons belonging to class C*, they will constitute the sample proportion as follows.

The sample proportion for the class C will be indicated by p and this is the total number of units, which is a divided by the total number of unit which is n. And the sample proportion of the units belonging to class C* that will be denoted by q and obviously, this is the total number of units in the class C* which is n - a divided by the total number of units in the sample which is n. So, this can also be written as 1 - a/n, which is the same as 1 - p.

So, you can see here we are denoting the population proportion by letters P and Q and the sample proportions are denoted by p and q, right. And there was a relationship for the population P + Q equal to 1 and similarly we have a condition p + q is equal to 1 from the sample. So, this is our basic setup.

(Refer Slide Time: 17:35)



Now, let me try to give you one more example to make you understand better. So, suppose we have here a population of size N is equal to 10. So, what I have done? Suppose there was an experiment to taste the coffee. So, N is equal to 10 number of people who were asked to drink the coffee and after that their reactions were recorded.

So, you can see here, I have used here 2 symbols, 2 smileys, 2 faces; one is here in yellow and another here is in yellow and another here in red. So yellow faces are indicating that ok, the taste was good and the person is happy after drinking the tea; whereas the red faces which are indicating that the person is not happy after drinking the tea.

So, you can see here that, I have here six persons 1, 2, 3, 4, 5 and here 6, which are happy with the taste of the coffee and there are four- 1, 2, 3 and here 4 persons who could not like the taste of the coffee. So, I can divide them into two possible classes; one is the class of people who are happy.

So, there are here A is equal to 6 and second class where the people are not happy with the taste of the coffee; this consists of class C* with N - A which is equal to 10 - 6 is equal to 4 number of units. Now, I try to draw a sample from this population and I try to draw say 5 number of units.

So, my n is equal to here 5. And in order to draw the samples, I have two options; I can use simple random sampling without replacement or simple random sampling with replacement. Now, in the sample also there will be 2 types of people, one who are happy and one who are not happy.

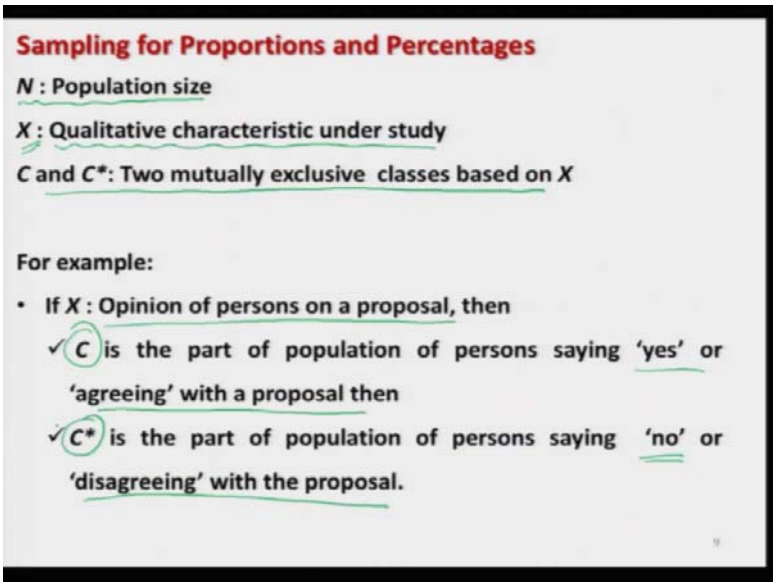
So, you can see here in the sample there are 1, 2 and 3 persons who are happy, who belong to the class C and there are 2 persons who are not happy, so they belong to the class C*. So, here a is equal to here 3 and n - a will be equal to 5 - 3 is equal to 2, right.

So, if you come back to population; in the case of population, we have the population proportion of the people who are happy is 6/10 and the people who are not happy with the taste of the coffee this is Q, this is 4/10. So, now, the problem is this that, these two values of P and Q which I am marking here in color red; they are unknown to us, because they belong to population.

Now, what I try to do? I try to compute the sample proportion from the sample say p, number of people who are happy with the taste of the coffee which is 3/5 and the number of persons who are not happy with the taste of the coffee which is 2/5.

So, now my question is this here that, this P and Q that is not known to us; you can see here I have hided it with the blue pen. So, you do not know, you simply know that these proportions are P and Q and you want to estimate the value of P on the basis of the sample. So, the question is this, how are you going to estimate this population proportion P or equivalently the Q?

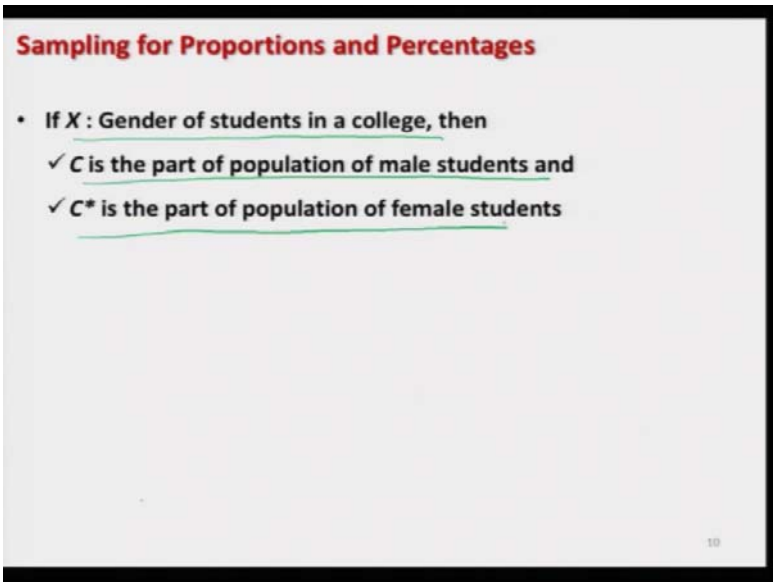
(Refer Slide Time: 21:54)



So, now, N is the population size from where we are going to draw the sample and I am introducing here one symbol X; so this is going to indicate the qualitative characteristic understudy. Remember I am saying this is qualitative variable; qualitative variables means, it is something like yes or no; or good or bad.

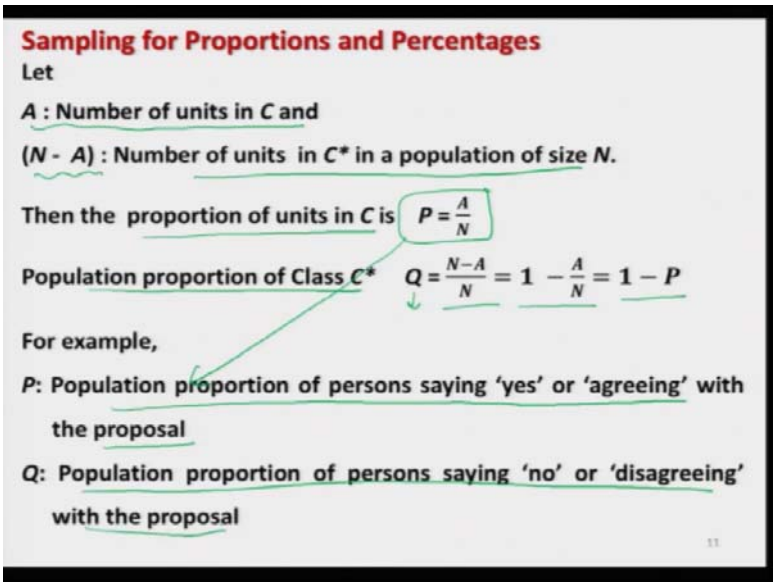
And these two characteristics have been divided into two mutually exclusive classes which are based on X, the qualitative characteristic. For example, X can be a variable like opinion of persons on a proposal. So, there will be some person who will be saying yes or they will agree with the proposal; then this is my here class C and the remaining people who are saying no or they are not agreeing or they are disagreeing with the proposal that is inside the class C*, ok.

(Refer Slide Time: 22:57)



So, similarly if I say X is the gender of the students in the classes; so C is the part or the population of male student and C* will be the part of the population corresponding to the female students.

(Refer Slide Time: 23:12)

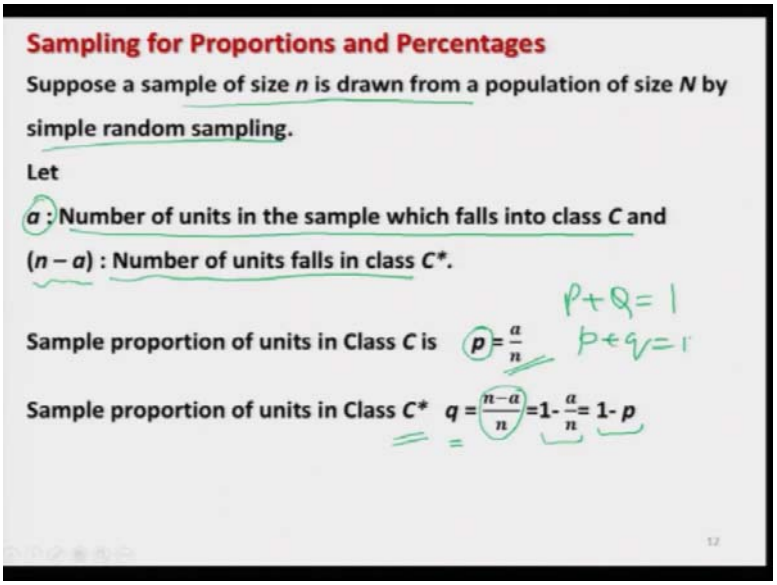


So, similarly there can be many more examples. So, as we have discussed in the earlier diagram, A is the number of units in class C and then obviously, N - A is the number of units in class C*. And then the proportion of the units in class C is denoted by P is equal

to A/N and the population proportion of the units in the class A star is denoted by Q as we have already discussed. This is N - A/A and which is same as 1 - A/N and this is same as 1 - P, right.

So, this P is something like the population proportion of persons who are saying yes or they are agreeing with the proposal as in the earlier example and Q is the complement of this. So, Q is the population proportion of person saying no or they are not agreeing with the proposal.

(Refer Slide Time: 24:13)

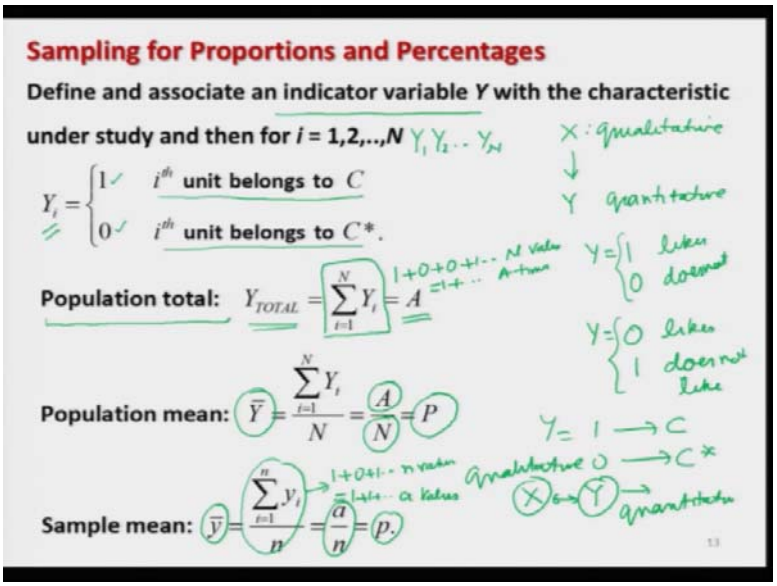


Now, from this population of size N, we are going to draw a sample of size n by simple random sampling that can be SRSWR or SRSWOR. And from the sample, we try to count the number of units which fall into the class C and we are indicating this number by a; then obviously n - a will be the number of units that falls in class C*.

So, now from the sample also, we try to estimate the or we try to find out the sample proportion. The sample proportion of the units belonging to class C is indicated by p, which is equal to a/ n and the sample proportion of the units in class C*; this is q is equal to n - a/n, which is 1 - a/n, which is equal to 1 - p.

So, you can see here, once again I will reiterate that P P + Q is equal to 1 and p + q is also equal to 1, ok.

(Refer Slide Time: 25:31)



So, now, this is our setup. But you have to remember one thing, we have defined here X and X is qualitative. The problem with the qualitative variable from the statistical point of view is that, that no statistical tool can be used on these values, which are in terms of say yes or no or good or bad. But we need to convert these values on the X to some quantitative variable say, let me denote this quantitative variable by Y.

So, now I am saying that Y is the variable which is quantitative. And in such a cases you will agree with me that, it is a big issue that how to indicate this good or bad or yes or no; because they are only indicating the absence on or the presence of the characteristic, they are not quantifying it. For example, if I say height is equal to 150 centimeter and say and height is equal to 170 centimeter; that indicates that the person with height 170 centimeter is taller than the person who is have having the height 150 centimeter.

But in case of indicator variable this is not possible; because somebody is saying say Y. So, how Y is good or bad? So, how to quantify it? So, we assume that Y is an indicator variable. What do we mean by indicator variable? So, then the meaning of indicator variable is this, it is only indicating the absence or presence by some numerical value; there is no concept of magnitude for example, if I say Y is equal to 1 if person likes the coffee and 0 if the person does not like the coffee, right.

So, 1 and 0 they are only indicating the absence or presence of the taste for a particular person. And in case if you ask me, I can also define it in the opposite way; for example, I can define Y equal to 0, if a person likes the coffee and 1, if the person does not like the coffee. So, that is the choice of the experimenter that what he chooses. Here in order to proceed further, we assume that Y takes value 1 and 0 and 1 is indicating the class C and 0 is indicating the class C*.

So, now I have collected a sample from a population of size N. So, what will happen in the case of population; there are units which we have denoted earlier by $Y_1, Y_2, \ldots, Y_N$. So, now, this $Y_1, Y_2, \ldots, Y_N$ they are the N number of persons in the population and now they are trying to indicate their response, whether they like the coffee or not.

So, suppose I assume that Y_i, which is the response of the ith unit this takes value 1 or 0; 1 is indicating that that the response belong to class C and 0 indicates that the response belongs to class C*, right ok. Now, based on this Y, we have to proceed further. So, now, what you can do?

I have made here a one to one correspondent between X and Y; X was qualitative, and Y is quantitative. And I have make here the one to one correspondence between the two; one to one correspondence does not mean that by looking at the value of Y, you can determine the value of X, means unless and until their order or position is not known to us, right ok.

So, first let me define the population value. So, the population total is indicated by here $Y_{TOTAL} = \sum_{i=1}^{N} Y_i = A$ and which is going to take a value here something like 1 + 0 + 0 + 1 up to here N values. So, now, you can see here you have assumed that the number of units which are belonging to class C are here A and these values are denoted by 1.

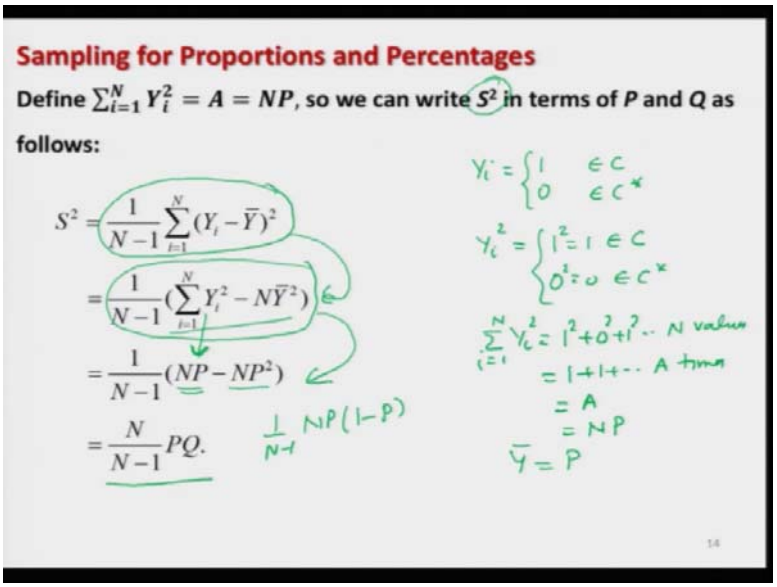
So, now this is going to be 1 + 1 + 1 up to A times. So, this will be equal to here A. So, the y total becomes here A. Now, in case if you go for population mean; population mean $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$.

So, this value here is A in the numerator, and in the denominator this is N the population size. So, you can recall that this quantity A/ N is population proportion P. And similarly for the sample, the sample mean say $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

So, this number you have seen that in the sample, there are n number of units which are in class C. So, this $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ will be something like 1 + 0 + 1 up to here n values. And out of this n values, there will be 1 + 1 + 1 a a number of values. So, this sum will become here a/ n, which is your sample proportion p.

So, now you can see here that, I have made here one to one correspondence between the X and Y in terms of population total, population mean and sample mean.

(Refer Slide Time: 32:21)



Similarly I can make a one to one correspondence between the quantities related to the variance. So, for example, you had earlier considered a quantity S^2, which was defined here like this $\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$.

Now, so what you have to see here is the following; this quantity can be expressed as like this, which we discussed in the earlier lecture $\frac{1}{N-1}\left(\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2\right)$

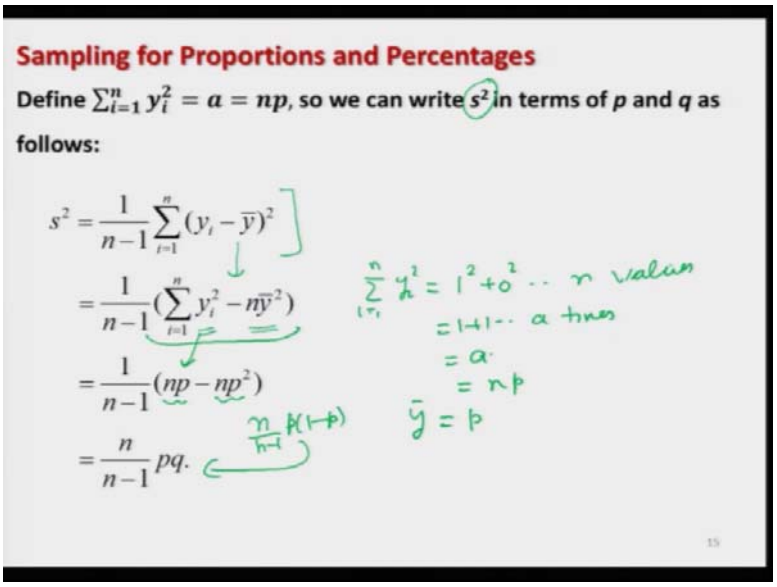
Now, if you try to see this Y i takes here value 1 and 0 if a unit belongs to class C or if the unit belong to class C*. Now, in case if you try to find out the value of Y_i^2; this will also be 1^2 which is equal to here 1 and 0 square which is equal to here 0, if the person or the unit belong to class C or if the unit belong to class C*.

So, in case if you try to find out the value of $\sum_{i=1}^{N} Y_i^2$; this is going to be a value like $1^2 + 0^2 + 1^2$ … N values. And so, this is going to be 1 + 1 + 1 how many times? A times. So, this is your here see here A; and A here is nothing, but N times P. So, that is what I am trying to use here right. And for about $\overline{Y}$, you already have proved that this is equal to P.

So, this quantity can be written here as say in place of summation Y i square; i can have here N into P. And in case of N times $\overline{Y}^2$, I have here NP^2. So, this quantity is nothing, but your $\frac{N}{N-1}P(1-P)$, which is nothing, but your $\frac{N}{N-1}PQ$.

So, you can see here, now I have made one to one correspondence between the quantity S square when the variables are transformed from quantitative to qualitative.

(Refer Slide Time: 34:51)

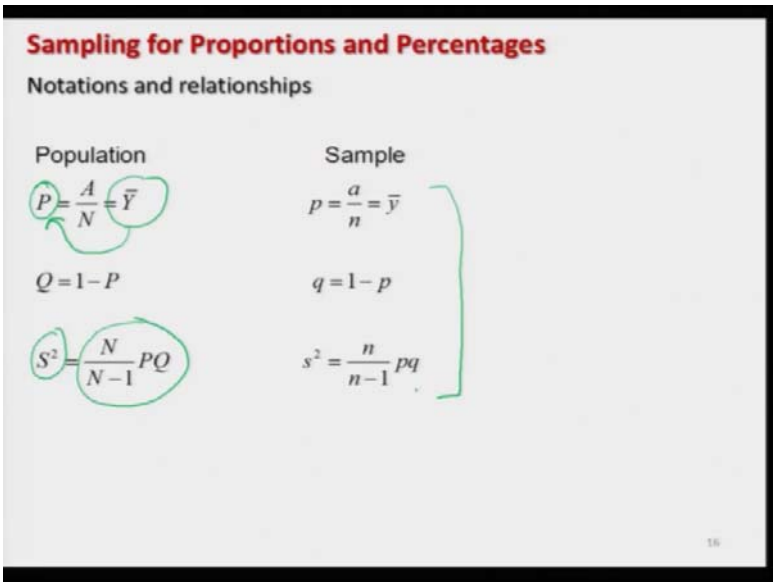


So, the same exercise can be done on the s^2 also; if you remember say, s^2 we have defined as the sample version of the S^2 is $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

So, this quantity will come out to be $\sum_{i=1}^{n} y_i^2 - n\bar{y}^2$. So, I mean as exactly on the same way $\sum_{i=1}^{n} y_i^2$ will be something like 1 square + 0 square and there will be such n values. And so, this will be 1 + 1 + 1 a times; so this is nothing but your, a. And a is you simply here n into p, and $\bar{y}$ was already you have proved this is p.

So, this quantity can be written here like this, summation y i square will be n times p, and $n\bar{y}^2$ will be n p^2 and finally, this will come out to be here $\frac{n}{n-1}p(1-p)$ which is over here $\frac{n}{n-1}pq$, right.

(Refer Slide Time: 36:08)



So, briefly I can comprehend these values over all the values in this slide; whatever was your population mean in case of qualitative variable $\bar{Y}$, this is now converted to P. And whatever was your S square in case of qualitative quantitative variable, this is now

converted to $\frac{N}{N-1}PQ$ in case of qualitative variable. And similarly is the story in the case of sample also.

So, now in this lecture I will stop here. What I have done that, I simply have transform the quantities that we use in the case of simple random sampling like as a $\bar{y}$, $\bar{Y}$, and S^2, which were defined for a quantitative variable, I have transformed then to the setup of qualitative variable.

Now, you can recall that, whatever properties we had studied in case of simple random sampling and that either finding out the properties of sample mean, properties of sample variance, the variance of sample mean, the estimate of variance of sample mean; they can directly be converted into the setup of a qualitative variable.

So, this topic I will try to take up in the next lecture and I am stopping here because I want to give you some time, so that you can refresh your concepts and terms that you studied during the simple random sampling. So, on the next turn, please come with a quick revision of what was a $\bar{y}$, what were the properties of $\bar{y}$, like as bias or say unbiased; how we have found the variance of $\bar{y}$ in case of SRSWOR and WR, and how we have estimated the variance of a $\bar{y}$ in case of SRSWR and WOR, right.

We also constructed the confidence interval. So, you just try to have a quick review of those concept and I will see you in the next lecture, till then good bye.