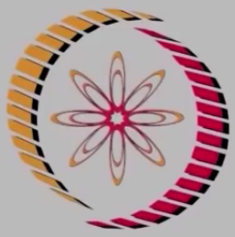
Indian Institute of Technology Kanpur



National Programme on Technology Enhanced Learning (NPTEL)

Course Title

Descriptive Statistics with R Software

Lecture - 31

Association of Variables

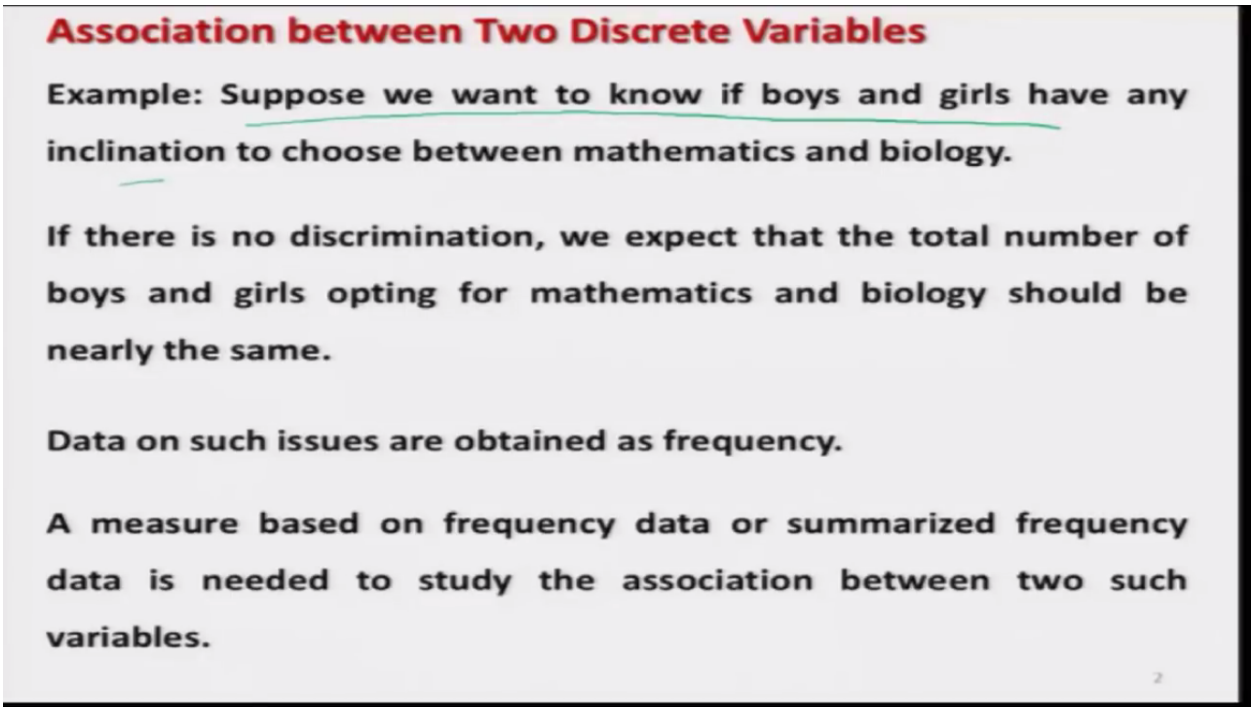
Measures of Association for Discrete and Counting Variables : Bivariate Frequency and Contingency Tables

by
Prof. Shalabh
Department of Mathematics and Statistics
IIT Kanpur

Welcome to the course descriptive statistics with R software. You may recall that in the last couple of lectures we have understood how to measure the association between continuous variable and rank data and we had the concept of correlation coefficient and rank correlation coefficient. Now the third case that is left is how to measure the association between two variables which are counting in nature or they are discrete in nature.

So in this lecture we are going to understand two concept; bivariate frequency tables and contingency tables and we will also see how to implement the concepts on the R software and in the next lecture we will continue with some more measures of association in the case of counting variables or counting data. So let's just start our discussion here. First of all you have to understand that what is a bivariate frequency table. You may recall that we have already discussed the concept of a frequency table and in that case we have essentially considered the univariate frequency table. Univariate frequency table means there is only one variable on which the data has been collected and the data was tabulated. Now suppose there are two variables and we want to tabulate the data. So this type of table will be called as bivariate table.

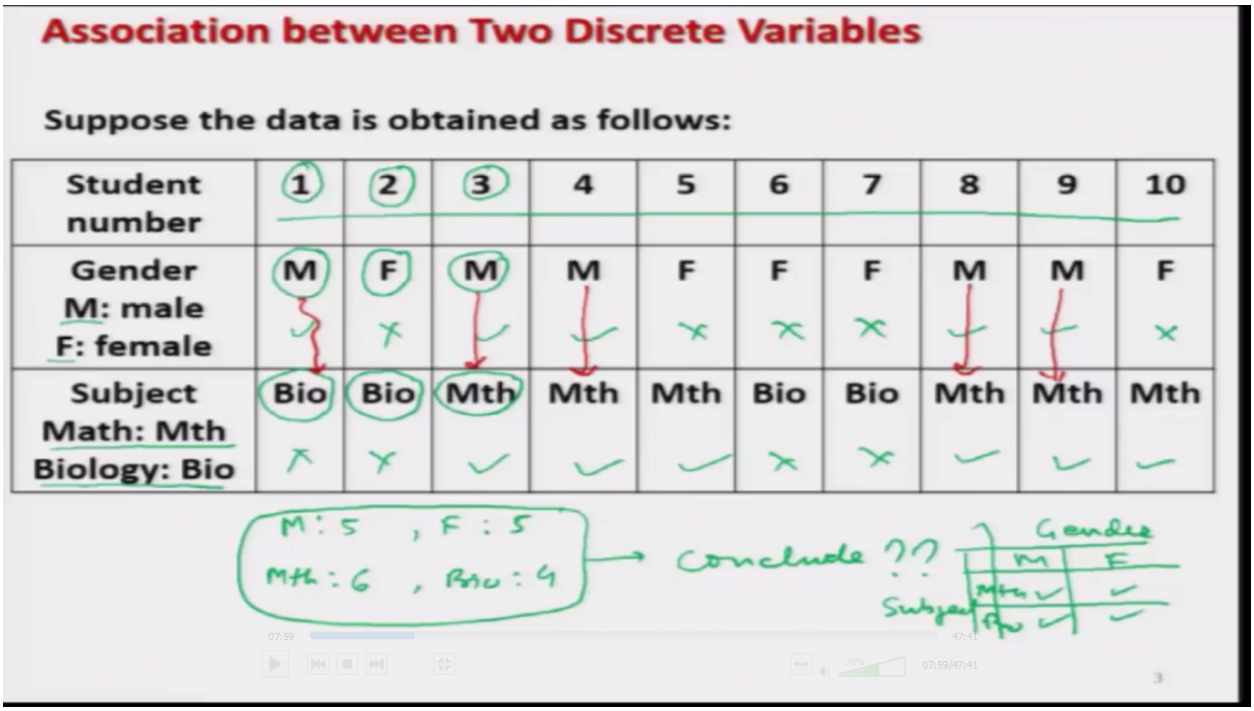
Now the question is this how does this come into picture and how the measures of association in this case of counting variable comes into picture. Now suppose you want to know in a college that what is the choice of the subjects between say mathematics and biology between boys and girls students, male and female students. Now what are we going to do? We will take a sample of students consisting of both boys and girls and we will ask their choices; do they like mathematics or biology. Now what we expect that in case if there is no choice of the subject with respect to the gender then the number of students who are considering to study mathematics or biology they should be the same number of boys and same number of girls but if a particular gender gives more choices for a particular subject that would indicate that yes the choice of the subjects between biology and mathematics that is being dominated by the gender.



So what are we going to do here let us take this simple example and it's the same example that suppose we want to know if boys and girls in a college have say an inclination to choose between mathematics and biology.

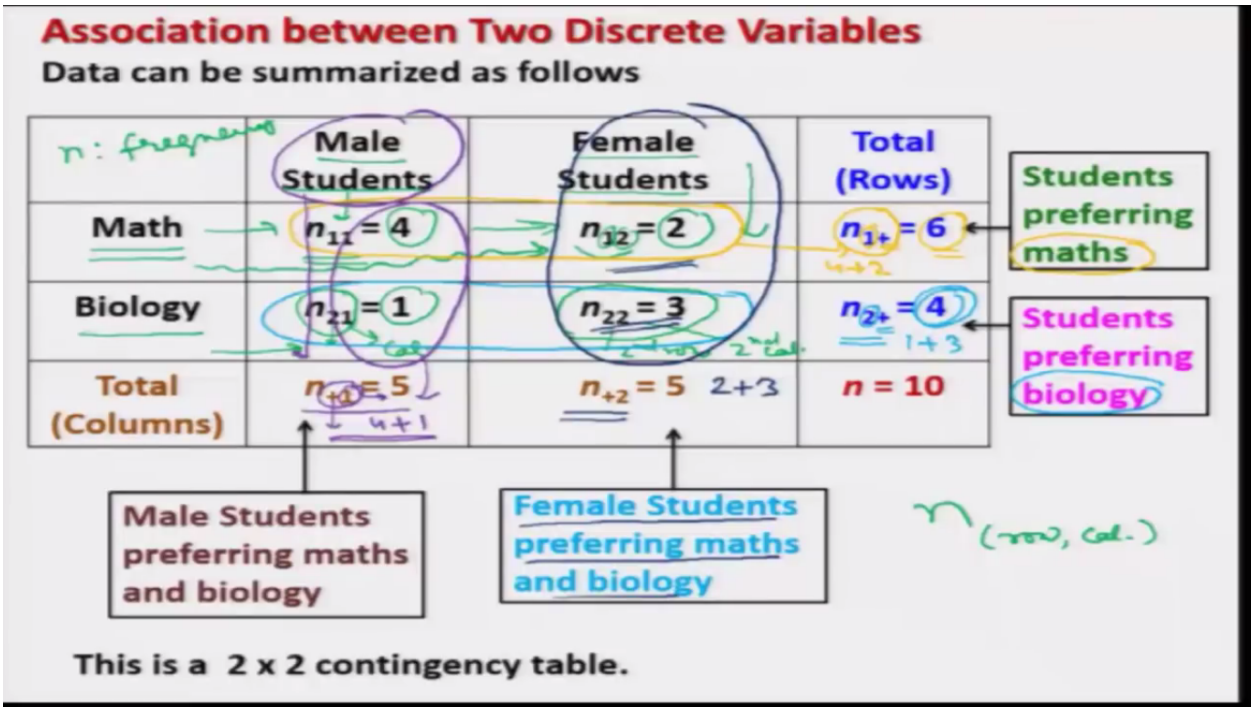
So obviously as we said that if there is no choice no discrimination between the two subjects then we expect that the total number of boys and girls opting for mathematics and biology should be nearly the same. And now we have to collect the data to make a conclusion on such an opinion.

So the data is collected on boys and girls with respect to biology and mathematics and this data is obtained in the form of a frequency and now we need to summarize this data in to a frequency table and based on that we need to devise a measure based on this frequency data or summarized data to study the association between two such variables which are basically counting in nature.



Now suppose I take a sample of 1, 2, 3, 4, up to 10 students and we ask individually each and every student what is their choice and we note down their gender. So I am denoting the gender of male students as M and gender of female student as F and the subjects; the mathematics is being denoted by here Mth and biology is being denoted here as a Bio. So now the data is collected as follows. Suppose we ask the first student and first student is a male and he says that yes he prefers biology. After this we ask the second student and second is girl. So we write here F female and she answered that she prefers to have biology. And similarly we try to take the third student who is a male and this boy answers that he prefers mathematics and so on and similarly we try to ask all the students.

Now you can see here how many boys are here 1, 2, 3, 4, and 5. So number of males here are 5 and obviously the number of females here are once again 5. 1, 2, 3, 4, and 5 and now we try to see what is the data on the subject mathematics. This is here 1, 2, 3, 4, 5, 6. So there are 6 student who prefer mathematics and there are 1, 2, 3, 4 students who are preferring biology but now you can see from this type of data, from this type of frequency we are unable to conclude anything. So what we need to do here we try to create a bivariate frequency like this one where on one side we will express the gender and on other side we will make the subject.

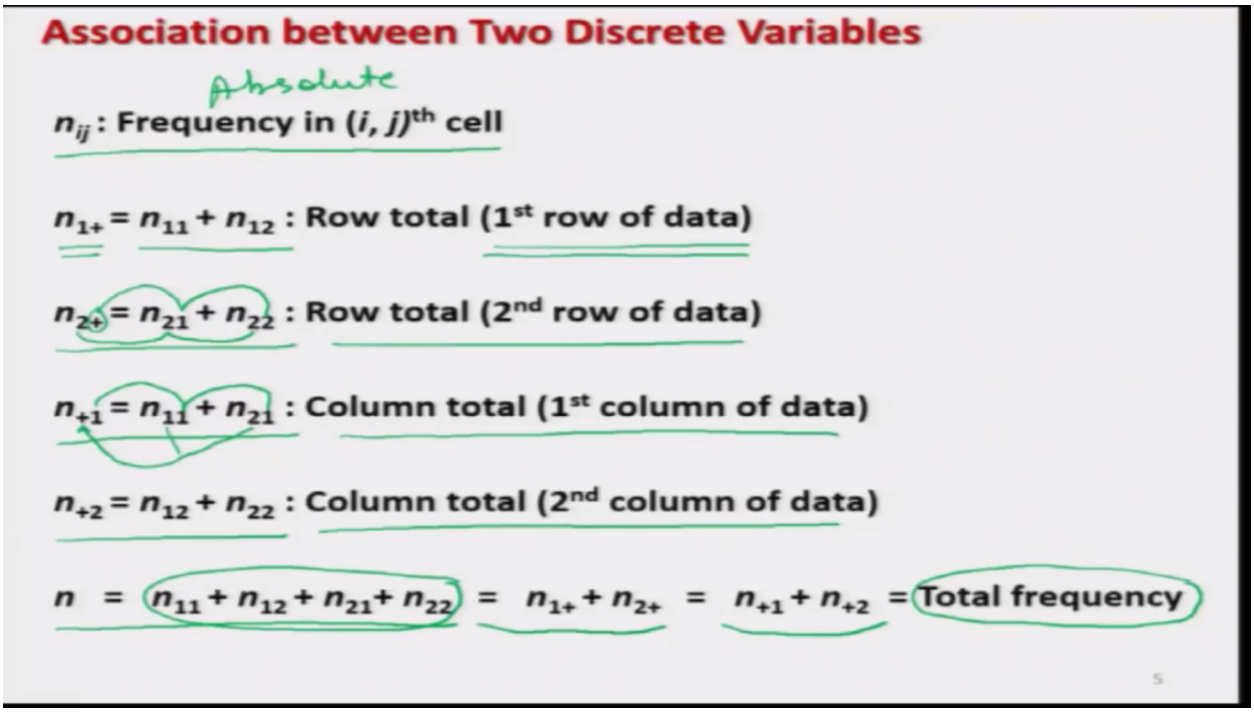


Now there are two genders male and female and there are two subjects see here mathematics and here biology and then we try to count these numbers that how many males are preferring maths and with how many females are preferring maths how many males are preferring biology and how many females are preferring biology. So you can see here that in this case the number of males who are choosing mathematics is number one, number two, number three, and number four. And males who are choosing biology is here in the first student only. And similarly the female students who are referring maths they are here one and two. And similarly if we collect the other data also and all this data is compiled here in such a table. So you can see here the number of male students who are choosing maths is four and this number we are denoting by n11, n is indicating the frequency and what is the meaning of 11, 1 is corresponding to first row and first column. Similarly the number of female students who are preferring here mathematics this number is given in this cell and this number here is 2 and so we write this frequency as n12 and 12 means 1 is the row this is the first row and 2 is the column this is the second column. And similarly the male students who are choosing biology this number here is one and this is denoted as the frequency n21 so once again 2 is going to denote here the second row and this 1 is going to denote the column. And similarly the number of female students who are choosing biology is here n22 is equal to 3 this means the second row and second column the data in the second row and second column. So you can see here I have denoted here the frequency as here say row and column. So the address is given by two numbers row and column for a particular type of frequency in any class. So these are essentially the absolute frequencies.

Now the next step is this I try to count the numbers row and column wise. Suppose I count the numbers in the first row. This is here 4 plus 2 and this is here 6 and if you try to see I am trying to denote this number here as say n1 plus. So n1 is going to indicate that the subject in the first row which is here mathematics, this one and this plus is indicating that we have added the

frequency over the column. N1 plus equal to 6 is going to give us information that how many students are referring maths and similarly in case if I try to sum the frequencies in the second row, this is denoted here as say n2 plus and this number is 1 plus 3 which is equal to here 4. So once again here this 2, in this n2 plus 2 is indicating the subject in the second row which is here biology and plus is indicating here that this addition has been obtained on the second subscript which is column. So this n2 plus is equal to 4 this is giving us the information that there are four students who are choosing or preferring biology.

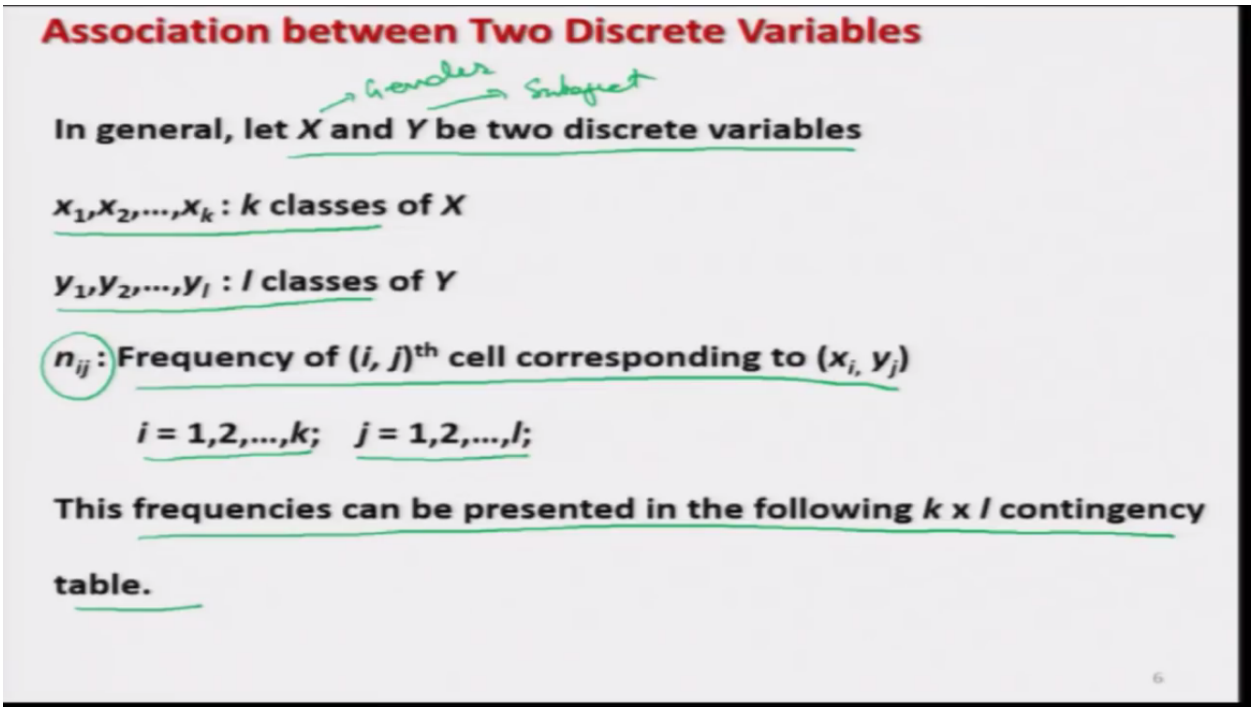
Now the same exercise can be done in columns. So when I try to add the numbers in the columns here like this then this comes here 4 + 1 and using the same philosophy of the symbols I am denoting this sum as n+1, +1 is coming as a subscript. So this plus is indicating that the sum has been obtained on the first column or this is the sum of the frequencies in the first column and this 1 here this is indicating the first column. So if you try to see this number here n plus 1 this is equal to 5 is indicating the male students here who are choosing maths as well as biology or any subject out of this. And similarly if I try to come on the second column here and then I try to add here n12 and n22 this is here 2 + 3 which is equal to here 5. So this number here n+2 this is going to indicate that the sum of the frequencies has been obtained based on the second column by n+2. So this is essentially giving us the number that how many female students are preferring maths and biology. So you can see here that in this case the entire data whatever we had obtained here on the basis of the sample has been classified into a two by two table and this type of two by two table is called as contingency table and in particular this will be called as two by two contingency table.



Now if you try to see what are the different symbols in general what are the indicating this I have summarized here. And you see here what is here nij in general this is the frequency in the ijth cell

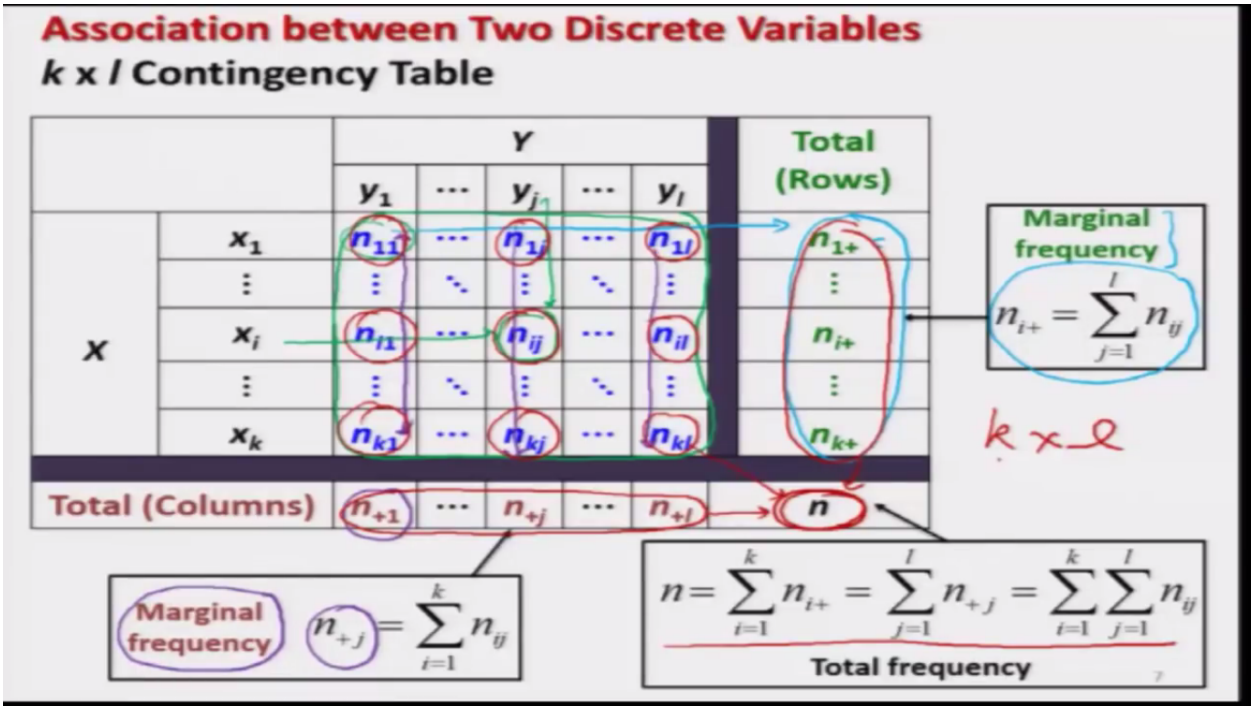
and essentially this is the absolute frequency in better terminology and similarly when I'm trying to take here n1 plus this is indicating the row total and 11 plus n12. So this is indicating the first row of the data or the sum of the frequencies in the first row of the table. Similarly n2 plus is equal to n21 plus n22. So you can see here this 2 remains the same and this plus is indicating that the sum of this 1 and this 2 has been obtained.

So this is going to give us the sum of the frequencies in the second row of the table and similarly in the case of columns n+1 is equal to n11 plus n21. So you can see here the sum is obtained over this 1 and 2 and this is indicated by this plus sign and this one and this one they remain the same and this is going to give us the column total of the frequency of the first column. And similarly n+2 is equal to n12 plus n22 and similarly this quantity is going to give us the sum of the frequencies in the second column. Now if you try to look in this table and try to see here n equal to 10 what is this n equal to 10. This is indicating the sum of all the frequencies and if you try to see here this can be obtained in different ways. First is this this is n11 + n12 + n21 + n22. This can also be obtained as here sum of the frequencies in the rows. So this is n1 + + and 2+ this is again equal to 10. So the sum has been obtained from the row and similarly if you try to take the sum of the columns then again this can be represented as n+1+n+2. So this is what I have mentioned here in the last line that n is equal to sum of all the frequencies which is here and this is the same as sum of the frequencies in the row and sum of the frequencies in the column and so n is going to indicate the total frequency. So that is going to be our general symbols and notations in contingency table.



So now if I try to make it more general. Suppose I try to take in general two variables, two discrete variables on which the observations are obtained as counting. So you can see here for example in the earlier case X was denoting the gender and Y was denoting the subject. And we had divided the data into two classes for X and two classes for Y. Two classes of X are male and female and two classes of Y are maths and biology. Similarly I can make it more general and we

assume that on the data on X variable we have created K classes which are denoted as X1, X2, XK and similarly for the Y we have created L classes say Y1, Y2, YL and this nij this is the absolute frequency of the ijth cell corresponding to the observations Xi, Yj and obviously i goes from 1 to K and j goes from 1 to L. So now in general these frequencies in general case they can be represented in k cross l contingency table similar to what we have represented in the two-by-two contingency table and this contingency table will look like this. I have used here different types of colors so that you can get an idea.



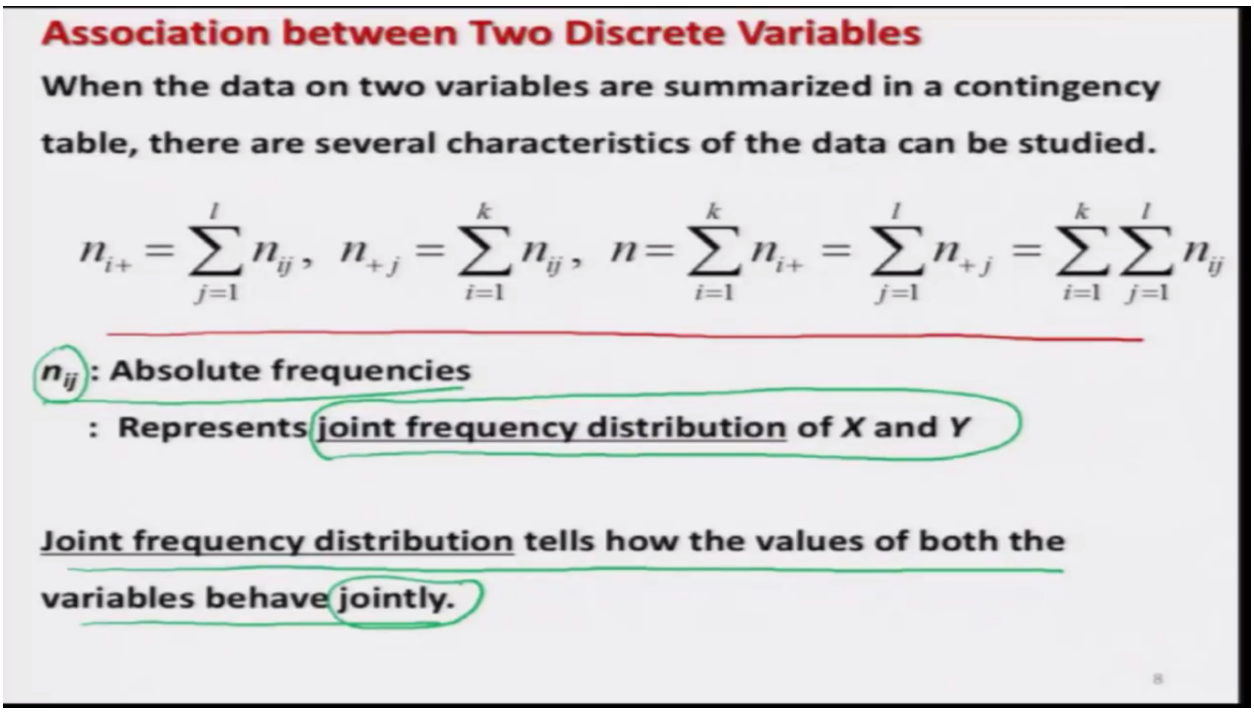
So you can see here this part in blue color this is going to indicate the absolute frequencies of different classes. So and 11 here is indicating the absolute frequency in X1, Y1 cell. Similarly here say nij that is going to indicate the frequency in the Xi and Yjth cell and so on. So all these values n in blue color they are going to indicate the absolute frequency of all the classes.

Now we try to find out the row sums and column sums. So we try to add all these frequencies here. So this is equal to n1 + is equal to n11 + n12 + n1j up to here n1l. So this is denoted here and this quantity is called as marginal frequency for the values here in this column which are indicating the total of the rows or the sum of the frequencies in different rows they are called as marginal frequencies and more precisely this is called as marginal frequencies of X or X1 X2 Xk different classes. Now if we try to do the same operation on the columns. Suppose I take the frequencies in the first column and I add them here. So this will be n11 plus n21 up to here and nl1 plus in general nkl.

So this sum is going to be denoted here as a n+1 and similarly if we try to do the same operation for each and every column here, here and so on. So these frequencies are denoted by here n+j

and if you try to see there the summation from 1 to K and ij means sum of all the frequencies in the column. So this is also called as marginal frequencies and they are essentially denoting the marginal frequencies of Y that means the module frequency of the class y1, y2, yl and so on.

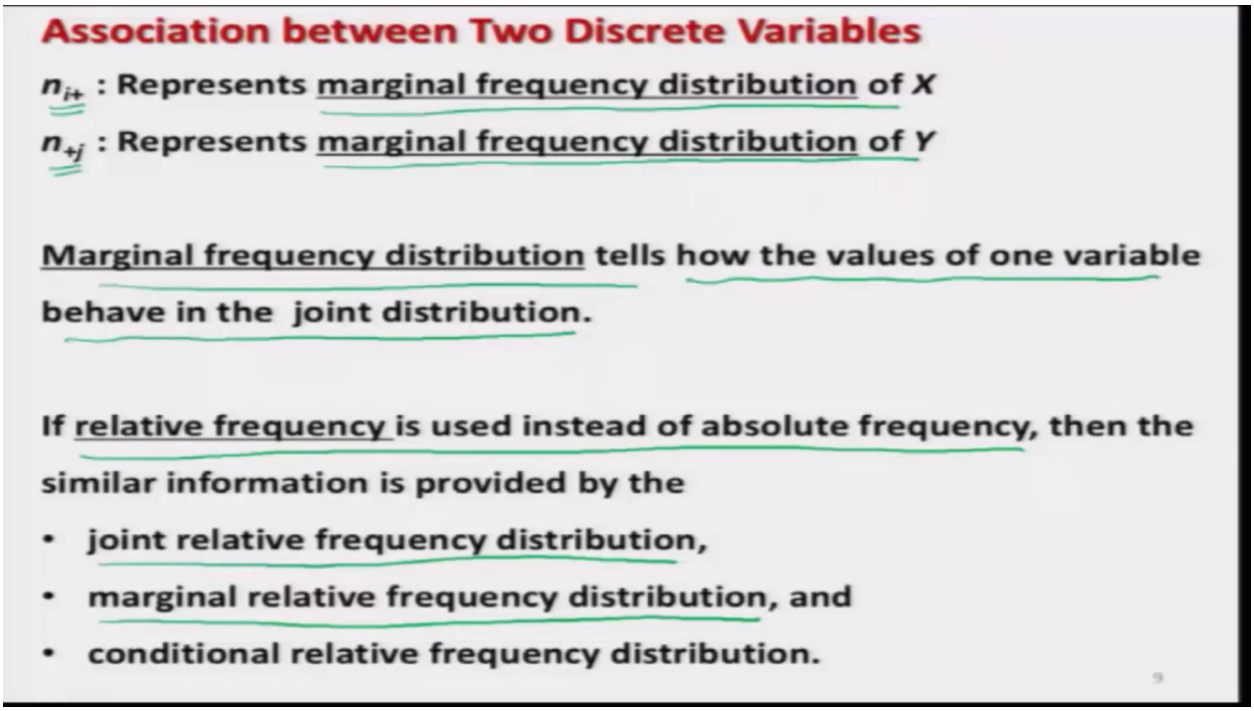
Now finally if you see this here n now this n can be obtained as a sum of all the frequencies which I am denoting here in red color or say all the frequencies which are denoted here in blue font, blue color. So this is the sum of all here n and this sum can also be obtained by adding the values in the rows that is n1+n2+ni+ up to here nk+ and this value will be going to be the same as n and similarly if you try to add here the marginal frequencies of the columns that will also give you the same value here and this I have denoted here in this expression. So this is called the total frequency. So this is how we try to interpret and we try to construct the contingency table and this contingency table is our k cross l contingency table. Why? Because there are K number of rows and there are L number of columns.



So now we have understood that all the data this can be represented in different types of frequencies and we simply try to summarize it here once again. This thing here and as we have discussed this nij are going to discuss about the absolute frequencies. Now in case if you try to see what are these nijs representing. These nijs are giving us different numbers about the choices between X and Y which are occurring together. So these values of nij they represent the joint frequency distribution of X and Y, Now you may recall that when we had discussed the frequency table in a univariate case then we had only one variable but now here I have two variable X and Y and we had discussed that how the frequencies are distributed in different class intervals that was compiled in a univariate frequency table but now since we have here two

variables or the frequencies inside the cell they are determined by two values; the value of X and the value of Y that is why this is called as joint distribution.

So this joint frequency distribution tells how the values are both the variables behave jointly and just to inform you here that here I am trying to take only two variables but these variables can be more. There can be three variable. There can be four variables and corresponding to those numbers so we can create the suitable contingency table. Suppose we have three variables X, Y, Z. X having two classes. Y having three class and Z having four class. So then we will create a table of the order 2 by 3 by 4 or 2 into 3 into 4 contingency table.

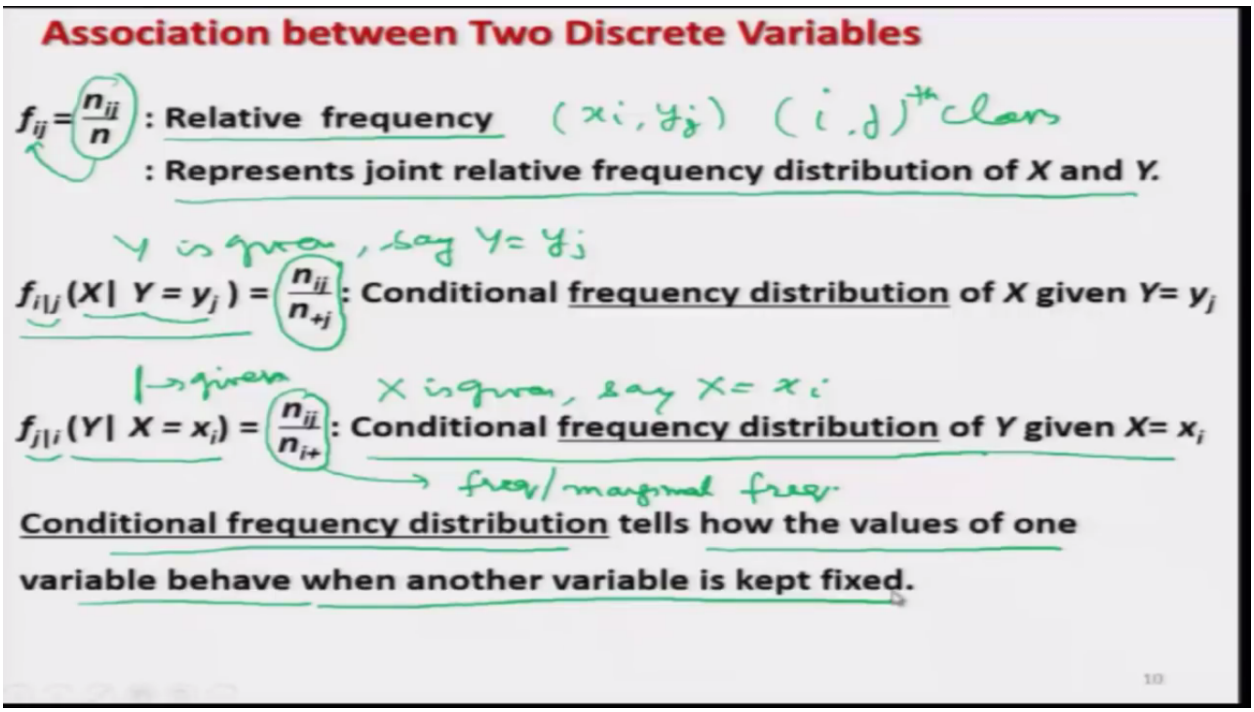


Now the next symbol here ni+ this was the sum of the frequencies and similarly n+j. This was again the sum of the frequencies in rows and columns. So these values are representing the marginal frequency distribution of X and marginal frequency distribution of Y. What does this marginal frequency distribution tells us? The marginal frequency distribution tells how the values of one variable behave in the joint distribution of X and Y. So now you can think here that these values are going to be determined by two values and we assume that the value is affected by two variables X and Y. So obviously one question comes that when we have the joint Distribution of two variables X and Y what is the contribution of X and what is the contribution of Y. So this information can be digged out from this bivariate frequency table or the contingency table by finding out the marginal frequencies.

Similarly when we had discussed the concept of frequency table then we had two types of frequencies; absolute frequencies and relative frequencies. The advantage of using the relative frequencies was that that the sum of all the relative frequencies will always be equal to one and

the relative frequency of every cell or any value will always be between 0 and 1. So this is very-very similar to the concept of probability. So in fact this frequency tables in univariate or say bivariate case they are indicating or they are representing the probability distribution of discrete variable in the case of say this probability theory.

So now in this case also in place of absolute frequency we can also use the relative frequency and relative frequency will be obtained simply by dividing the absolute frequency by the total frequency and a new table or a new contingency table can also be created using the bivariate frequency table based on the relative frequency. So in case if we try to use the relative frequency in place of absolute frequency then the similar information is provided and we call as joint relative frequency distribution. And similar to modular frequency distribution. Now we will have marginal relative frequency distribution and there will be one more concept what we call as a conditional relative frequency distribution.

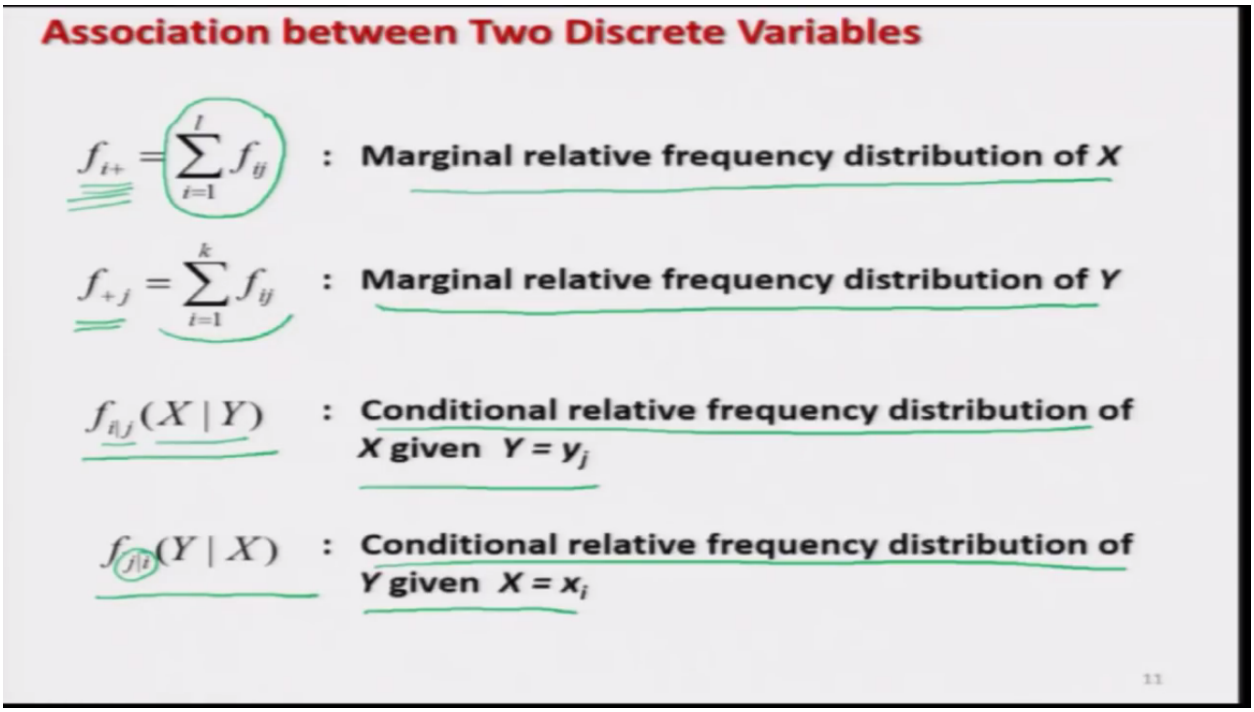


So what is this we try to understand here. You see the relative frequency of any class or say any class corresponding to xi, Yj or ijth class this will be obtained by nij upon n and this is indicated by the symbol fij. o similar to nijs s now this fijs will represent the joint relative frequency distribution of X and Y. Now we try to obtain one more quantity which is called as conditional frequency distribution. This conditional frequency distribution is obtained in two cases. When the value of Y is given or the value of X is given. So when the value of Y is given say Y equal to some particular value Yj then in this case the conditional frequency distribution of X given Y equal to Yj is obtained by nij divided by n+j that is the frequency divided by the marginal frequency of that class and this is denoted here say F of x given Y and Y is given as Y equal to Yj

and there is a subscript here i given j that's a standard symbol for indicating that conditional frequencies.

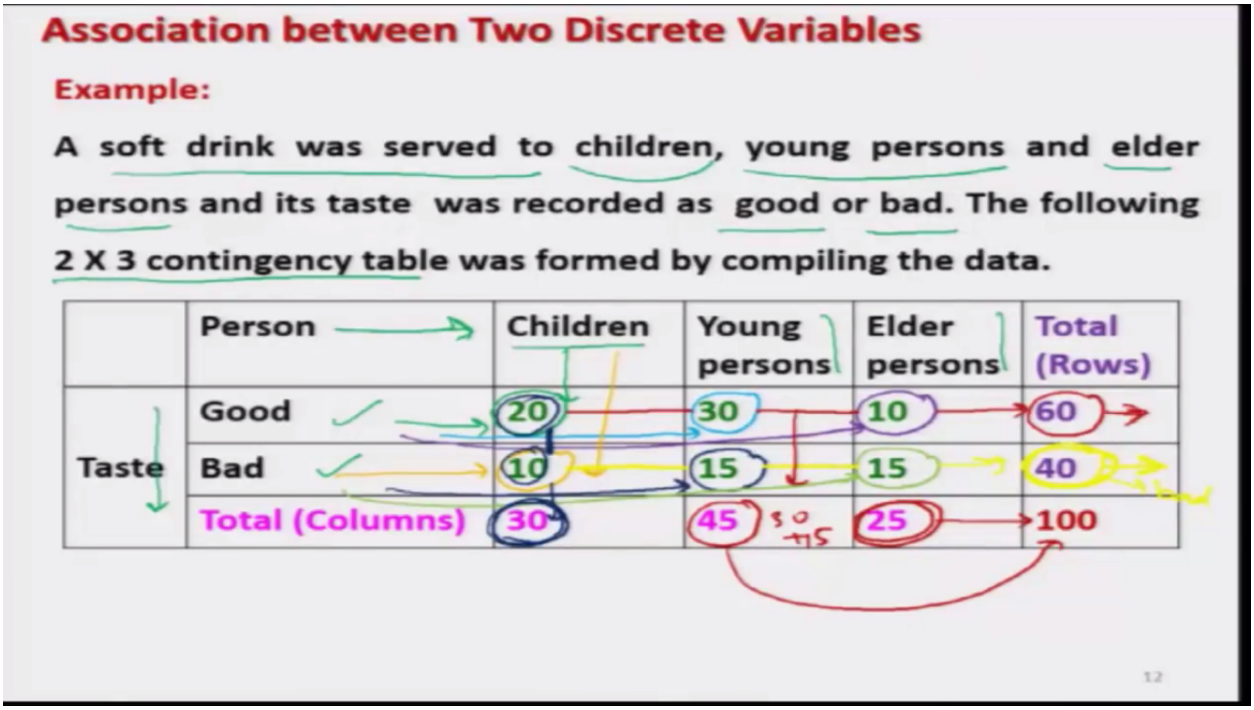
Now the second case will be that in case of Y is given suppose X is given. The value of X is given and suppose X is given as Xi then in that case the conditional frequency distribution of Y given X equal to Xi is obtained by nij upon ni+ so that is again the ratio of say this here frequency divided by marginal frequency and this is denoted as say F of y given X equal to Xi and in the subscript we write j given i. This symbol here vertical line this is thus this is called as given.

So these conditional frequencies or the conditional frequency distribution gives us an information that how the values of one variable behave when another variable is kept fixed. For example we have considered the case of gender versus subject. Now suppose I want to know what is the behavior of the subjects for a given gender. Then this type of information can be obtained by the concept of conditional frequency distribution. So I will try to take an example to show you that how to interpret such values. But before that let me just write all the symbols in general.



So when I try to find out the sum of all the frequencies in the rows and columns corresponding to X and Y they will give us the marginal relative frequency similar to the concept of marginal frequency. So can I try to sum all the relative frequencies corresponding to X then for the ith class I get fi+ which is here and because of the sum of all the frequencies in that particular class or say particular row. Similarly the modular relative frequency distribution of Y values or the classes in Y this is denoted as f+j and this is the sum of i goes from 1 to k fijs. And similarly the

conditional relative frequency distribution of X given Y equal to Yj this will be denoted by fx given Y and here i given j in the subscript and similarly the conditional relative frequency distribution of Y given X equal to Xi this is denoted by here f of Y given X and in the subscript j given i.

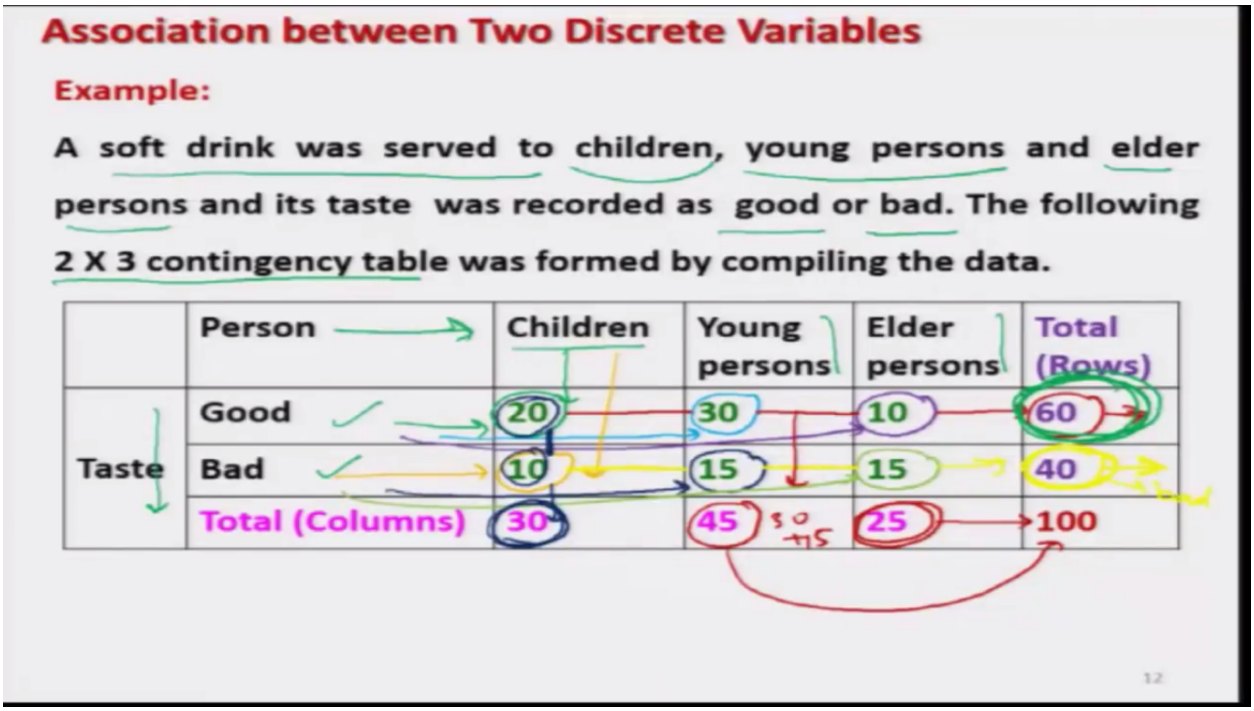


So now let me take a very simple example and we try to first understand that how these values are obtained and how to interpret them and after that I will show you that how to create the contingency table and this type of marginal frequencies inside the R software. Suppose I have an example here where soft drink was served to some persons and those persons have been divided into three groups depending on their age. First group is children. Second group is young person. And third group is elder persons. And they would ask that how the drink taste and they were given two options whether the taste is good or the taste is bad and based on that we have obtained the data. For example you can see here in the row whatever the data I have obtained this has been counted and compiled in a 2 cross 3 contingency table like as follows. You can see here in the row I'm writing here three classes of children, young persons, and elder persons and in the column I am taking another variable taste and which has two classes good and here bad and then based on the data collected from such hundred persons if we try to count that how many children said that the drink is good and this number is supposed 20. So how many children and taste is good here and similarly we try to count that how many young person said that the drink is good and similarly we try to count that how many elder persons said that the drink is good and this number is here 10. And similar information was obtained for children and there are 10 children who said that the drink is bad. Similar to this there are 15 young person who said that the drink is bad and similar to this there are 15 elder person who said that the drink is bad. Now you see in

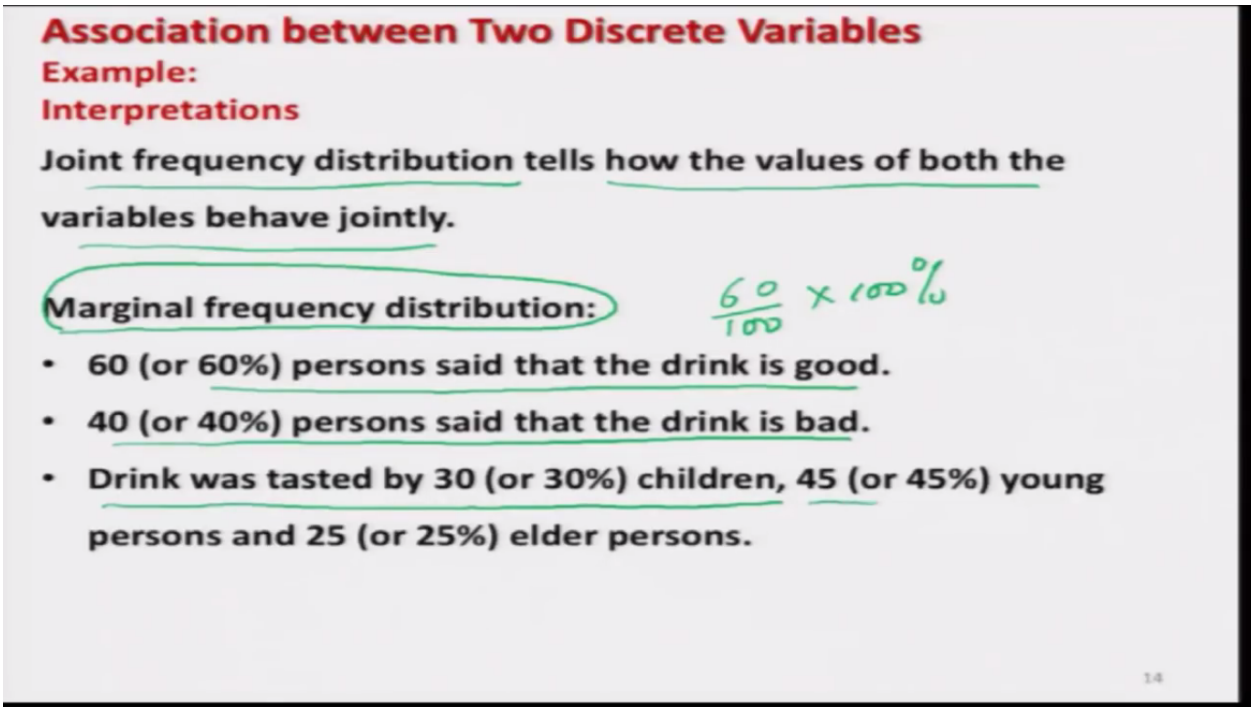
this data we have three groups on each and two groups on taste. So there are three classes of age and two classes of taste and one can see here that the taste and age are not independent. Different people in different age groups they are giving a different opinion. Had this drink been very good then we expect that all the person, all the 100 person would have said the drink is good but this is not happening here. So this is indicating that the variables X and Y are not independent but they are correlated and my issue is and my question here is this how to measure this association.

So there are different types of measures which have been suggested and all those measures tries to measure this association in different ways. So the objective here is to understand what are those measures and how are they going to give us information.

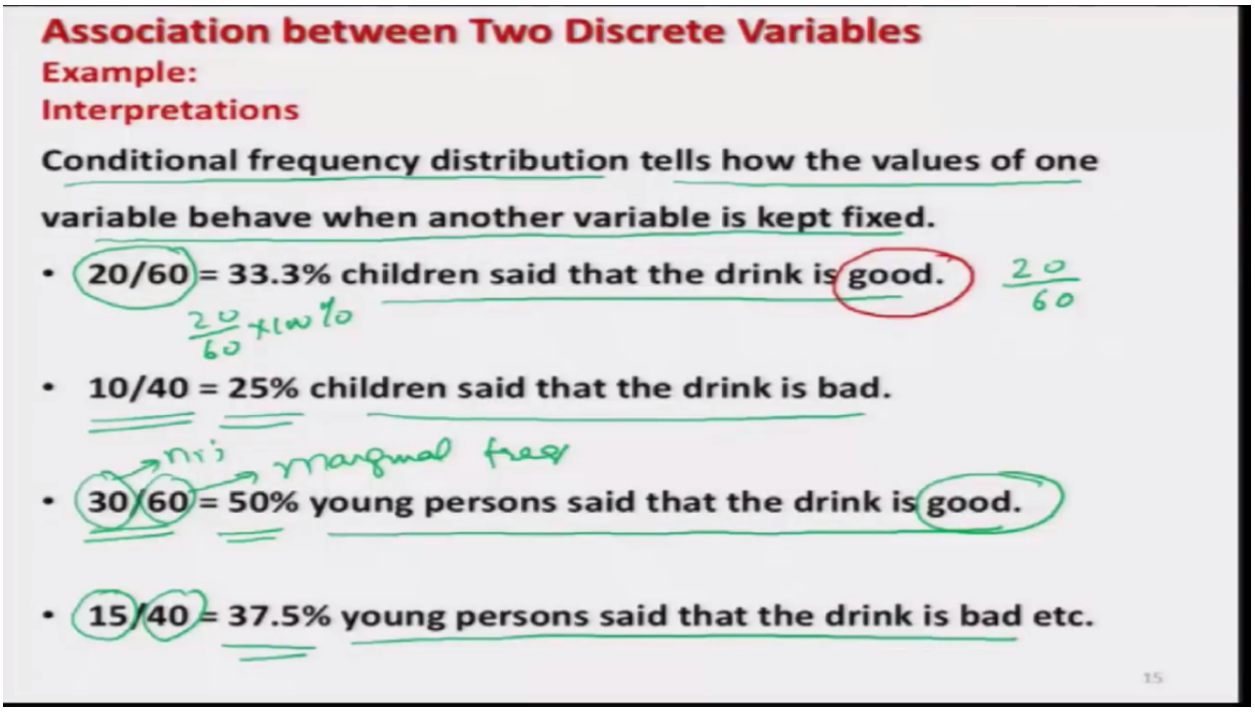
Now after this I try to find out their row sums and row columns so you can see here the sum of 20 plus 30 plus 10 here is 60. So this 60 is going to give us the information on the marginal frequency. So I can see here by looking at this number 60 that there are 60% out of 100 who said that the drink is good and similarly this marginal frequency which is here 40 this is obtained as the sum of this 10, 15, and 15 and this is here 40 so this number 40 is essentially indicating that out of 100 there are 40 persons which are saying that the drink is bad. Now on the same lines let me try to add the numbers in the column. So you can see here I try to add here this 20 and this here 10 and this gives me value here 30. 20 plus 10 is equal to 30. So this number here at 30 this is giving us an information that out of 100 persons there are 30 children and similarly in the second column I get this number 45 which is equal to 30 plus 15. This and this number. So that is indicating that there are 45 young persons out of this hundred persons and similar to this and the last column of elder person this value is 25. So that is indicating that there are 25 elders percent in the sample of 100 persons. So you can see here that this marginal frequencies in rows and columns they are giving us a particular type of information and this information has been obtained by making one of the effect to be the constant. For example when I say that how many person said that the drink is good then we are suppressing the information on the age and we try to add simply children plus elder person plus young persons together. And similarly I want to find out that how many persons are there in different categories then I am trying to suppress the variable taste. I am not bothering who said good or who said bad but I am simply counting that how many children said the drink is either good or bad. Similarly how many young person or elder person said that the drink is good or bad. So this is the type of information what we obtained from the marginal frequency and now the similar information can also be obtained in terms of relative frequencies. What we have to do in the same frequency I simply have to divide each and every frequency nij by total frequency n.



So you can see here that total frequency here is 100. So first cell has a relative frequency 20 by 100 and similarly other cells have 30 by 100. 10 by 100, and 10 by 100, 15 by 100 and 15 by 100. Now once again what I try to sum them row wise then the sum is 60 by 100 and 40 by 100 in the first and second rows. So they are essentially trying to give us the marginal relative frequencies and in the columns when I try to add it here this number is 30 by 10. 40 by 100 and 25 by 100. So this is trying to give us the same information in terms of relative frequencies and obviously here the sum of all the frequencies will always be equal to one.



Now so if you try to see what type of information I have got here from this table well I'm not going to discuss here each and every information but I will try to give you some information. So one thing is clear that this is a joint frequency distribution and it is informing us that how the values or both the variables behave jointly. So when I try to see here about the marginal frequency distribution so you can see here in this case where I am now making here as circle if you try to see here 60. So there are 60 person who said that the drink is good and there are 40 persons who said that the drink is bad. So in general I can also write it there as 60 out of 100 and into if I multiply by 100 this will give us the value in percent. So I can say in general that 60% of the person said that the drink is good and similarly 40 or 40% percent person said that the drink is bad. And similarly if I try to look at the values in the column say here, here, and here then 30%, 45% and here 25% persons in the sample are children, young persons, and elder person. So you can see here that I can say that there are 45% young person and 25 or say 25% elder persons in the sample.



Similarly if I try to find out the frequency distribution in terms of conditional frequencies then this conditional frequency distribution is giving us an information that how the values of one variable behave when another variable is kept fixed. So you can see here I am obtaining here a value 20 by 60 and I am saying that 20 upon 60 into 100% children said that the drink is good. How this has been obtained? If you try to see how this 20 and 60 values are coming. So now you can see here I will try to make here a circle in a red color if you try to see this data here, this is here that 20 and this is here the 60, the marginal frequency. So you can see here I'm trying to fix one variable here which is good. This I'm now fixing and that is the idea of the conditional frequency distribution that I have fixed the variable here good and then I am trying to find out the conditional frequency by taking the absolute frequency 20 and the marginal frequency 60 and this is going to give us an information that how many children said the drink is good and similarly there will be another information that well how many children said that the drink is bad so for that I try to take the information on here the bad and here the frequency is 10 and the modular frequency here is 40. So I try to take here 10 upon 40 or this is equivalent to 25% of the children said that the drink is bad and similarly if I try to come under columns you can see here that I'm trying to take here different values from here and I am trying to say here that 30 upon 60 which is equal to 50% of the young person said that the drink is good. So I'm again fixing the variable here good and then I'm trying to take the frequency here say nij divided by marginal frequency and similarly I can have the information about the young children, sorry the young person who said the drink is bad. So this can be divided by the total number of young persons divided by the marginal frequency and this comes out to be 37.5% people said that the drink is bad.

So now if you try to see what we have done here we have understood the concept of how to create the bivariate frequency table and in turn how to convert them into a contingency table and

this contingency table is going to give us different type of information. Now the next question is how to create this frequency table and contingency table or the contingency table from the frequency table using the R software. So this I would try to discuss in the next lecture.

In this lecture you please try to revise all these concepts and try to understand them what they are trying to say, what they are trying to tell. Once you understand them then getting the output from the software is very simple but the main thing will be how to interpret those values. So you practice here and I will see you in the next lecture. Till then good bye.