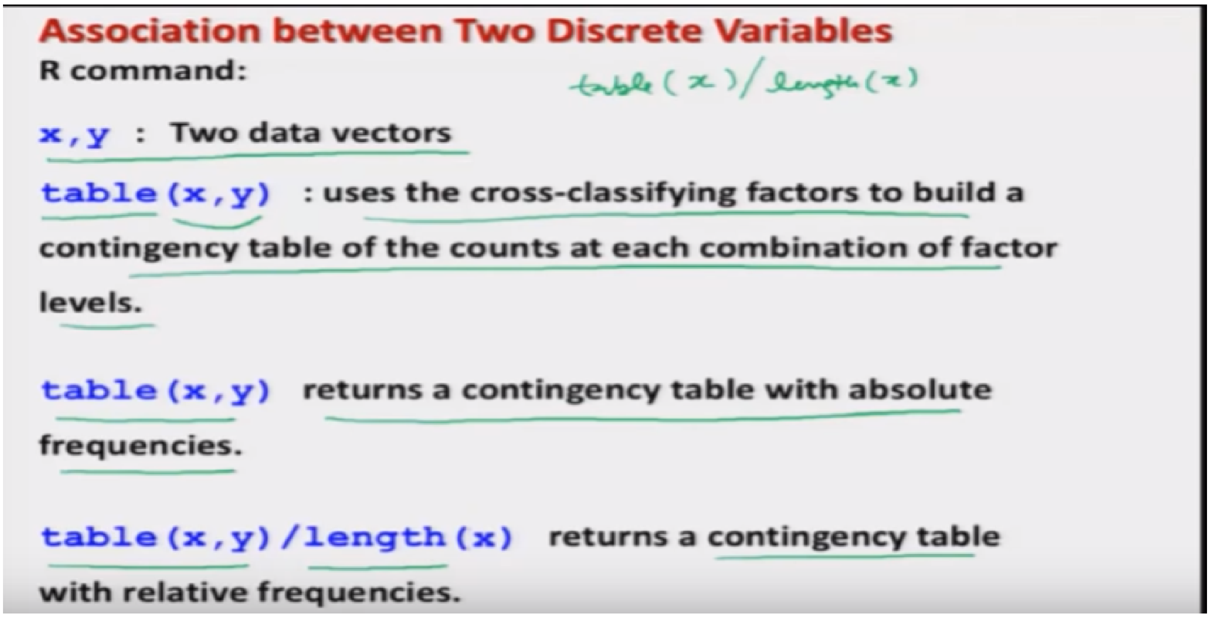
Lecture-32
Association of variables

Measures of association for Discrete and Counting

Welcome to the next lecture on the course, descriptive stats it with R software. You may recall that in the last lecture we started our discussion on measuring the association between two discreet variable, on which the observations were obtained as numbers obtained by counting, and in that lecture we had discussed that from the given set of data we can create a table, and a contingency table. From that contingency table we can obtain marginal frequency distribution and the conditional frequency distributions, and this marginal and frequency distributions can be obtained in terms of absolute frequency and relative frequency, and we had taken an example and we understood how these values are coming and we understood how to compute them manually.

Now, in this lecture I will try to show you that these contingency tables can be obtained inside the R software. So I will take an example and I will try to show you that how to obtain the contingency table and after this I will introduce some measures, some quantitative measures to find out the magnitude of the Association or the degree of association between the two variables. So now first I try to take the topic that how to construct the contingency table in R software. Right?

Refer Slide Time: (2:03)



Association between Two Discrete Variables

R command:

table(x)/length(x)

x,y : Two data vectors

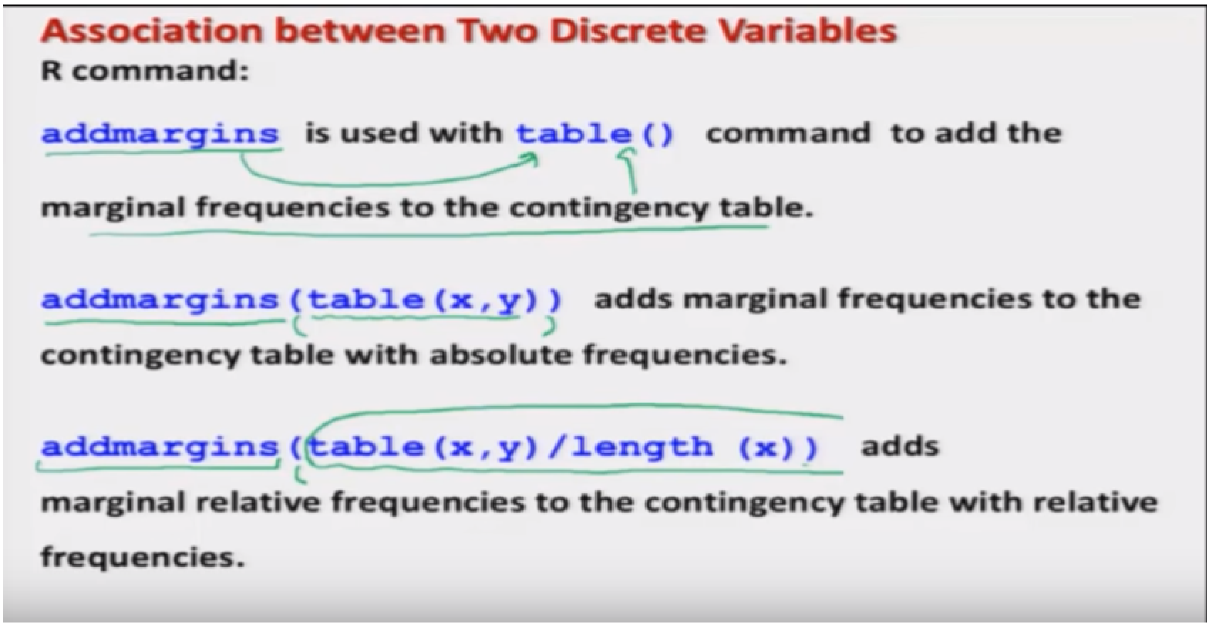
table(x,y) : uses the cross-classifying factors to build a contingency table of the counts at each combination of factor levels.

table(x,y) returns a contingency table with absolute frequencies.

table(x,y)/length(x) returns a contingency table with relative frequencies.

So as usual I will assume that we have got here a data vector and suppose we have two data better x and y, you may recall that in the case of univariate frequency distribution if I have the data vector as x then we had used the command table to find out the frequency table. And when this table was divided by the length of x then we had got the frequency table in terms of relative frequencies. Similar to that when we want to tabulate the bivariate data the same command is used that is table ‘table’ the only difference is that now inside the argument you have to give that two variables or two data vectors, and similarly if you have more than two you can express all those data vectors here separated by comma. So, this table (x, y) is used to cross classify the factors to build a contingency table of the count set each combination of the factor levels. Right? and if you try to use this command table (x, y) this will give you an output in the form of a contingency table with absolute frequencies, and similarly if you try to divide this table by the length command then it will return a contingency table with relative frequencies.

Refer Slide Time: (3:36)



Association between Two Discrete Variables

R command:

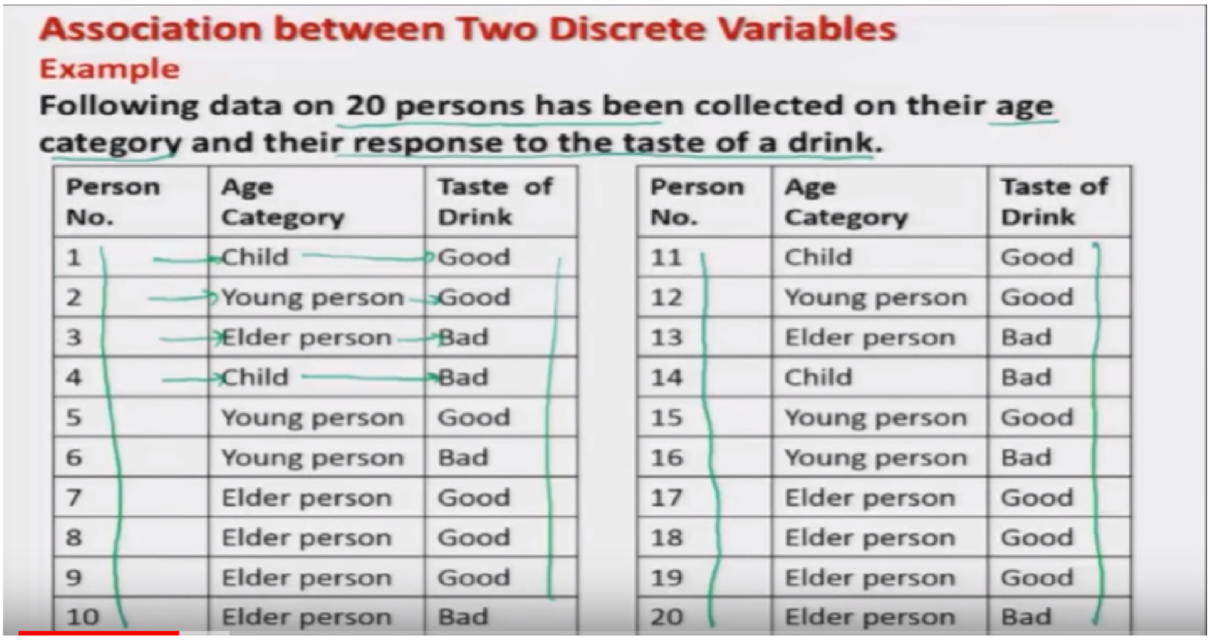
`addmargins` is used with `table()` command to add the marginal frequencies to the contingency table.

`addmargins(table(x,y))` adds marginal frequencies to the contingency table with absolute frequencies.

`addmargins(table(x,y)/length (x))` adds marginal relative frequencies to the contingency table with relative frequencies.

Now in case if you want to find out the marginal frequencies then there is a command here addmargins 'addmargins' and this command addmargins is used along with that table command, and this gives us the marginal frequencies to the contingency table that was constructed by the command table. So the entire command to obtain or to add the marginal frequencies in the contingency table will become addmargins and inside the argument try to tell adding margin to what? So add margin to the contingency table which was provided by the command table (x, y), and in case if you want to obtain the marginal frequencies in terms of relative frequencies, then you simply try to use this addmargins command inside the argument and inside that argument tried to write down the contingency table of which you want to obtain the marginal relative frequencies. Okay?

Refer Slide Time: (4:51)



Association between Two Discrete Variables

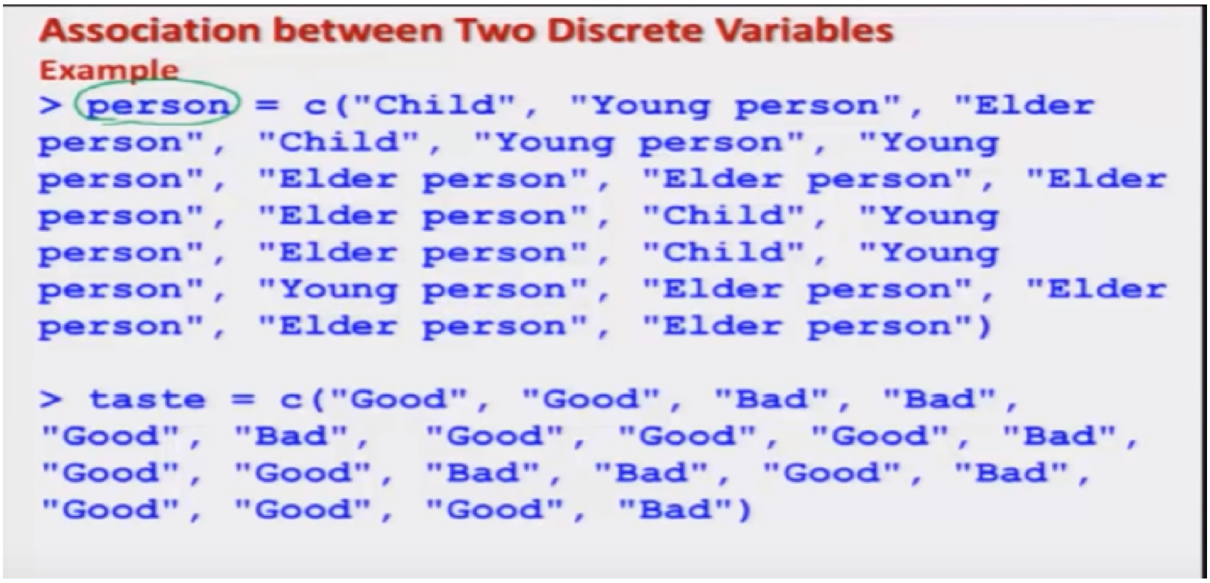
Example

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink
1	Child	Good
2	Young person	Good
3	Elder person	Bad
4	Child	Bad
5	Young person	Good
6	Young person	Bad
7	Elder person	Good
8	Elder person	Good
9	Elder person	Good
10	Elder person	Bad
11	Child	Good
12	Young person	Good
13	Elder person	Bad
14	Child	Bad
15	Young person	Good
16	Young person	Bad
17	Elder person	Good
18	Elder person	Good
19	Elder person	Good
20	Elder person	Bad

So now I try to take a very simple example, and I try to convert the given data into a contingency table and then I would try to obtain the marginal frequencies. Suppose there are twenty persons and they have been divided in three categories with respect to their age as a child young person and elder persons and all of them were given a drink and the taste of the drink was asked? you must note here that this is a similar example which I took in the earlier lecture in the last lecture where I took hundred persons but my objective here is to show you that how the contingency table is constructed from the raw data. So, showing you here hundred observation is more difficult so that is why I am taking here only twenty observations. So, you can see here that there are twenty persons one to ten and then here eleven to twenty and first person is a child, and the child has been asked that how's the drink? and he responds good. Similarly, the second person is a young person who said the drink is good. Similarly, the third person is an elder percent who said that drink is bad, and then the fourth person is a child who said that the drink is bad and so on and this is how we have collected the data on the age and taste here. Right?

Refer Slide Time: (6:30)



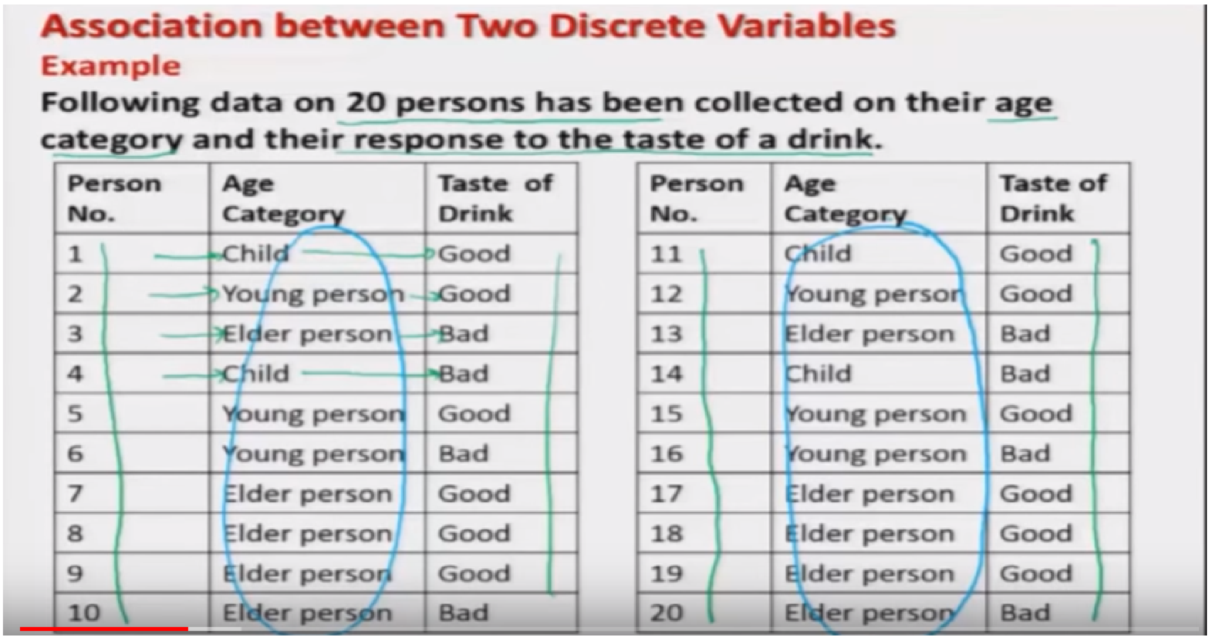
Association between Two Discrete Variables

Example

```
> person = c("Child", "Young person", "Elder
person", "Child", "Young person", "Young
person", "Elder person", "Elder person", "Elder
person", "Elder person", "Child", "Young
person", "Elder person", "Child", "Young
person", "Young person", "Elder person", "Elder
person", "Elder person", "Elder person")

> taste = c("Good", "Good", "Bad", "Bad",
"Good", "Bad",  "Good", "Good", "Good", "Bad",
"Good", "Good", "Bad", "Bad", "Good", "Bad",
"Good", "Good", "Good", "Bad")
```

So now you see I would try to store this data into two data vectors. One is here person in which I would try to store the data set which is here this and here this.

Refer Slide Time: (6:41)



Association between Two Discrete Variables

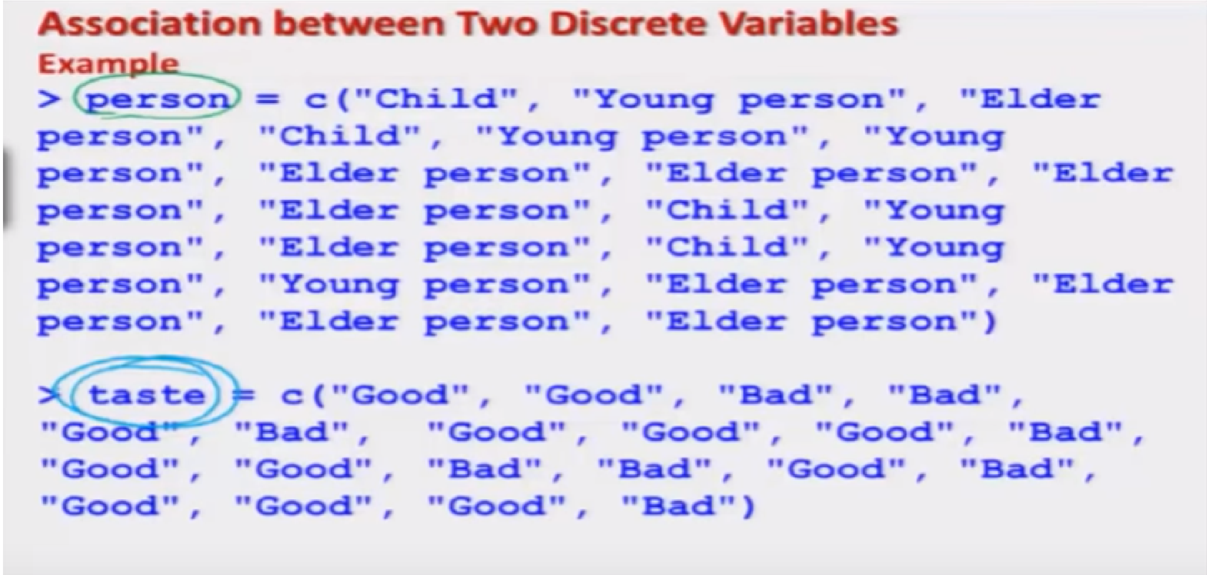
Example

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink	Person No.	Age Category	Taste of Drink
1	Child	Good	11	Child	Good
2	Young person	Good	12	Young person	Good
3	Elder person	Bad	13	Elder person	Bad
4	Child	Bad	14	Child	Bad
5	Young person	Good	15	Young person	Good
6	Young person	Bad	16	Young person	Bad
7	Elder person	Good	17	Elder person	Good
8	Elder person	Good	18	Elder person	Good
9	Elder person	Good	19	Elder person	Good
10	Elder person	Bad	20	Elder person	Bad

So, I I have simply typed it.

Refer Slide Time: (6:48)



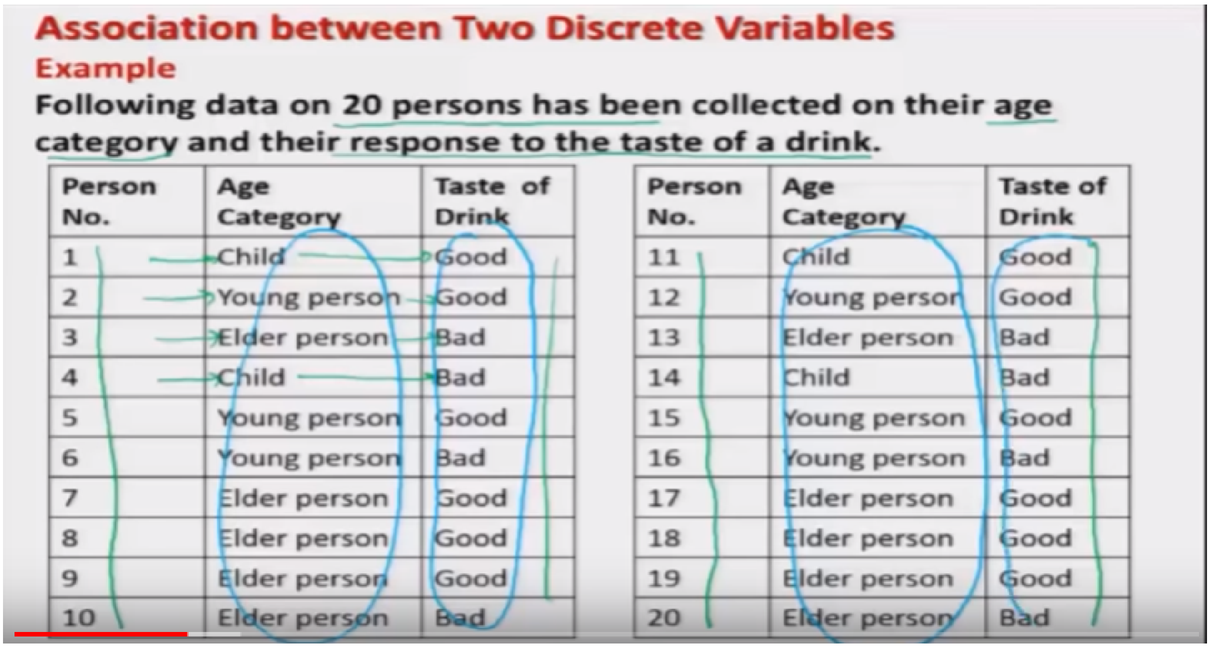
Association between Two Discrete Variables

Example

```
> person = c("Child", "Young person", "Elder
person", "Child", "Young person", "Young
person", "Elder person", "Elder person", "Elder
person", "Elder person", "Child", "Young
person", "Elder person", "Child", "Young
person", "Young person", "Elder person", "Elder
person", "Elder person", "Elder person")

> taste = c("Good", "Good", "Bad", "Bad",
"Good", "Bad",  "Good", "Good", "Good", "Bad",
"Good", "Good", "Bad", "Bad", "Good", "Bad",
"Good", "Good", "Good", "Bad")
```

And then in the second variable which is here taste here I have collected the data,

Refer Slide Time: (6:52)



Association between Two Discrete Variables

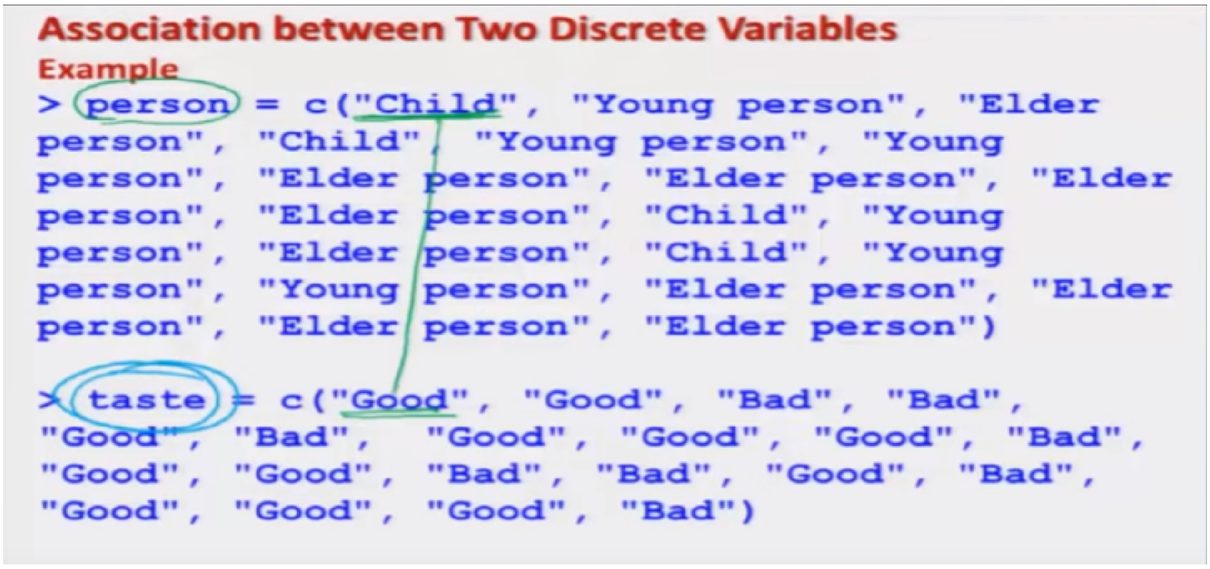
Example

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink	Person No.	Age Category	Taste of Drink
1	Child	Good	11	Child	Good
2	Young person	Good	12	Young person	Good
3	Elder person	Bad	13	Elder person	Bad
4	Child	Bad	14	Child	Bad
5	Young person	Good	15	Young person	Good
6	Young person	Bad	16	Young person	Bad
7	Elder person	Good	17	Elder person	Good
8	Elder person	Good	18	Elder person	Good
9	Elder person	Good	19	Elder person	Good
10	Elder person	Bad	20	Elder person	Bad

On this tale good bad and so on. And this data has been assigned to these vectors in the same order.

Refer Slide Time: (7:00)



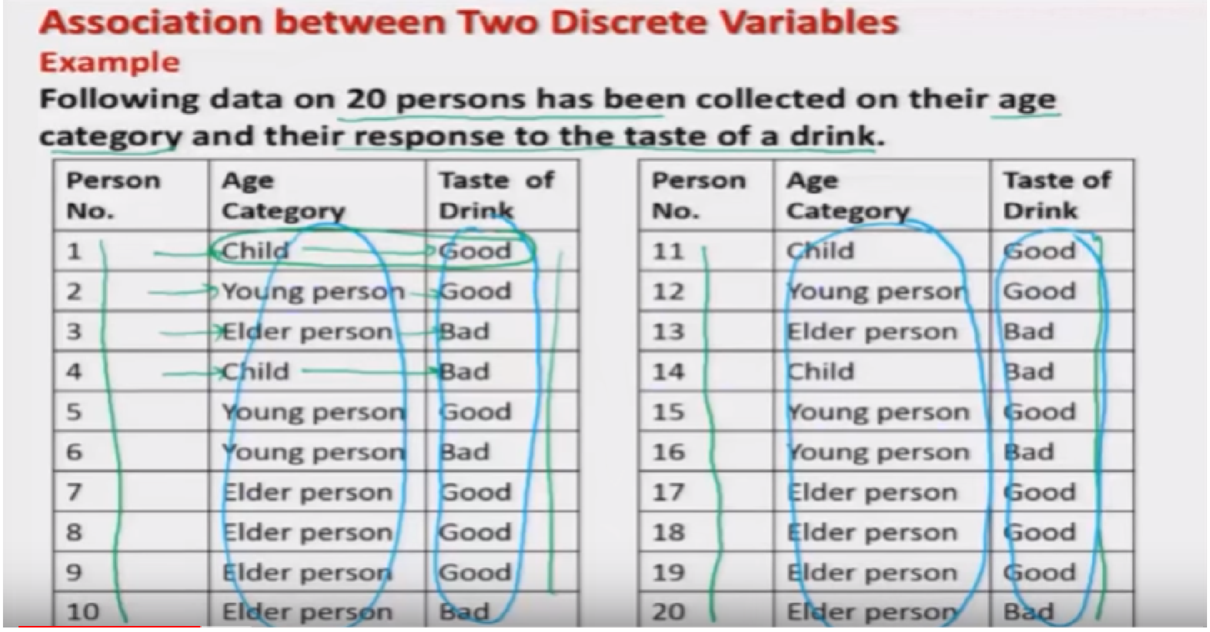
Association between Two Discrete Variables

Example

```
> person = c("Child", "Young person", "Elder
person", "Child", "Young person", "Young
person", "Elder person", "Elder person", "Elder
person", "Elder person", "Child", "Young
person", "Elder person", "Child", "Young
person", "Young person", "Elder person", "Elder
person", "Elder person", "Elder person")

> taste = c("Good", "Good", "Bad", "Bad",
"Good", "Bad", "Good", "Good", "Good", "Bad",
"Good", "Good", "Bad", "Bad", "Good", "Bad",
"Good", "Good", "Good", "Bad")
```

Same order means if you try to say here first person here a child and this child said that the taste is good you can see here and now if you try to see here in the data vector,

Refer Slide Time: (7:13)



Association between Two Discrete Variables

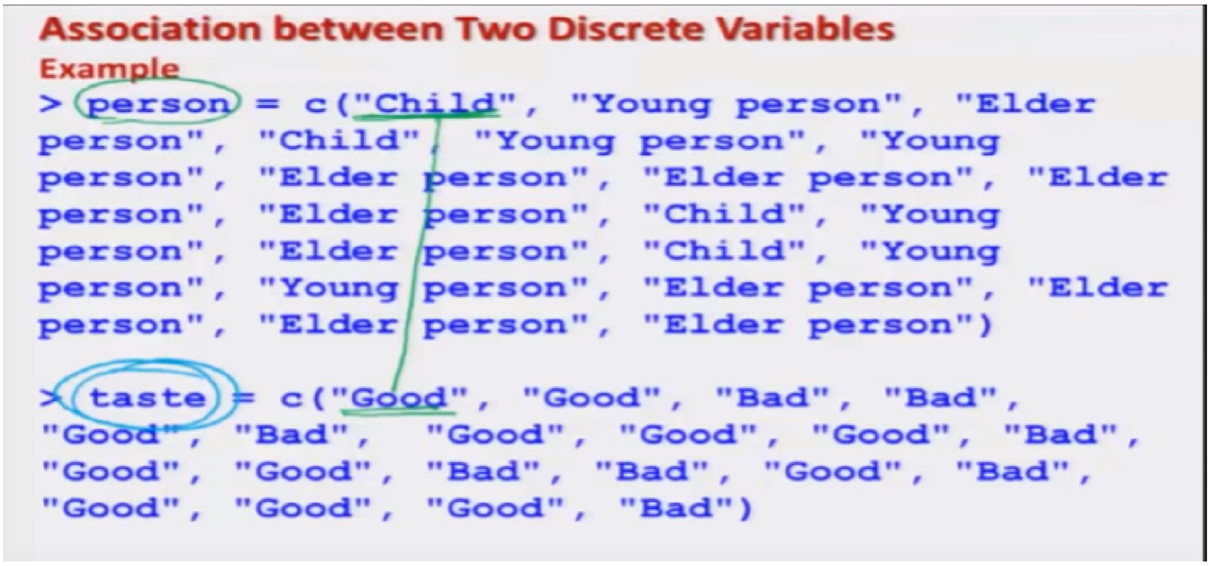
Example

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink
1	Child	Good
2	Young person	Good
3	Elder person	Bad
4	Child	Bad
5	Young person	Good
6	Young person	Bad
7	Elder person	Good
8	Elder person	Good
9	Elder person	Good
10	Elder person	Bad
11	Child	Good
12	Young person	Good
13	Elder person	Bad
14	Child	Bad
15	Young person	Good
16	Young person	Bad
17	Elder person	Good
18	Elder person	Good
19	Elder person	Good
20	Elder person	Bad

This is the thing here child is saying the taste is good and so on.

Refer Slide Time: (7:20)



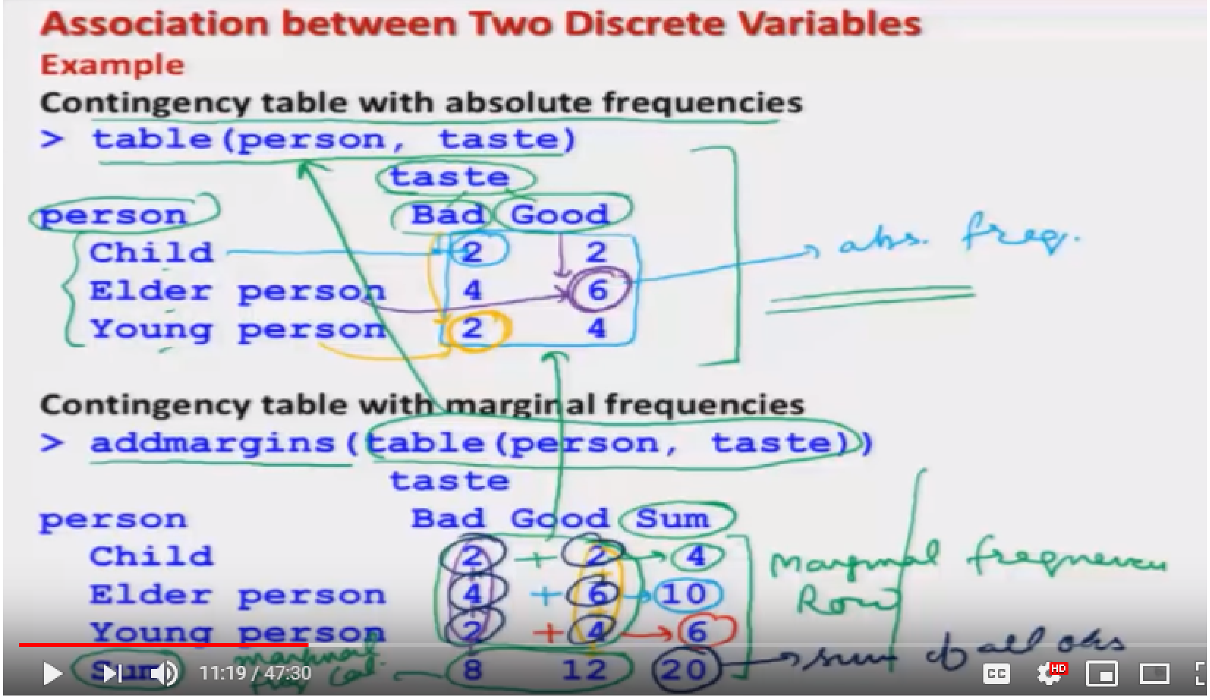
Association between Two Discrete Variables

Example

```
> person = c("Child", "Young person", "Elder
person", "Child", "Young person", "Young
person", "Elder person", "Elder person", "Elder
person", "Elder person", "Child", "Young
person", "Elder person", "Child", "Young
person", "Young person", "Elder person", "Elder
person", "Elder person", "Elder person")

> taste = c("Good", "Good", "Bad", "Bad",
"Good", "Bad", "Good", "Good", "Good", "Bad",
"Good", "Good", "Bad", "Bad", "Good", "Bad",
"Good", "Good", "Good", "Bad")
```

So, these observations are written exactly in the same order as in the table.

Refer Slide Time: (7:25)



Association between Two Discrete Variables

Example

Contingency table with absolute frequencies

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
```

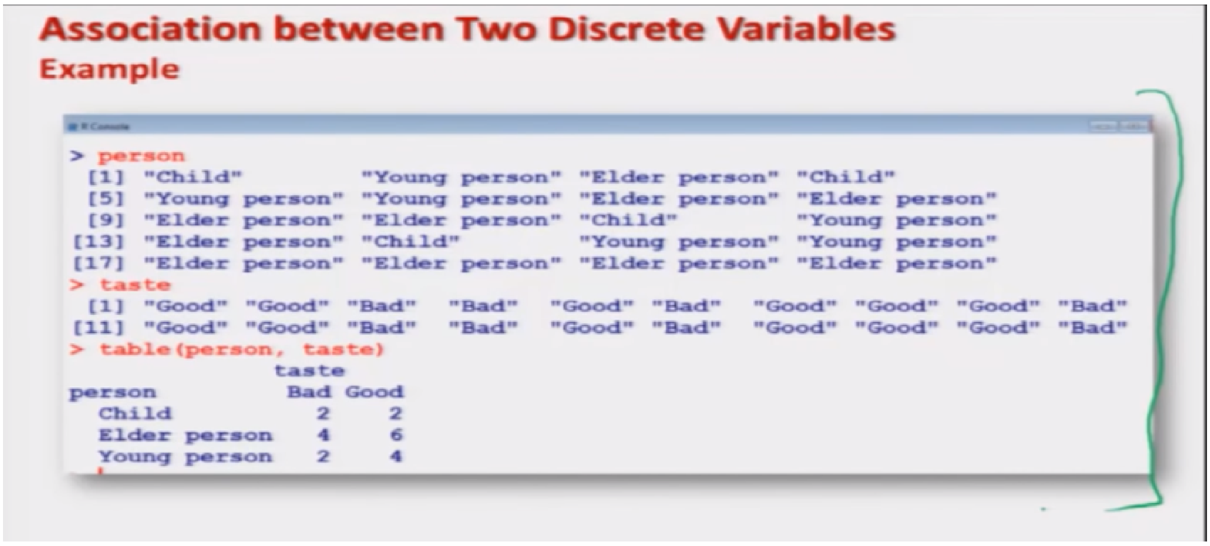
Contingency table with marginal frequencies

```
> addmargins(table(person, taste))
              taste
person         Bad Good Sum
  Child          2    2   4
  Elder person   4    6  10
  Young person   2    4   6
  Sum            8   12  20
```

And now after this I have to use the command table and inside the arguments person separated by comma taste, and this command will provide a contingency table with an absolute frequency and when I try to execute this command on the R console I get here a table like this one. So now how? to interpret this table and how to read this table that is more important to learn, you can see here one where you will here is taste and another variable here is person, and this person has three categories child elder person and young person and these categories are the same what you have denoted in the data, and similarly that taste is also divided into two categories bad and good. And this classification has been done by the R software automatically by Counting that how many persons are in which category. Now this is showing you here for example, if you try to see this is here your contingency table data are the frequencies, so these values are your absolute frequencies.

For example this two is indicating that there are two children which are saying that taste is bad and similarly if you try to see here I will use a different colour pen say here six so six means that there are six elder person who are saying that taste is good, and similarly if you try to take another data here say here 2, so this 2 is indicating that there are two young persons out of 20 who are saying that the taste is bad. Now next we would like to obtain the marginal frequencies so you can see here I am using here the command addmargins and inside the argument I am using the same command which was obtained here to get this contingency table. Now you can see here this contingency table is the same which is obtained here but now there is one more column and one more row which is added in this case here you can see here this is here sum and sum so what is this sum this value here is 4 you can see here this has been obtained by 2 plus 2 is equal to 4, and similarly if you see here 10 this is here 4 plus 6 is equal to 10, and similarly if you see here is this 6 the 6 is here 2 plus 4 is equal to 6, and similarly if you try to take here this first column 2 plus 4 plus 2 this is equal to here 8, and similarly for the second column if you try to add 2 plus 6 plus 4 this is here till. So you can see here this sum is indicating the marginal frequencies, so these are the based on row and similarly here this is the marginal frequencies based on columns, and finally this value here 20, this 20 is the sum of all the observations or sum of all the frequencies its frequencies 2 plus 2 plus 4 plus 6 plus 2 plus 4 which is equal to 20. So, this is how we are going to obtain the table with marginal frequencies.

Refer Slide Time: (11:22)



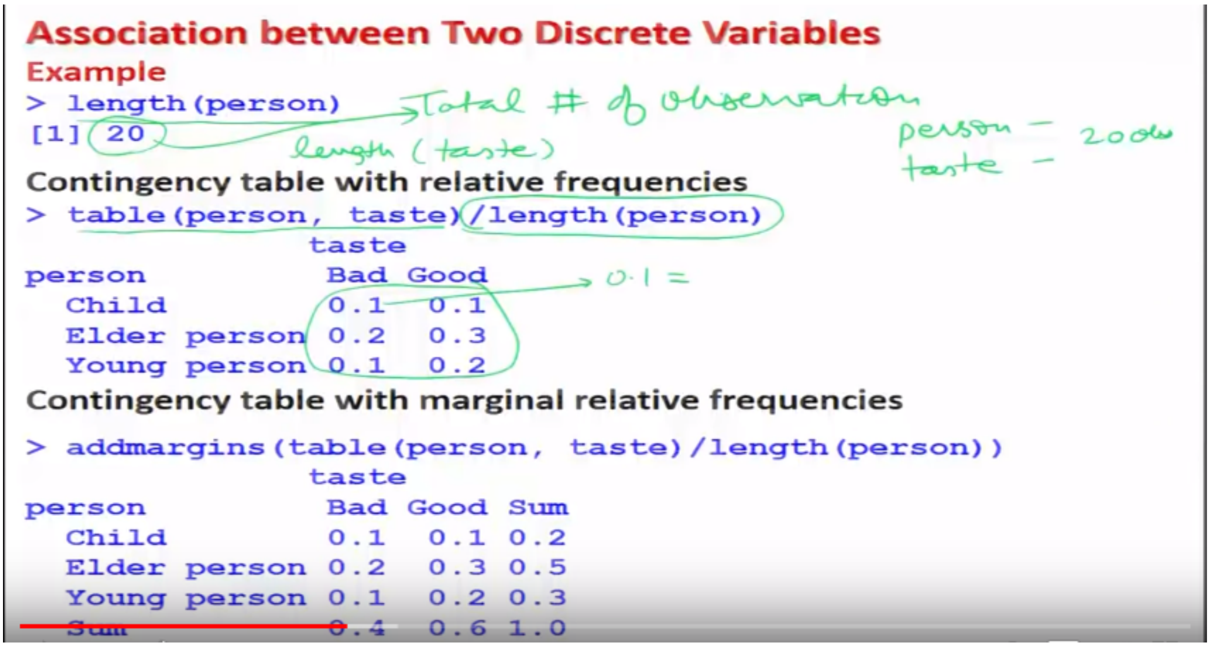
Association between Two Discrete Variables

Example

```
> person
 [1] "Child"          "Young person" "Elder person" "Child"
 [5] "Young person"   "Young person" "Elder person" "Elder person"
 [9] "Elder person"   "Elder person" "Child"        "Young person"
[13] "Elder person"   "Child"        "Young person" "Young person"
[17] "Elder person"   "Elder person" "Elder person" "Elder person"
> taste
 [1] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
[11] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
```

Now this is here the screenshot which I will show you that how I have obtained on the R console,

Refer Slide Time: (11:31)



Association between Two Discrete Variables

Example

```
> length(person)
[1] 20
```

Contingency table with relative frequencies

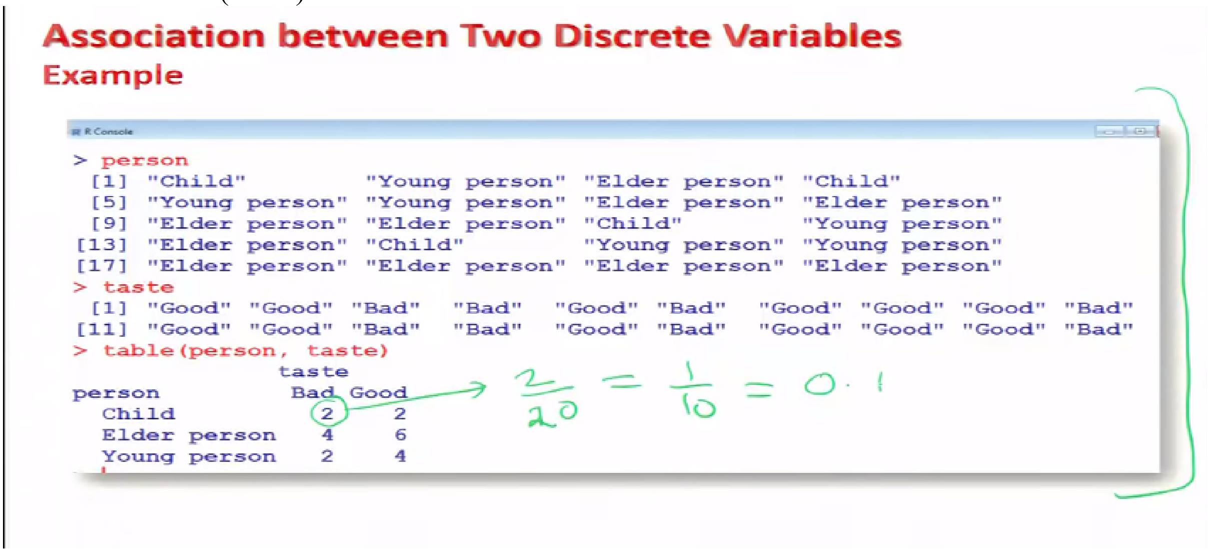
```
> table(person, taste)/length(person)
              taste
person         Bad Good
  Child        0.1  0.1
  Elder person 0.2  0.3
  Young person 0.1  0.2
```

Contingency table with marginal relative frequencies

```
> addmargins(table(person, taste)/length(person))
              taste
person         Bad Good Sum
  Child        0.1  0.1 0.2
  Elder person 0.2  0.3 0.5
  Young person 0.1  0.2 0.3
  Sum          0.4  0.6 1.0
```

Now we try to find out the same thing with respect to the relative frequencies. In order to find out the contingency table with the relative frequencies first we need to find out the, the total number of observations. In order to find out the total number of observations we have two options since we have got here two variables one is here person, and another here is taste and you observe that both these variables have got 20 observation. So now I can use here the context the length of the person or I can also use here length of taste both are going to give us the same values because both the variables have got the same number of observations. So, once I try to operate here the command length of this vector, I will get here a value here 20 which is indicating the total number of observation. Now in order to find out the contingency table I have to use the same command but now I have to divide it by the length of the data vector, so I try to use here the same command table, person taste and now it is divided by length of the data vector, and once you try to do it you will get here an outcome like this one. So firstly, let me try to show you that how you are getting this value so suppose if I take here this value here 0.1 how it is coming if you try to see here,

Refer Slide Time: (13:18)



Association between Two Discrete Variables

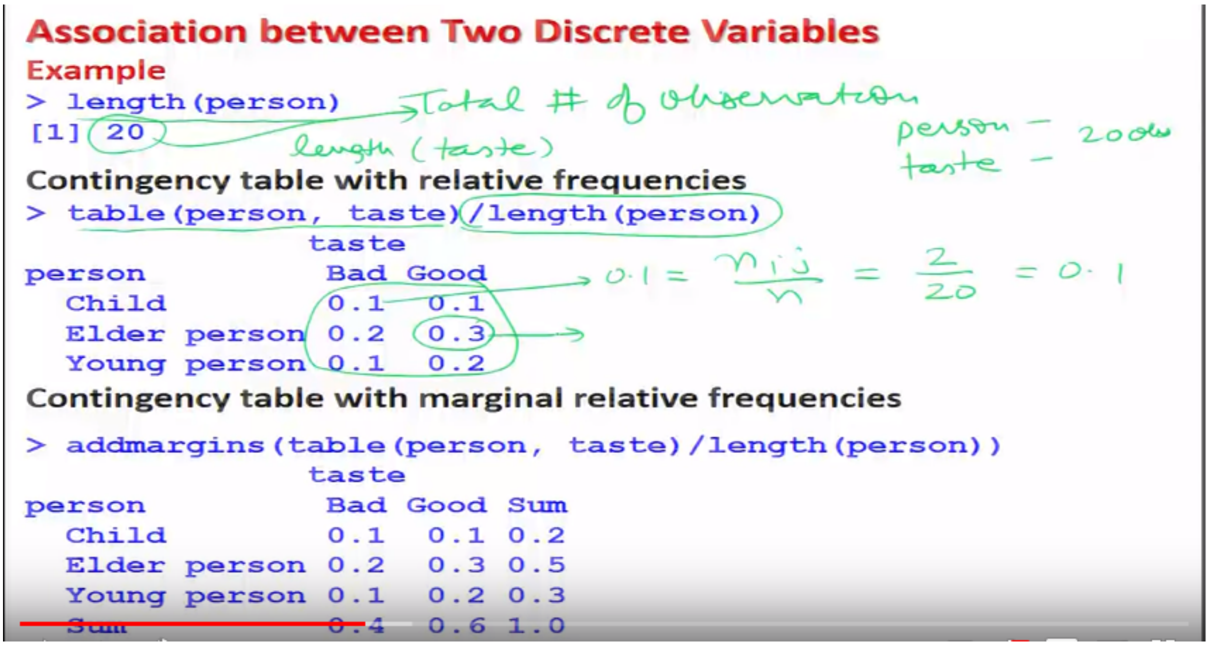
Example

```
> person
 [1] "Child"        "Young person" "Elder person" "Child"
 [5] "Young person" "Young person" "Elder person" "Elder person"
 [9] "Elder person" "Elder person" "Child"        "Young person"
[13] "Elder person" "Child"        "Young person" "Young person"
[17] "Elder person" "Elder person" "Elder person" "Elder person"
> taste
 [1] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
[11] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
```

$\frac{2}{20} = \frac{1}{10} = 0.1$

In the earlier slide we had obtained the frequency here this year two which is corresponding to bad taste and a child. Right? So now this two is being divided by the total number of observations which is here 20 and this will be equal to here 1 upon 10 and which is equal to here 0.1.

Refer Slide Time (13:47)



Association between Two Discrete Variables

Example

```
> length(person)
[1] 20
```

Contingency table with relative frequencies

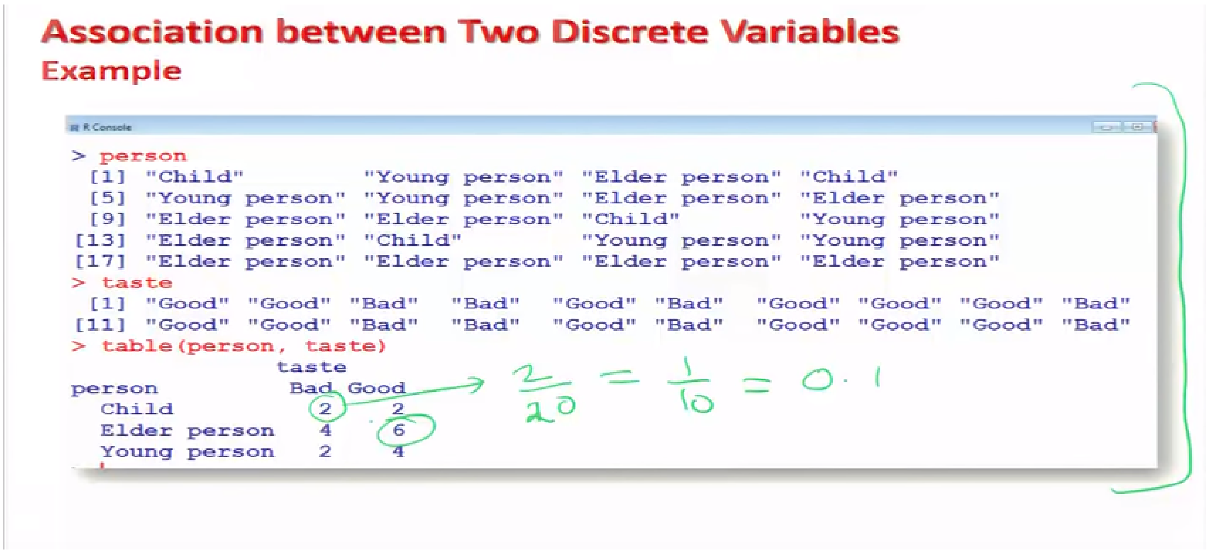
```
> table(person, taste)/length(person)
                 taste
person            Bad Good
   Child          0.1  0.1
   Elder person   0.2  0.3
   Young person   0.1  0.2
```

Contingency table with marginal relative frequencies

```
> addmargins(table(person, taste)/length(person))
                 taste
person            Bad Good Sum
   Child          0.1  0.1 0.2
   Elder person   0.2  0.3 0.5
   Young person   0.1  0.2 0.3
   Sum            0.4  0.6 1.0
```

And this is here the same value, so this is no nothing but your nij upon n which was your h equal to 2 upon here 20 is equal to 0.1, and similarly if you try to find out here how this value has been obtained 0.3

Refer Slide Time: (14:06)



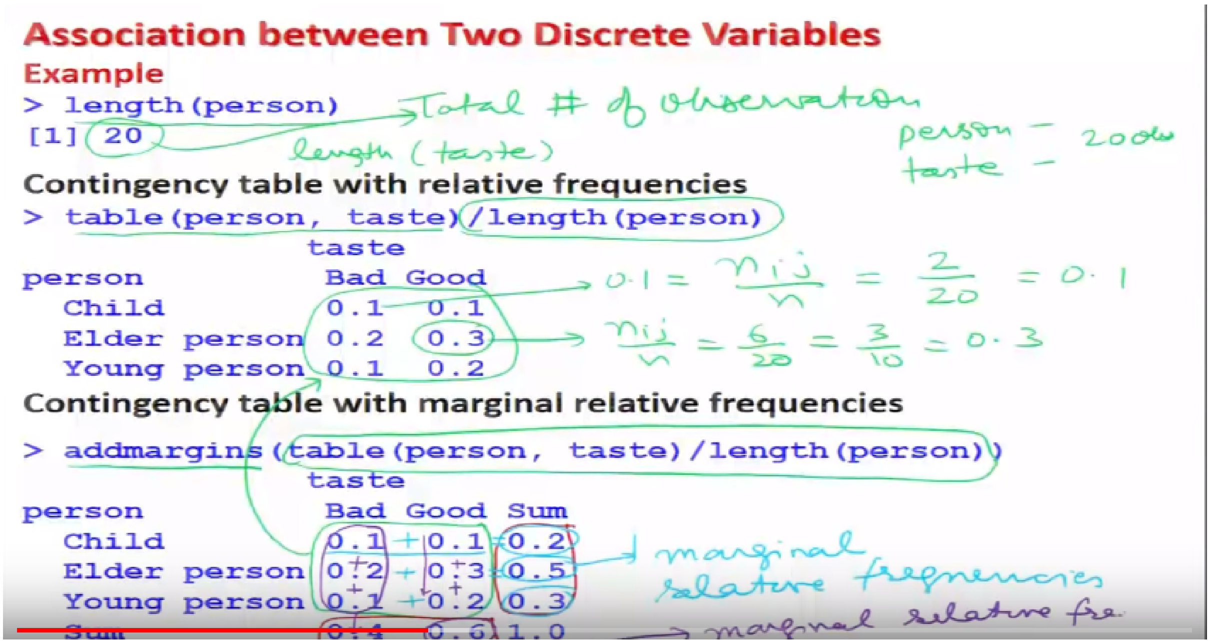
Association between Two Discrete Variables

Example

```
> person
 [1] "Child"        "Young person" "Elder person" "Child"
 [5] "Young person" "Young person" "Elder person" "Elder person"
 [9] "Elder person" "Elder person" "Child"        "Young person"
[13] "Elder person" "Child"        "Young person" "Young person"
[17] "Elder person" "Elder person" "Elder person" "Elder person"
> taste
 [1] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
[11] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good" "Good" "Good" "Bad"
> table(person, taste)
                taste
person           Bad Good
  Child            2    2
  Elder person     4    6
  Young person     2    4
```

So, you can see here the corresponding value here is 6.

Refer Slide Time: (14:10)



Association between Two Discrete Variables

Example

```
> length(person)
[1] 20
```

Contingency table with relative frequencies

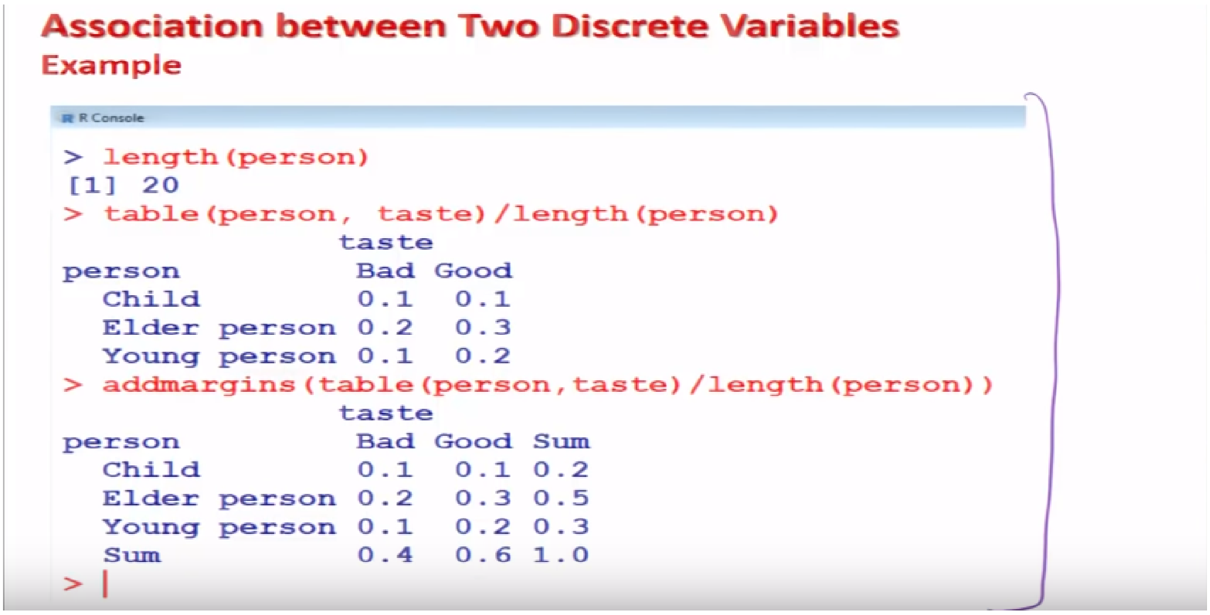
```
> table(person, taste)/length(person)
              taste
person         Bad Good
  Child        0.1  0.1
  Elder person 0.2  0.3
  Young person 0.1  0.2
```

Contingency table with marginal relative frequencies

```
> addmargins(table(person, taste)/length(person))
              taste
person         Bad Good Sum
  Child        0.1  0.1 0.2
  Elder person 0.2  0.3 0.5
  Young person 0.1  0.2 0.3
  Sum          0.4  0.6 1.0
```

So, I try to use here say here nij upon here n which is equal to here 6 upon 20 and this comes out to be 8 could be a 3 upon 10 which is equal to here 0.3. So, this is how all other values in this table are obtained. Now in case if I also want to find out the marginal relative frequencies so I have to use here the command addmargins and I have to use the same command which I you have used to find out the contingency table, and in case if you try to do it you can see here this part here, this is the same as this part here because this is corresponding to the contingency table, and now there is additional row and column here which are here like this this and here this. So the first question comes what are these additional rows and columns are indicating so I will try to show you here that suppose if I try to take here the first row so for the sum of first row which is 0.1 plus 0.1 is equal to zero point two and this is indicated here in the first value in this column. Similarly, if you try to add here 0.2 plus 0.3 which is equal to here 0.5 so this is the second value, and similarly the third value here is 0.1 plus 0.2 which is equal to here 0.3. So, they give us the values of marginal relative frequencies. Similarly if you try to look into the columns so if I try to sum here these things so 0.1 plus 0.2 plus 0.1 this is being given here as 0.4, and similarly in the second column 0.1 plus 0.3 plus 0.2 in this direction this is giving us the value 0.6. Right? So, these two values here 0.4 and 0.6 they are also the marginal relative frequencies.

Refer Slide Time: (16:40)



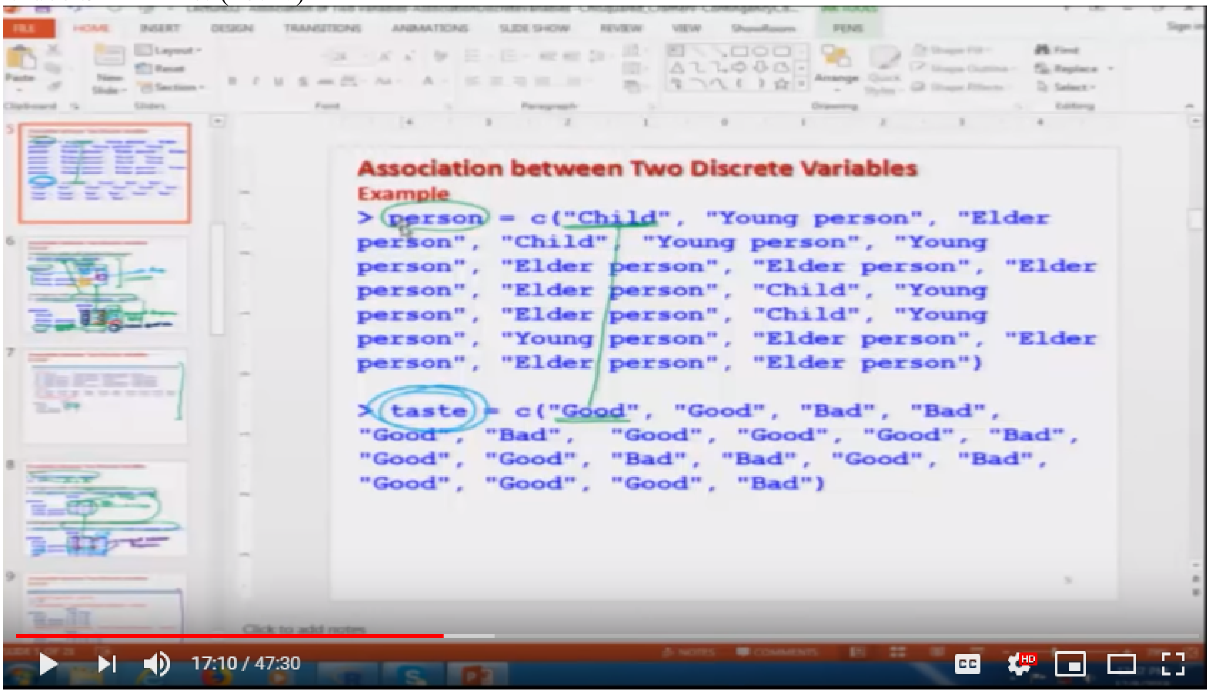
Association between Two Discrete Variables

Example

```
> length(person)
[1] 20
> table(person, taste)/length(person)
                taste
person           Bad Good
  Child          0.1  0.1
  Elder person   0.2  0.3
  Young person   0.1  0.2
> addmargins(table(person,taste)/length(person))
                taste
person           Bad Good Sum
  Child          0.1  0.1 0.2
  Elder person   0.2  0.3 0.5
  Young person   0.1  0.2 0.3
  Sum            0.4  0.6 1.0
> |
```

And this is here the screenshot of the operations which are to be done on the R console. Now if we're going further let me try to show you all these operations on the R console, so I will try to take the same example and I will try to enter the same data set, and I will try to obtain the contingency tables with respect to the absolute frequencies as well as with respect to the relative frequencies.

Video Start Time: (17:10)



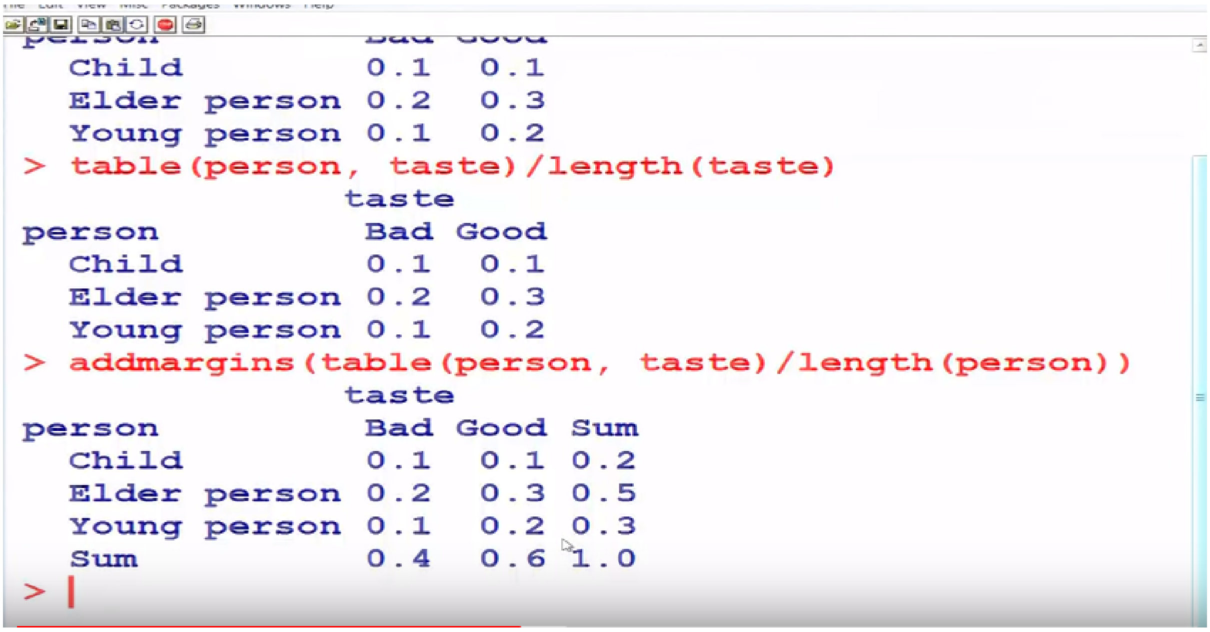
Association between Two Discrete Variables

Example

```
> person = c("Child", "Young person", "Elder
person", "Child", "Young person", "Young
person", "Elder person", "Elder person", "Elder
person", "Elder person", "Child", "Young
person", "Elder person", "Child", "Young
person", "Young person", "Elder person", "Elder
person", "Elder person", "Elder person")

> taste = c("Good", "Good", "Bad", "Bad",
"Good", "Bad", "Good", "Good", "Good", "Bad",
"Good", "Good", "Bad", "Bad", "Good", "Bad",
"Good", "Good", "Good", "Bad")
```

So you can see here first I will try to create these two data vectors here on the R console and similarly I try to create this data vector taste so you can see here now I have the data on person, and here taste. Right? Okay? I clear the screen and I try to now create here the contingency table so you can see here this is the person and here taste this is the same table what you have obtained. Now I just want to

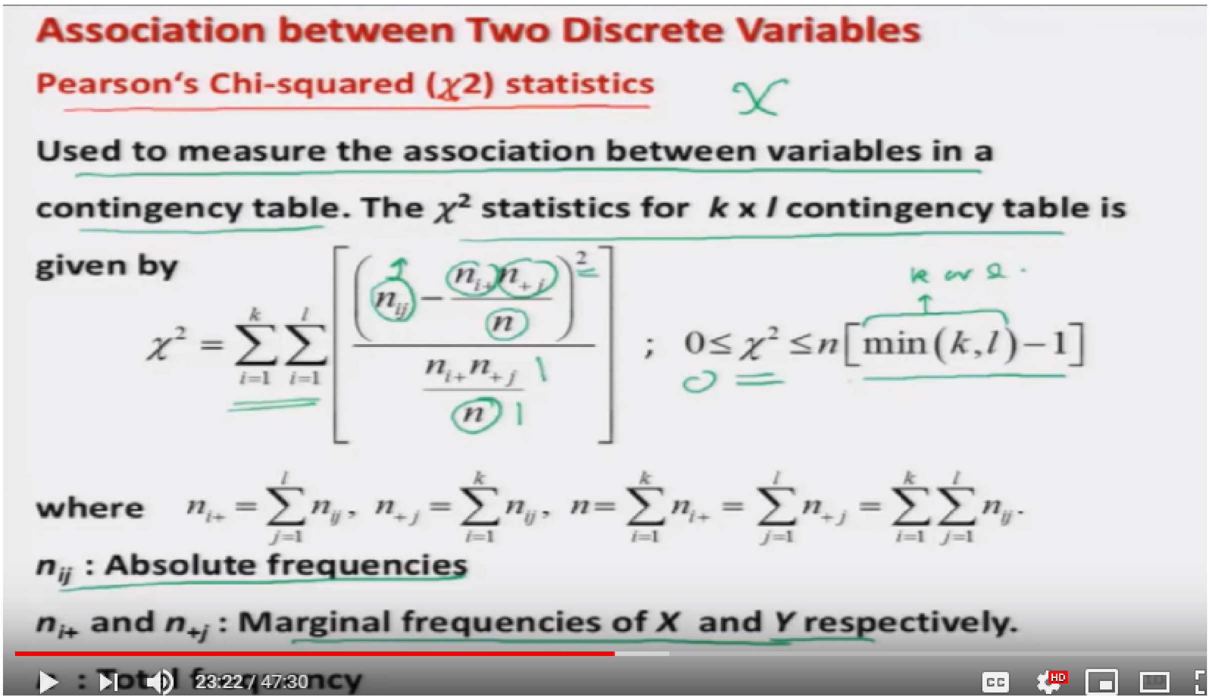
show you that if you try to interchange person and tastes so first, I try to give inside the argument taste and
then person you see what happens. So obviously the data will remain the same but only the rows and columns are interchange that you can observe here. So, in case if I want to find out the marginal frequency of anyone say person comma taste you can see here this is obtained here like this, so this is the same thing what we have just obtained. Now I try to find out the same contingency table with respect to the relative frequencies so you can see here I'm trying to take here the same command but now I'm trying to divide it by the length of the data vector and I'm choosing here the data vector to be person and you can see here that we have got here this type of command and these are the values which are the same which we have shown you on the slides, and now in case if I want to find out the same contingency table with respect to the length of another data vector so I try to choose here the data vector taste inside of here person and you will see here that in both the cases you are going to get the same command because length of the two data vectors here is the same which is here 20. Right? and now in case if I want to add the marginal relative frequencies so I have to use the same command but I have to add here one command addmargins so you can see here that this gives us this value. Right? So, you can see here now the sum of all the marginal relative frequencies is coming out to be here one. Right? and this is the same output which I had shown you on the screens so you can see here that finding out search contingency table with respect to absolute frequencies or relative frequencies is not difficult once you have some data set.
Video End Time: (20:16)



```
person              Bad Good
   Child            0.1  0.1
   Elder person     0.2  0.3
   Young person     0.1  0.2
> table(person, taste)/length(taste)
                    taste
person              Bad Good
   Child            0.1  0.1
   Elder person     0.2  0.3
   Young person     0.1  0.2
> addmargins(table(person, taste)/length(person))
                    taste
person              Bad Good Sum
   Child            0.1  0.1 0.2
   Elder person     0.2  0.3 0.5
   Young person     0.1  0.2 0.3
   Sum              0.4  0.6 1.0
> |
```

And now I would try to discuss one more new topic. So now I'm going to discuss a tool which is called as chi-square statistics and the role of this chi-square statistic is that it tries to give us an idea by quantifying the degree of Association, similar to what we had in the case of continuous variable we had correlation coefficient. Right? So one thing you have to keep in mind that when I am going to introduce the chi-square statistics actually this chi-square statistics is used for testing of hypothesis and chi-square test is based on a probability density function which is called as chi-square probability density function, and when we try to use this statistic there are certain conditions in the case of test of hypothesis. For example the cell frequency should be greater than 5 and so on but here you see I am taking an artificial example so in this example means I have kept the frequencies to be low means if you have more data then obviously these frequencies are going to be higher, so while computing the statistics on the R software you may get sort of warning but you need not to worry for these things you have to follow essentially the procedure and the concept.

Refer Slide Time (21:48)



Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics χ

Used to measure the association between variables in a contingency table. The χ^2 statistics for k x l contingency table is given by

$$\chi^2 = \sum_{i=1}^{k}\sum_{i=1}^{l}\left[\frac{\left(n_{ij} - \frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}\right] \quad ; \quad 0 \le \chi^2 \le n\left[\overbrace{\min(k,l)}^{k \text{ or } l\,\cdot} - 1\right]$$

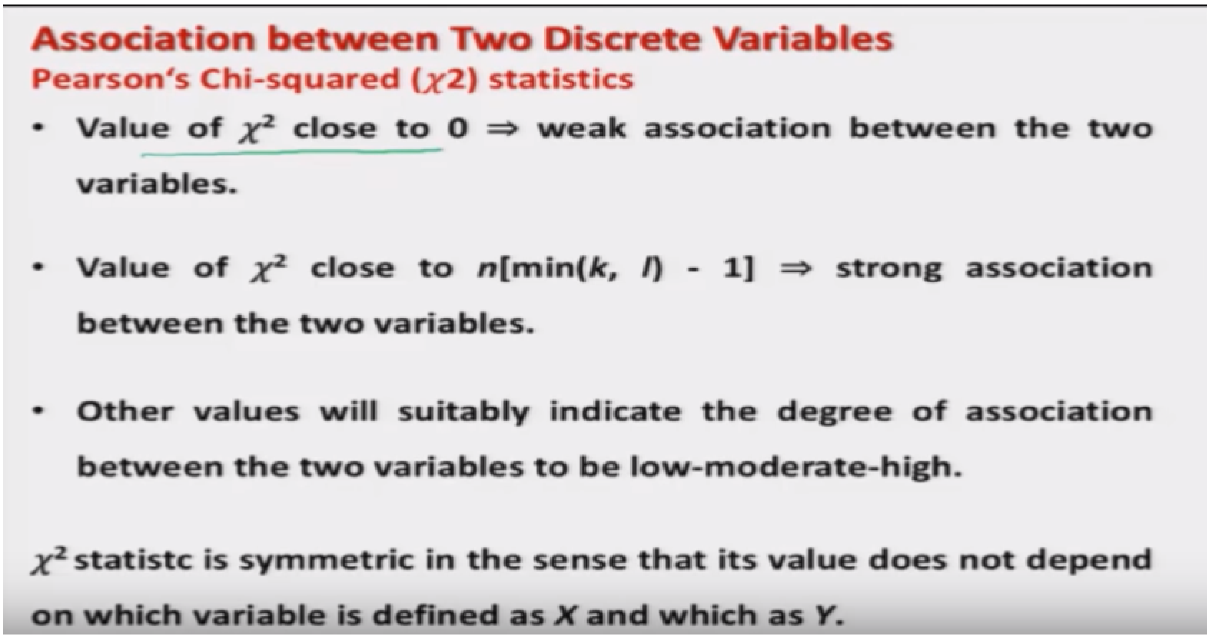
where $n_{i+} = \sum_{j=1}^{l} n_{ij}$, $n_{+j} = \sum_{i=1}^{k} n_{ij}$, $n = \sum_{i=1}^{k} n_{i+} = \sum_{j=1}^{l} n_{+j} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}$.

n_{ij} : Absolute frequencies

n_{i+} and n_{+j} : Marginal frequencies of X and Y respectively.

So this chi- squared statistics or this is called as Pearson's chi squared the statistics, this symbol here chi this is a Greek letter which is written here like this Chi and this statistics chi-squared statistic is used to measure the association between the variables in our contingency table suppose there are two variables so the chi-square statistics for a K cross L contingency table what we have created earlier is given by this quantity you can see here this is the summation over all the observations and in the numerator this is nij minus ni plus n plus j divided by here n whole square and divided by a similar quantity ni plus into n plus j divided by here n. So, you can recall that this nij is giving you the value of absolute frequency and this ni plus and n plus J they are the marginal frequencies of X and Y and small n here is the total number of observation or the total frequency. So this is the statistics which gives us an idea about the degree of Association and this value of chi-square lies between 0 and n into minimum value between k and l minus 1 minimum value k and l means whatever is the minimum value out of k and l they this can be k or l whose ever is minimum.

Refer Slide Time (23:24)



Association between Two Discrete Variables

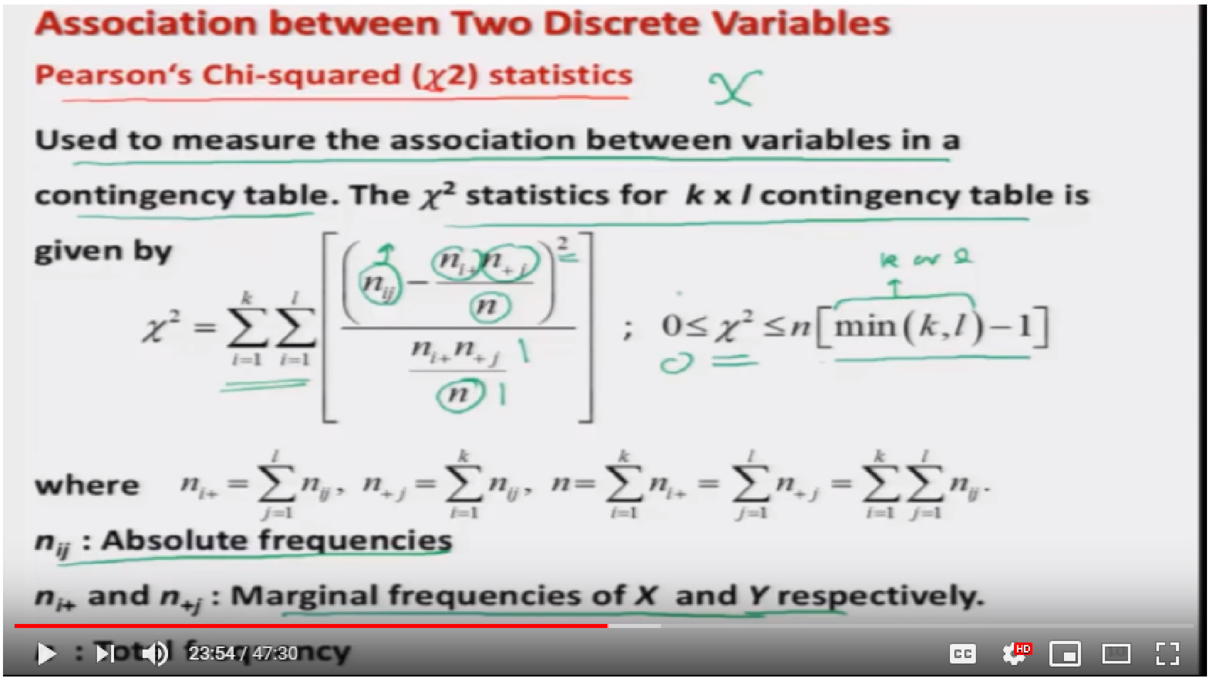
Pearson's Chi-squared ($\chi 2$) statistics

- Value of χ^2 close to 0 ⇒ weak association between the two variables.
- Value of χ^2 close to $n[\min(k, l) - 1]$ ⇒ strong association between the two variables.
- Other values will suitably indicate the degree of association between the two variables to be low-moderate-high.

χ^2 statistc is symmetric in the sense that its value does not depend on which variable is defined as X and which as Y.

Now what is the interpretation of this statistics now given the data one can compute this statistics now in case if the values of chi-square are coming close to zero or the value of chi-square is very close to zero this will indicate that the association between the two variable is weak. so, a value of chi-square close to zero indicates a weak association between the two variables x and y and similarly you can see here,

Refer Slide Time: (23:53)



Association between Two Discrete Variables

Pearson's Chi-squared (χ2) statistics

Used to measure the association between variables in a contingency table. The χ^2 statistics for $k \times l$ contingency table is given by

$$\chi^2=\sum_{i=1}^{k}\sum_{i=1}^{l}\left[\frac{\left(n_{ij}-\frac{n_{i+}n_{+j}}{n}\right)^2}{\frac{n_{i+}n_{+j}}{n}}\right] \quad ; \quad 0\le\chi^2\le n\left[\min(k,l)-1\right]$$

where $n_{i+}=\sum_{j=1}^{l}n_{ij},\ n_{+j}=\sum_{i=1}^{k}n_{ij},\ n=\sum_{i=1}^{k}n_{i+}=\sum_{j=1}^{l}n_{+j}=\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}.$

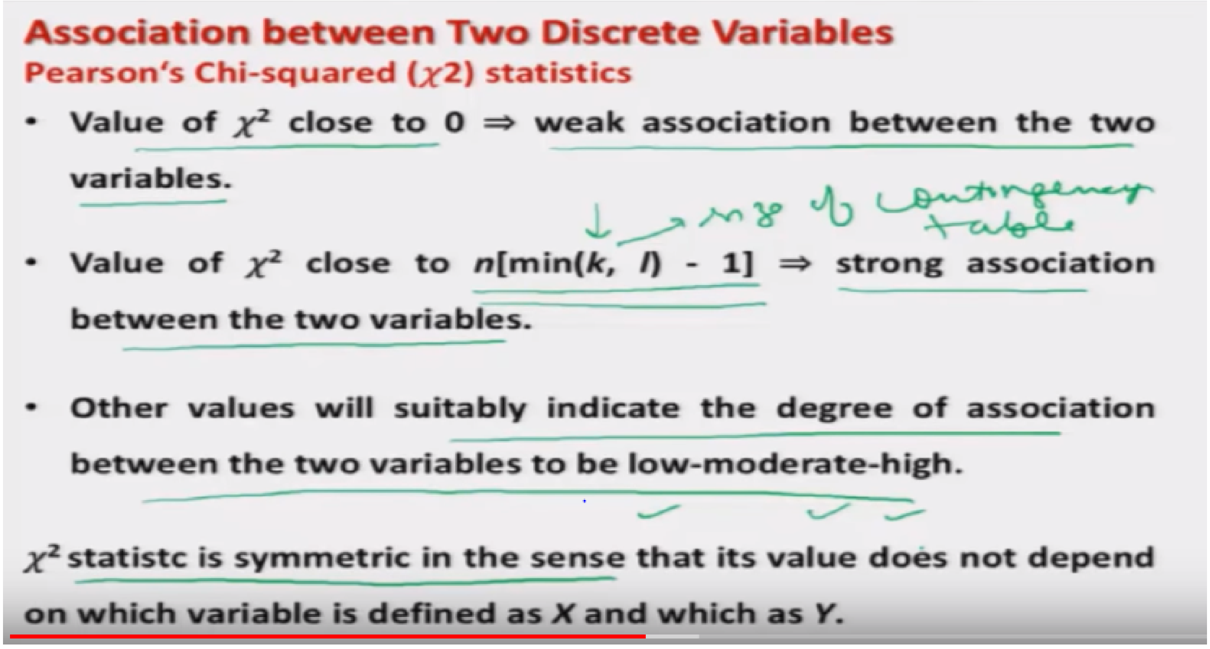
n_{ij} : Absolute frequencies

n_{i+} and n_{+j} : Marginal frequencies of X and Y respectively.

: Total frequency

that the chi-square has the limits zero and here this minimum k, l minus 1 into n. so, in case if the value of

Refer Slide Time: (24:01)



Association between Two Discrete Variables

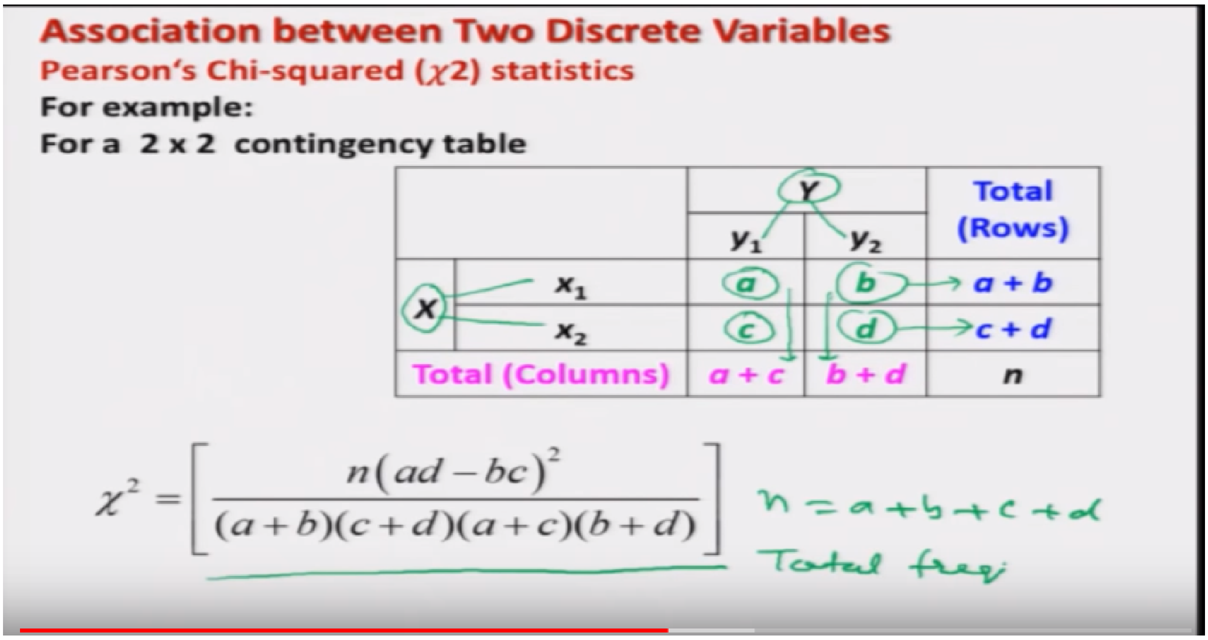
Pearson's Chi-squared (χ2) statistics

- Value of χ^2 close to 0 ⇒ weak association between the two variables.
- Value of χ^2 close to $n[\min(k, l) - 1]$ ⇒ strong association between the two variables.
- Other values will suitably indicate the degree of association between the two variables to be low-moderate-high.

χ^2 statistc is symmetric in the sense that its value does not depend on which variable is defined as X and which as Y.

chi-square is close to the second limit which is n into minimum of k, l, l minus one then this would indicate a strong association between the two variables remember this thing this is not between like zero or one or minus one or one something like this. So, this value depends on the size of the contingency table, size size of the contingency table well that's a drawback and that was overcome in the, in some further modification that we are going to discuss. Now, in case if you are getting any other value of chi-square which is not close to zero or this interval and into minimum of k, l minus one then suitably that will indicate the degree of association between the two variables say as low moderate or say high in general. One aspect of this chi-squared statics is that it is symmetric, symmetric means that which of the variables you are taking in the row or say column this will not make any difference for example, you had seen that we had constructed the frequency table will persons and taste and taste and person both the cases this Chi-square statistics will remain the same.

Refer Slide Time: (25:18)



Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics

For example:

For a 2 x 2 contingency table

		Y		Total (Rows)
		y_1	y_2	
X	x_1	a	b	a + b
	x_2	c	d	c + d
Total (Columns)		a + c	b + d	n

$$\chi^2 = \left[\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \right]$$

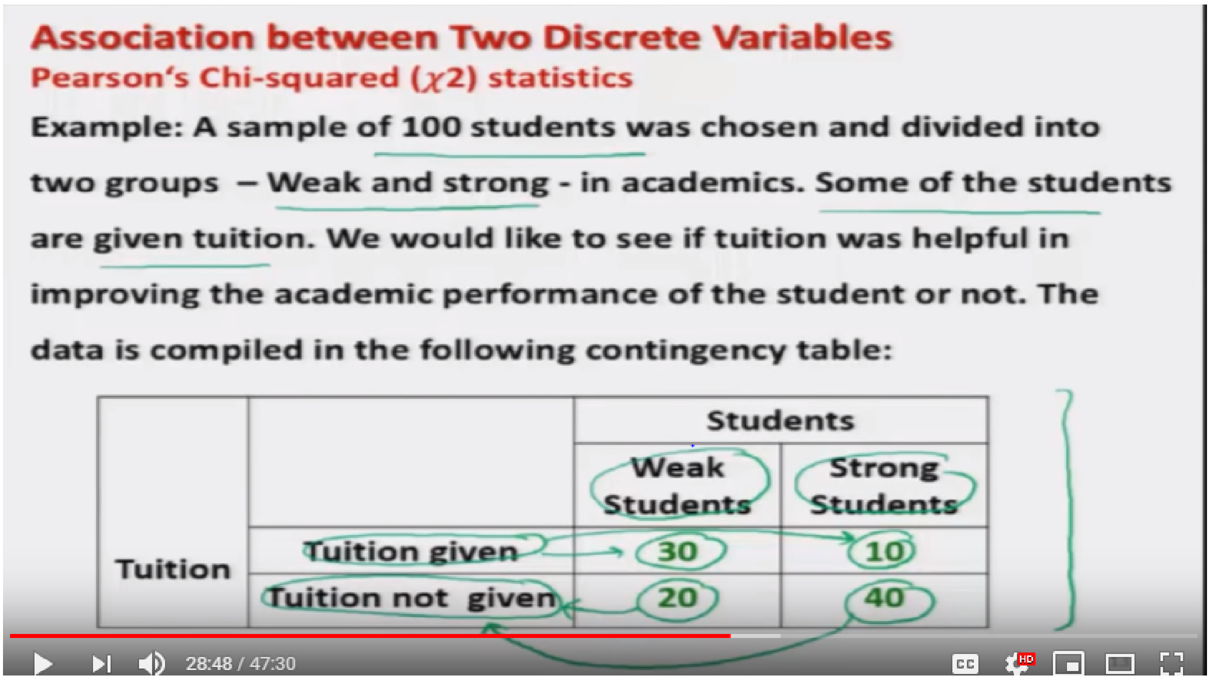
n = a+b+c+d

Total freq;

\

Now, you look at me just give a particular example which is more popular, say in case if I have only two by two contingency table. So, suppose there are a variable X and Y which have got two classes each x_1, x_2 and y_1, y_2 and absolute frequencies in these cells are a, b, c and d which I have written in green colour and now, if you try to find out the sum of rows if it is the marginal frequency this will come out to a plus b and similarly the marginal frequency of this second row corresponding to x two is c plus d and similarly the modular frequency with respect to the first column is a plus c and the modular frequency with respect to the second column y two is b plus d in this case if you try to substitute all the values inside the chi-square statistics this will simplify to this thing. where here obviously n is equal to a plus b plus c plus d which is indicating the total frequency or Right! Okay. Now, after this what I will do is the following that I will take a simple example being on this two cross two data and I will try to compute this chi-square statistic and later on I will introduce some more statistics and then I would try to measure the degree of Association in the same example using different statistics. So, this data is about that a sample of hundred student, the students was collected and they were judged whether they are weak in academics or good in academics, they are good in your studies or bad in your studies based on their I can make performance now after this a group of student was given a tuition and after the tuition they were just once again and the idea was that we wanted to know whether this tuition is going to be helpful or not, whether this tuition means extra studies are really helping the students and improving their academic performance? That is the question which I would like to know on the basis of given sample of data. So, what we have done here?

Refer Slide Time: (27:44)



Association between Two Discrete Variables

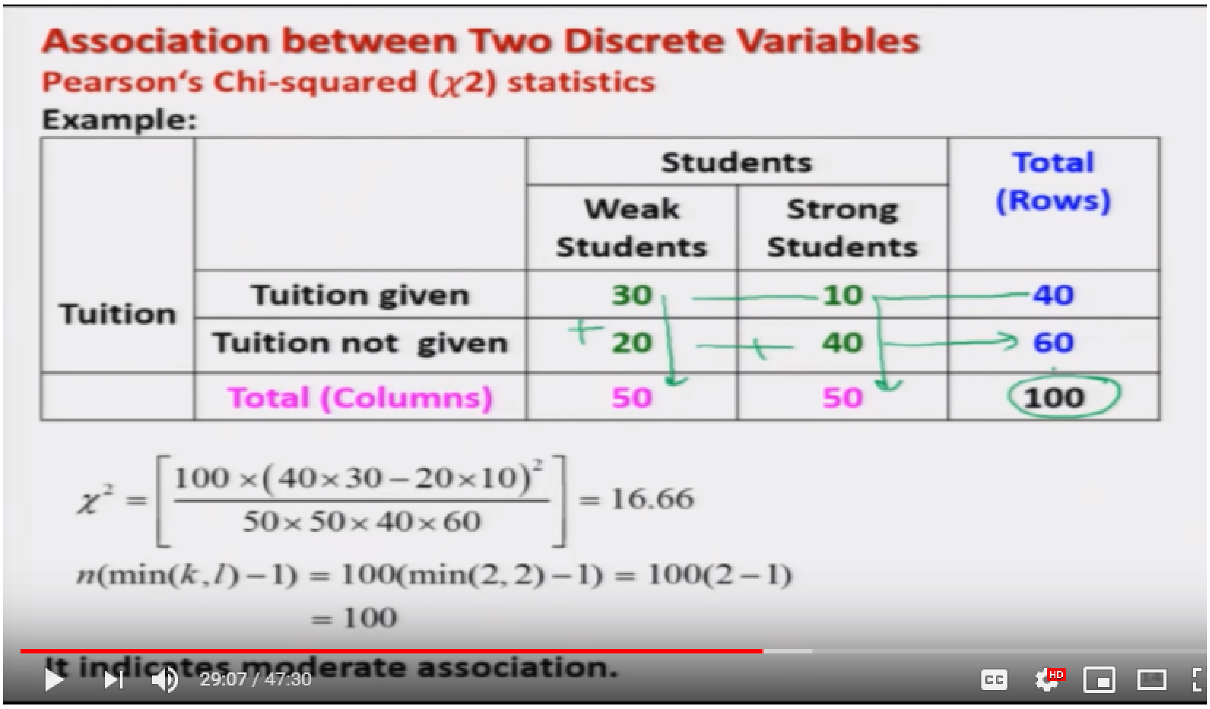
Pearson's Chi-squared ($\chi 2$) statistics

Example: A sample of 100 students was chosen and divided into two groups – Weak and strong - in academics. Some of the students are given tuition. We would like to see if tuition was helpful in improving the academic performance of the student or not. The data is compiled in the following contingency table:

		Students	
		Weak Students	Strong Students
Tuition	Tuition given	30	10
	Tuition not given	20	40

That this sample of handed student is divided into two groups, weak and strong in academics and some of the students from say weak and strong both, they are given tuition and after that their academic performance was just and after this that data was compiled in the following contingency table here. So, you can see here there are weaker students and there are strongest students and students who were given the tuition hours, they were not given that tuition. So, it was found finally that there are 30 weak students who were given the tuition and there were 10 strong students, 10 good students who were also given the tuition. Similarly, there were 20 weaker students who were not given the tuition and there were 40 strong students who were not given that tuition. and based on that we would like to find whether there is any association between tuition and the academic strength of the students.

Refer Slide Time: (28:50)



Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics

Example:

		Students		Total (Rows)
		Weak Students	Strong Students	
Tuition	Tuition given	30	10	40
	Tuition not given	20	40	60
	Total (Columns)	50	50	100

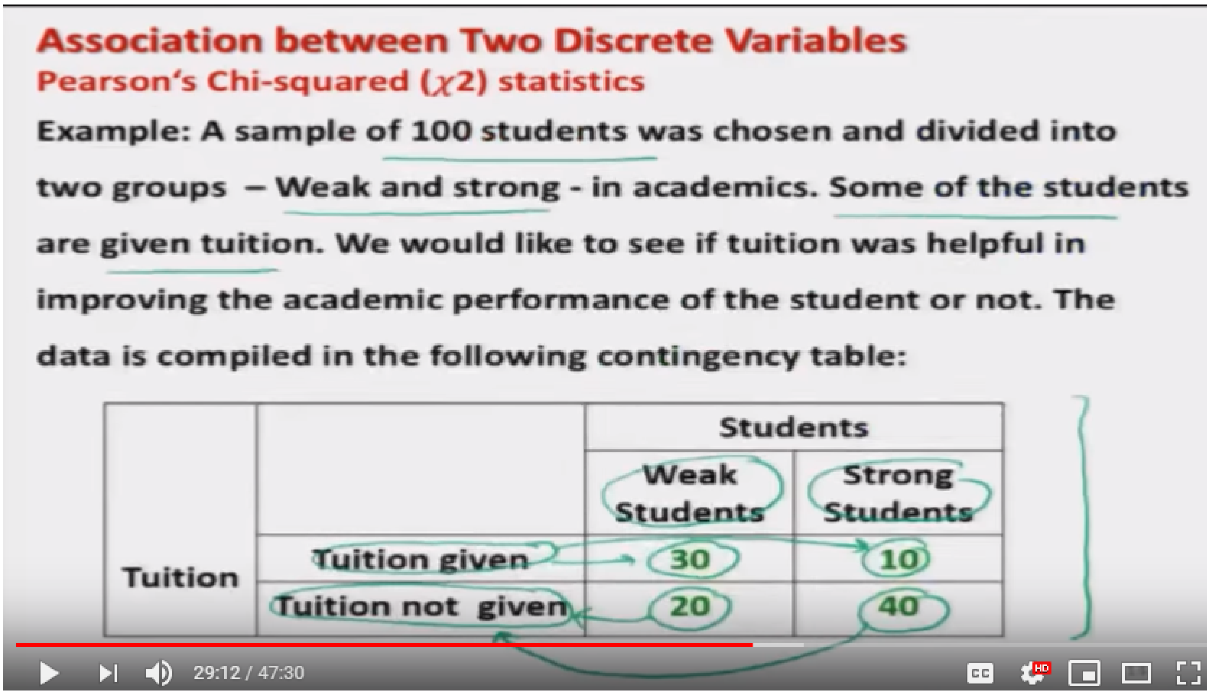
$$\chi^2 = \left[\frac{100\times(40\times30-20\times10)^2}{50\times50\times40\times60}\right] = 16.66$$

$$n(\min(k,l)-1) = 100(\min(2,2)-1) = 100(2-1)$$
$$= 100$$

It indicates moderate association.

So, we try to find out their modular frequency. So, you can see here modular frequencies are here 30 plus 10 is equal to 40, 20 plus 40 is equal to 60, and similarly at 30 plus 20 is equal to 50, and 10 plus 40 is equal to 50 and the total sum here is hundred. Based on that I try to substitute all these values of frequency's in terms of a, b, c, d.

Refer Slide Time: (29:12)



Association between Two Discrete Variables

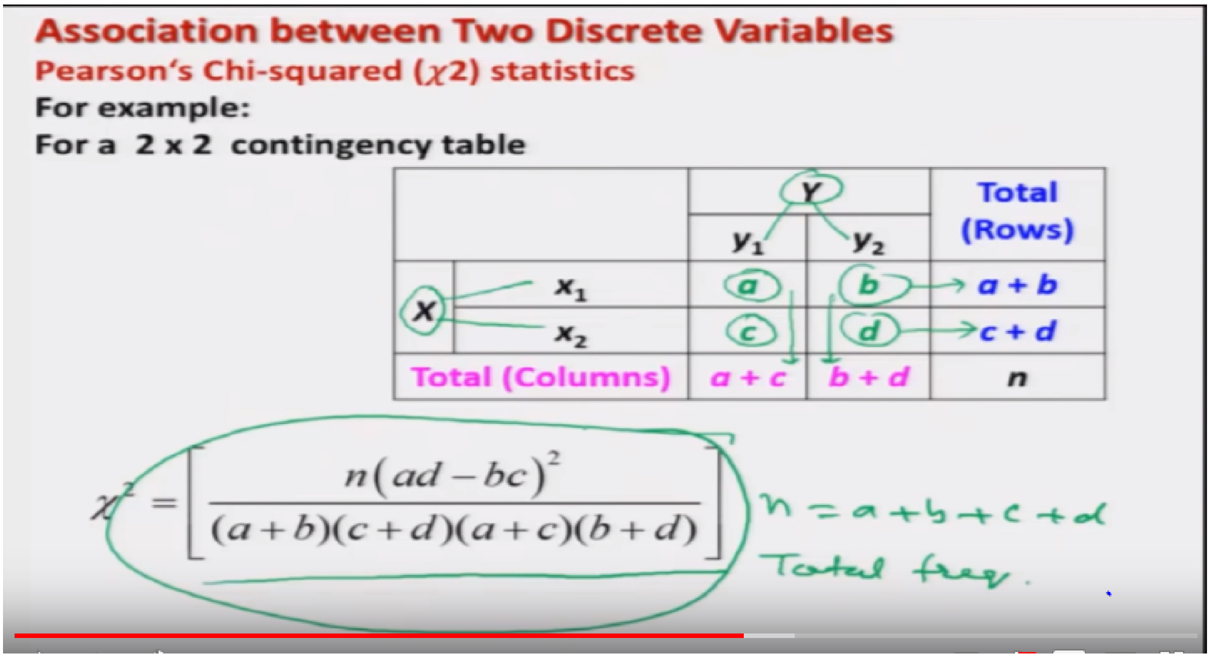
Pearson's Chi-squared ($\chi 2$) statistics

Example: A sample of 100 students was chosen and divided into two groups – Weak and strong - in academics. Some of the students are given tuition. We would like to see if tuition was helpful in improving the academic performance of the student or not. The data is compiled in the following contingency table:

		Students	
		Weak Students	Strong Students
Tuition	Tuition given	30	10
	Tuition not given	20	40

what we have done here

Refer Slide Time: (29:12)



Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics

For example:

For a 2 x 2 contingency table

		Y		Total (Rows)
		y_1	y_2	
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
Total (Columns)		$a+c$	$b+d$	n

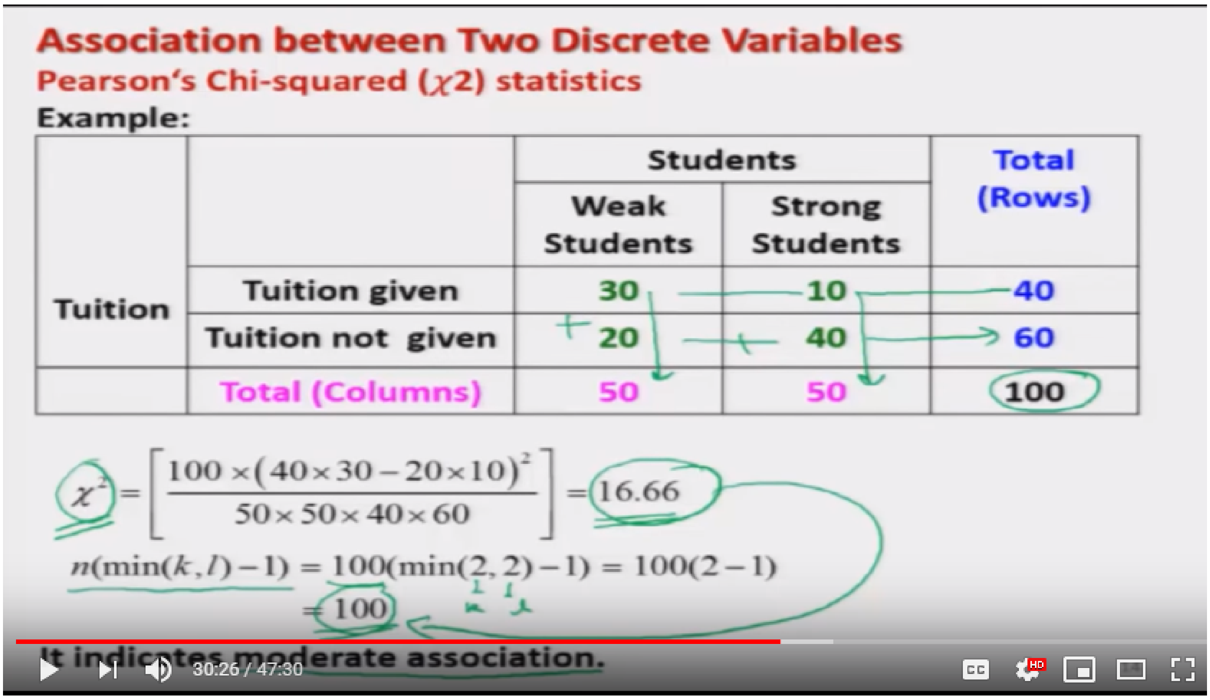
$$\chi^2 = \left[\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}\right]$$

$n = a+b+c+d$

Total freq.

in this table and I try to use the same formula here and

Refer Slide Time: (29:18)



Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics

Example:

		Students		Total (Rows)
		Weak Students	Strong Students	
Tuition	Tuition given	30	10	40
	Tuition not given	20	40	60
	Total (Columns)	50	50	100

$$\chi^2 = \left[\frac{100 \times (40 \times 30 - 20 \times 10)^2}{50 \times 50 \times 40 \times 60} \right] = 16.66$$

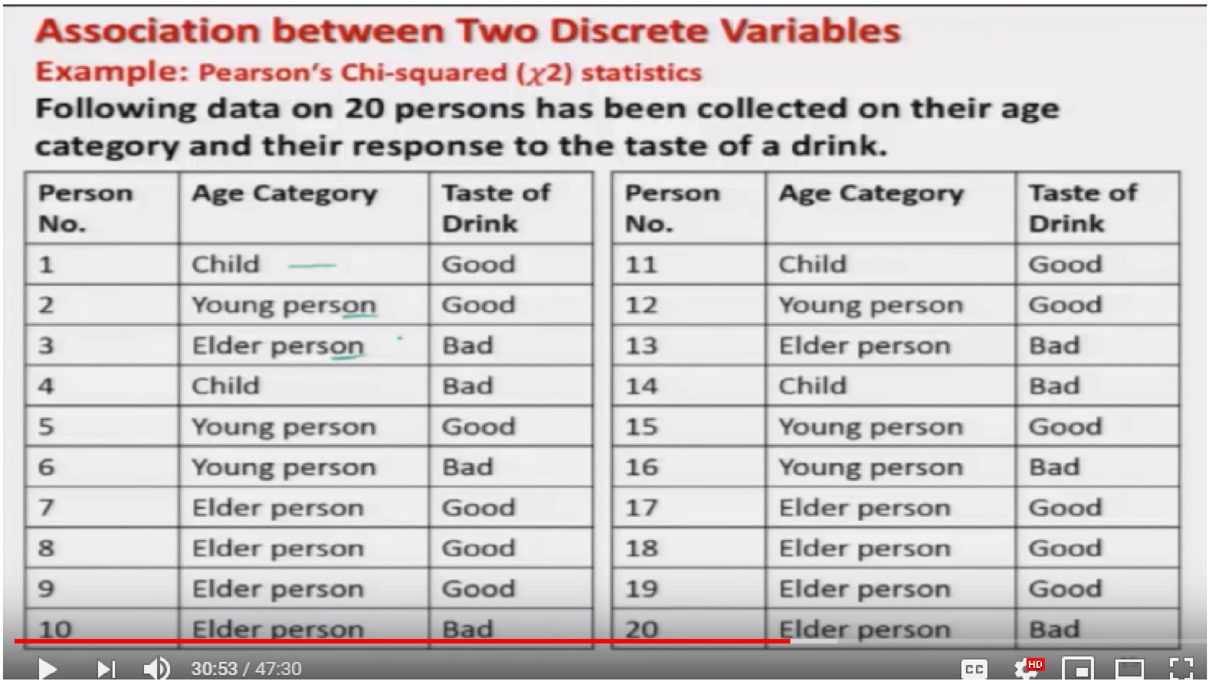
$$n(\min(k,l)-1) = 100(\min(\underset{k}{2}, \underset{l}{2})-1) = 100(2-1)$$

$$= 100$$

It indicates moderate association.

I try to compute this value this value of chi-square comes out to be 16.66, this is a manual calculation, Right! and then the value of the upper limit which is n into minimum of k, l minus one. So, you can see here the value of n here is hundred and the k here is two and l here is two, the minimum value between 2 and 2 is 1. So, 2 minus 1 is 1 and this value here is hundred. So, now you can see here whether this value 16.6 is close to zero or close to hundred this is what we have to see and based on that we have to take a call what's really happening. So, you can see here that this chi-square value is not close to zero, but on the other hand it is also not close to hundred. Right? So, one may conclude that well there can be a moderate association or a lower association it depends on you how you want to interpret it there is no hard and fast rule to decide what is low and what is moderate and what is strong but, yeah in my opinion I can say that well there is a moderate Association.

Refer Slide Time: (30:30)



Association between Two Discrete Variables

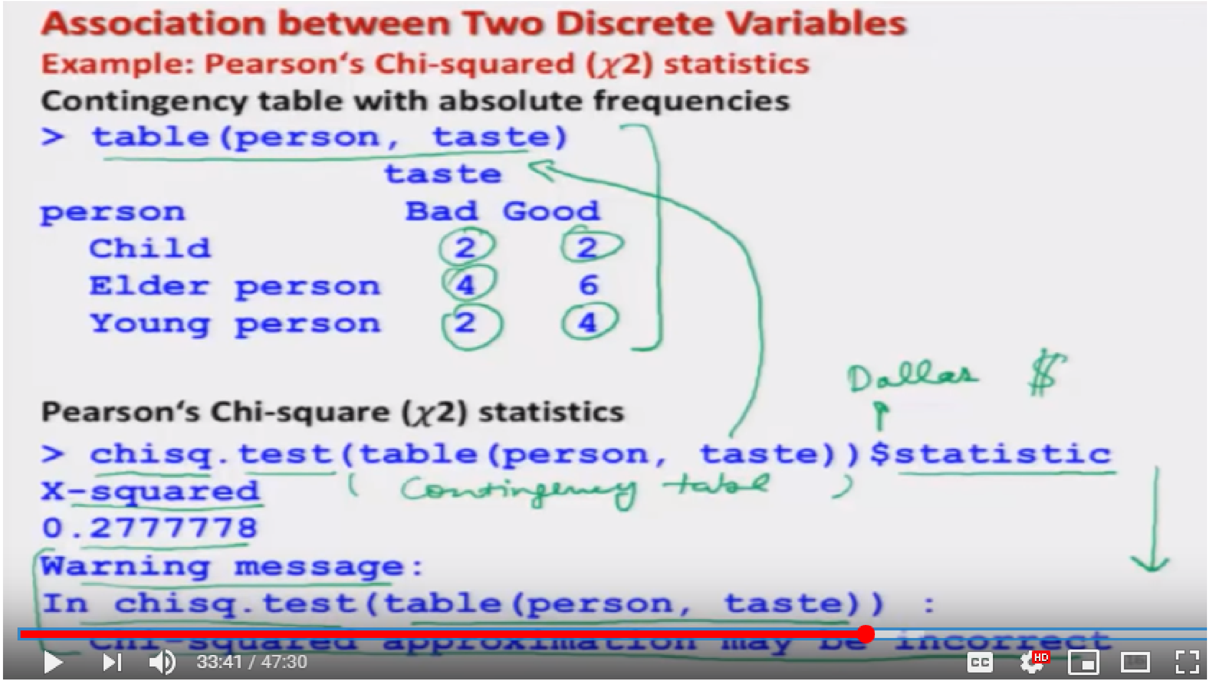
Example: Pearson's Chi-squared ($\chi 2$) statistics

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink	Person No.	Age Category	Taste of Drink
1	Child	Good	11	Child	Good
2	Young person	Good	12	Young person	Good
3	Elder person	Bad	13	Elder person	Bad
4	Child	Bad	14	Child	Bad
5	Young person	Good	15	Young person	Good
6	Young person	Bad	16	Young person	Bad
7	Elder person	Good	17	Elder person	Good
8	Elder person	Good	18	Elder person	Good
9	Elder person	Good	19	Elder person	Good
10	Elder person	Bad	20	Elder person	Bad

Now I try to take the same example which I had considered earlier, and I would try to find out the chi-square statistics. So, this is the same example which I just considered Right! and which we have collected the data on 20 persons and their high responses for the taste of a drink are recorded based on the category of the age that is child's young person or elder person.

Refer Slide Time: (30:54)



Association between Two Discrete Variables

Example: Pearson's Chi-squared ($\chi 2$) statistics

Contingency table with absolute frequencies

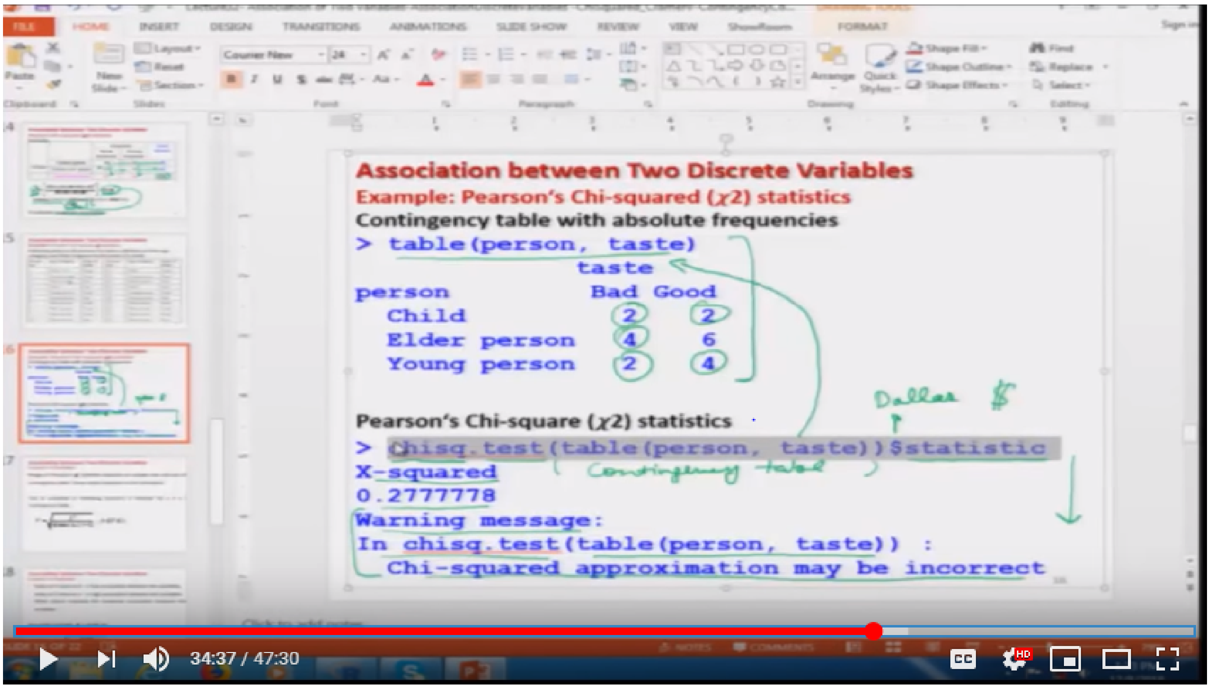
```
> table(person, taste)
                   taste
person             Bad Good
  Child              2    2
  Elder person       4    6
  Young person       2    4
```

Pearson's Chi-square ($\chi 2$) statistics

```
> chisq.test(table(person, taste))$statistic
X-squared
0.2777778
Warning message:
In chisq.test(table(person, taste)) :
  Chi-squared approximation may be incorrect
```

So, you have already done this job that you had obtained the contingency table here using the command table person dot taste. Now in order to obtain the, this chi-square statistics. I am giving you here two things, one is the command and second is the application on this example. So, the command here you can see here is chisq dot test. So, this is a short form of chi-square test and you can see here that inside the argument I am trying to give the data in terms of contingency table. so, this is table comma person comma test and this bracket close this is the same command which is given here. now after this I am using here a command dollar. Which is given on your keyboard and after this I have to write down statistic, statistic and this will give you this outcome. So, the outcome looks like this that this is showing that the value of chi-squared which is written here say X because Chi is a Greek letter so in R, we cannot type the Greek letter this value is coming out to be 0.277 and so on but here you will see that there is also a warning message that and it is saying that in chi-square test this data the chi-squared approximation maybe incorrect. why this is happening? I just informed you that the chi-square is statistics which we have used here to find out the association between the two variables persons and tastes this is actually used for test of hypothesis and to test the hypothesis that whether there is an significant association between the two variables see here in this case persons and tastes. So, when we try to apply the chi-square test of hypothesis then there are certain conditions for the applicability of the test and one of the condition is that that each and every frequency should be greater than five and here you can see here in this case there are several frequencies which are not greater than five like as two two four two and four and that is why this outcome is given here in terms of warning and the chi-square test is telling you well you are trying to find out the chi-square statistic but the number of frequencies are smaller than five so that means the values may not be so accurate and you may have a wrong conclusion but that is related to the test of hypothesis and here our objective is very simple I just want to show you how to calculate the chi-squared statistic. If I can take a bigger data also but then you will not be able to match what R is doing and what you can obtain manually rather my request will be you try to take the same example and try to create the same contingency table yourself and try to compute this chi-square value this will come out to be the same.

Refer Slide Time: (34:31)



Association between Two Discrete Variables

Example: Pearson's Chi-squared ($\chi 2$) statistics

Contingency table with absolute frequencies

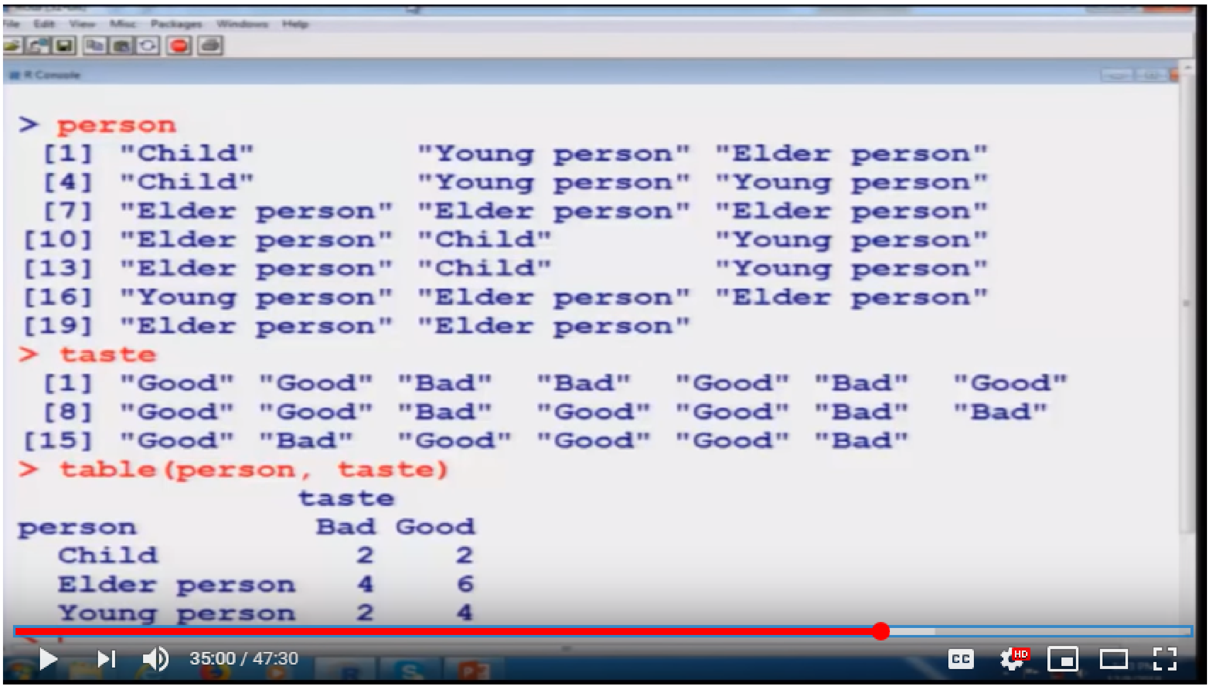
```
> table(person, taste)
                    taste
person              Bad  Good
   Child             2    2
   Elder person      4    6
   Young person      2    4
```

Pearson's Chi-square ($\chi 2$) statistics

```
> chisq.test(table(person, taste))$statistic
X-squared
0.2777778
Warning message:
In chisq.test(table(person, taste)) :
   Chi-squared approximation may be incorrect
```

So, now let us try to see that how you can compute these things on the R software on the R console.

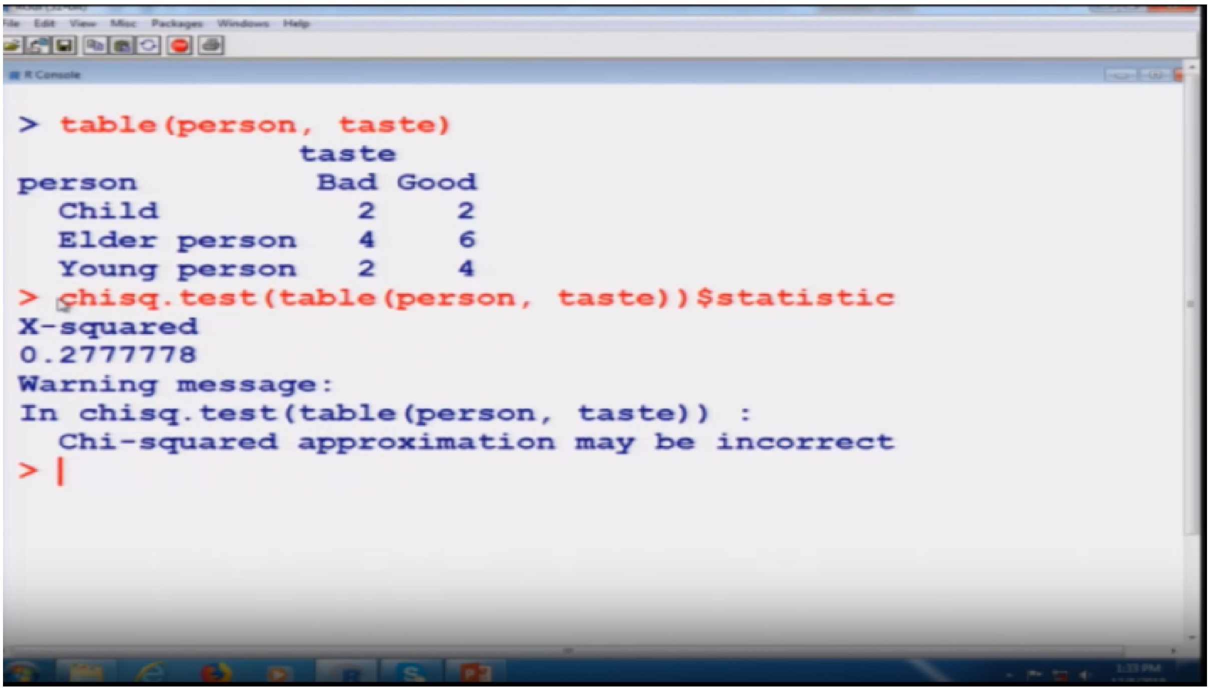
Refer Slide Time:(34:41)



```
> person
 [1] "Child"        "Young person" "Elder person"
 [4] "Child"        "Young person" "Young person"
 [7] "Elder person" "Elder person" "Elder person"
[10] "Elder person" "Child"        "Young person"
[13] "Elder person" "Child"        "Young person"
[16] "Young person" "Elder person" "Elder person"
[19] "Elder person" "Elder person"
> taste
 [1] "Good" "Good" "Bad"  "Bad"  "Good" "Bad"  "Good"
 [8] "Good" "Good" "Bad"  "Good" "Good" "Bad"  "Bad"
[15] "Good" "Bad"  "Good" "Good" "Good" "Bad"
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
```

So, you may see here that I already have this data on person we just used it taste and here like this and if you try to create here a table say here person and taste you get here the same contingency table. Right?

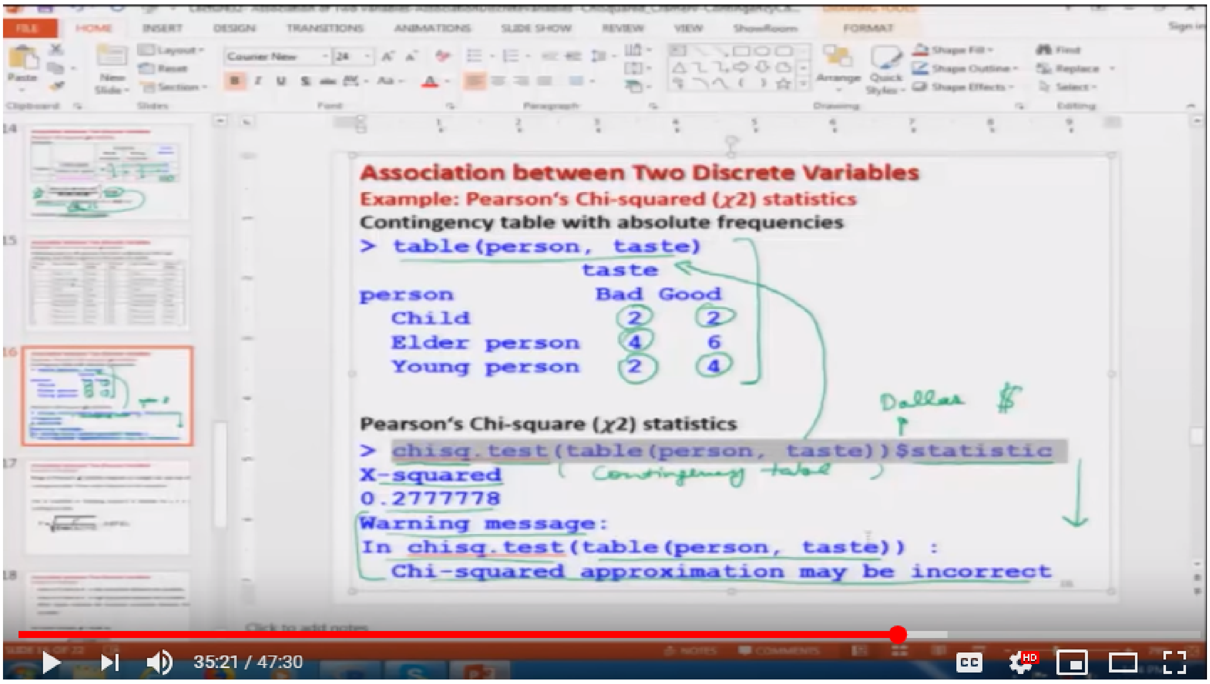
Refer Slide Time: (35:04)



```
> table(person, taste)
                  taste
person             Bad Good
  Child              2    2
  Elder person       4    6
  Young person       2    4
> chisq.test(table(person, taste))$statistic
X-squared
0.2777778
Warning message:
In chisq.test(table(person, taste)) :
  Chi-squared approximation may be incorrect
> |
```

So, I try to clear the screen so that you can see everything clearly and now I try to compute the chi-square statistics. So, you can see here I just use the same command what I have used in the slides and this is giving you the same outcome. Okay?

Refer Slide Time: (35:21)



Association between Two Discrete Variables

Example: Pearson's Chi-squared (χ2) statistics

Contingency table with absolute frequencies

```
> table(person, taste)
                  taste
person             Bad Good
  Child              2    2
  Elder person       4    6
  Young person       2    4
```

Pearson's Chi-square (χ2) statistics

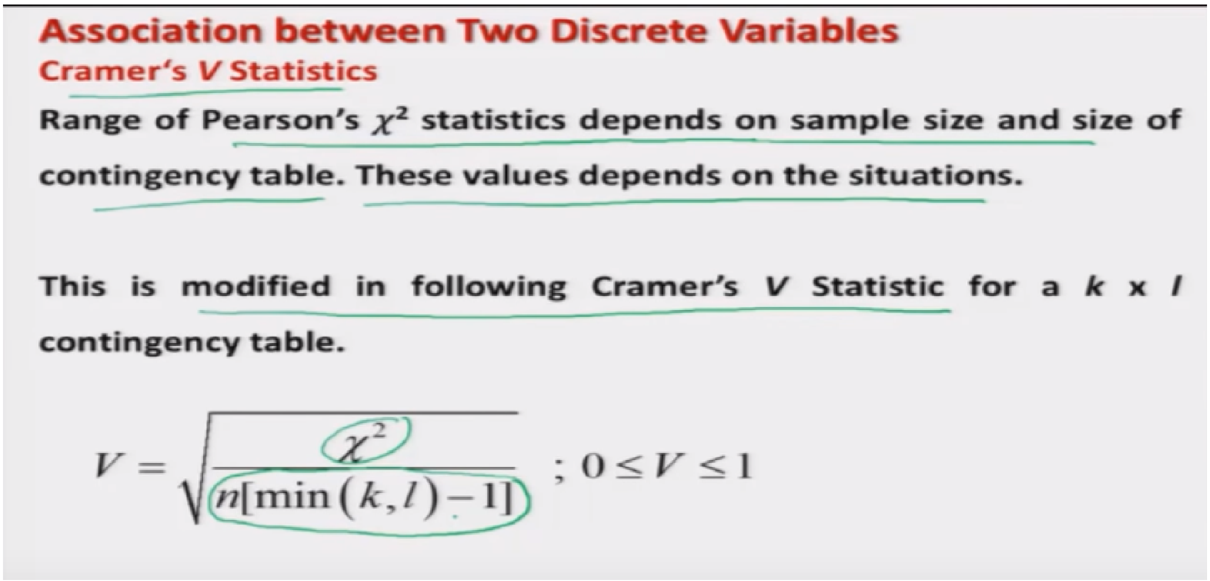
```
> chisq.test(table(person, taste))$statistic
X-squared
0.2777778
Warning message:
In chisq.test(table(person, taste)) :
  Chi-squared approximation may be incorrect
```

Dollar $

(Contingency table)

Now, let us come back to our slides now you see by looking at the value of chi-square you can judge whether the association between the academic performance and the tuition is significant or not. Right? just by computing the values of n into minimum of k, l minus one and so on. But, now I would try to address another aspect you can see here in this case you need to compute the values of the range, well. One is zero but another is n into minimum of k, l minus one. So, the value itself is not giving you a clear cut indication for example if you remember in the case of say correlation coefficient that was lying between minus one and plus one or the magnitude will lie between zero and one. So, just by looking at the value of odd you can very easily communicate whether the association is high or low and so on. So, here a modification was registered in the chi-square statistics and a new statistics which is a modified version of the chi-square statistics was defined as Cramer's V statistics and in this case what happened.

Refer Slide Time: (36:40)



Association between Two Discrete Variables

Cramer's *V* Statistics

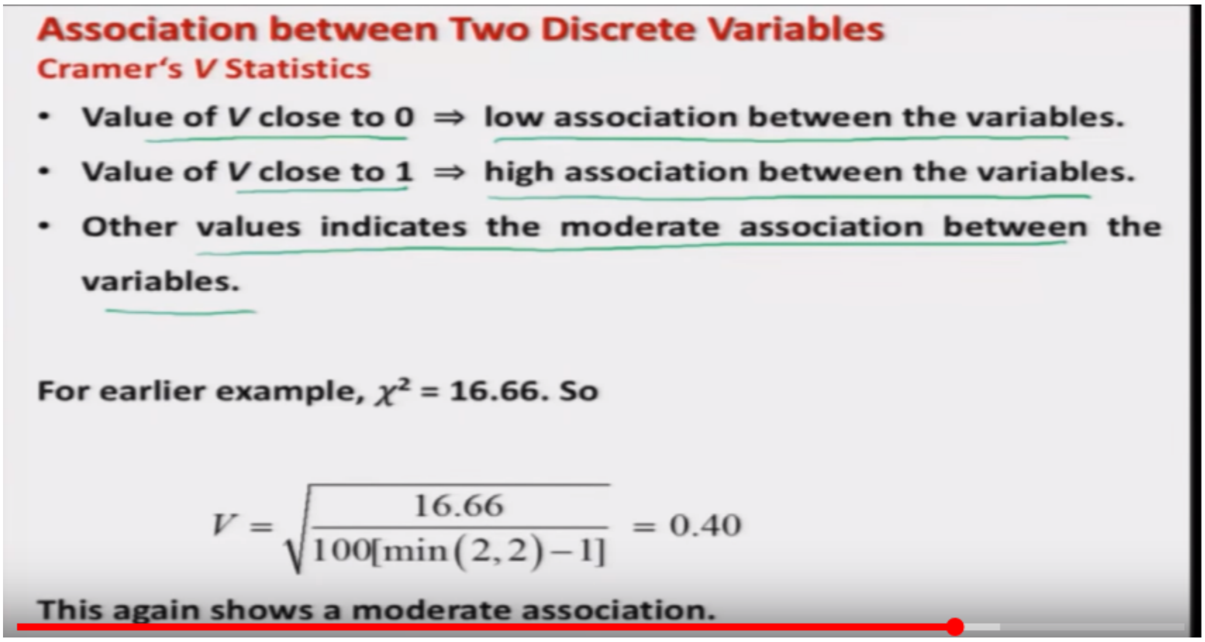
Range of Pearson's χ^2 statistics depends on sample size and size of contingency table. These values depends on the situations.

This is modified in following Cramer's *V* Statistic for a *k* x *l* contingency table.

$$V = \sqrt{\frac{\chi^2}{n[\min(k,l)-1]}} \quad ; 0 \le V \le 1$$

That in the case of this Cramer V statistics the range is if the range of the Pearsons Chi's statistic depends on the sample size and the size of the contingency table. So, these values depends on the situation that what is the number of row what is the number of column and so on. So, this issue was solved and a modified version of the Person chi-square statistics were presented as Cramer's V statistics for a k cross l contingency table, for the same table and this was defined as say square root of the chi-square value which we have just obtained and it was divided by the limit of this chi-square statistics n into minimum of k, l minus one and the advantage of this these statistics was that it lies between zero and one.

Refer Slide Time: (37:36)



Association between Two Discrete Variables

Cramer‘s *V* Statistics

- Value of *V* close to 0 ⇒ low association between the variables.
- Value of *V* close to 1 ⇒ high association between the variables.
- Other values indicates the moderate association between the variables.

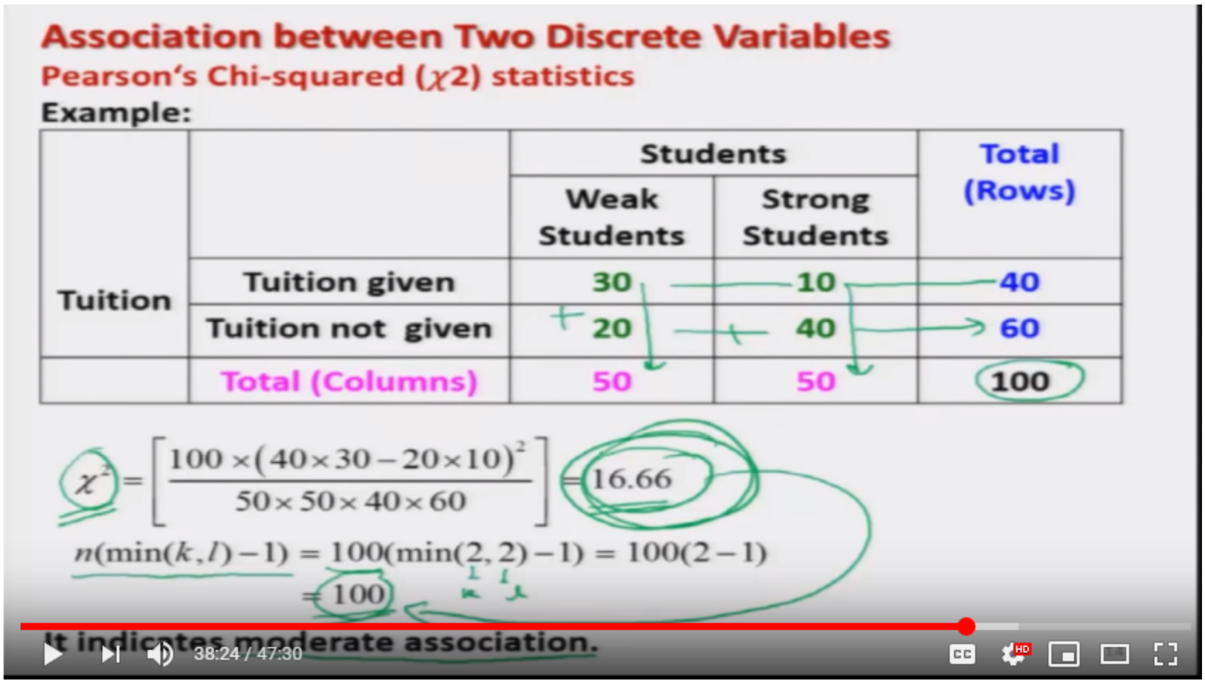
For earlier example, $\chi^2 = 16.66$. So

$$V = \sqrt{\frac{16.66}{100[\min(2,2)-1]}} = 0.40$$

This again shows a moderate association.

So, now making inference about the degree of Association became more simpler. For example, we can conclude that if the value of V is close to zero that would simply indicate the low association between the variables and in case of the value of V is close to one then this will indicate the high association between the variables and similarly if you have any other value say between zero and one then depending on the magnitude of the value that would indicate whether there has been a moderate association or a lower association or a strong association. So, now you can see here that in the earlier case we had obtained

Refer Slide Time: (38:16)



Association between Two Discrete Variables

Pearson's Chi-squared ($\chi 2$) statistics

Example:

		Students		Total (Rows)
		Weak Students	Strong Students	
Tuition	Tuition given	30	10	40
	Tuition not given	20	40	60
	Total (Columns)	50	50	100

$$\chi^2 = \left[\frac{100\times(40\times30-20\times10)^2}{50\times50\times40\times60}\right] = \boxed{16.66}$$

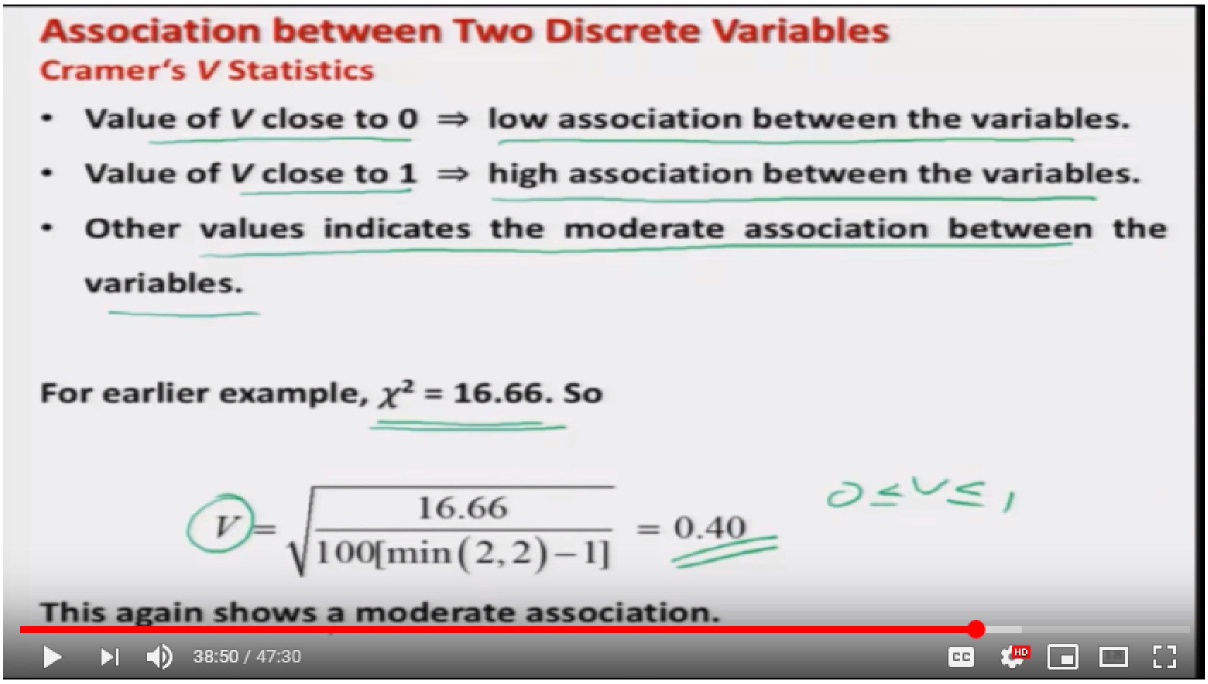
$$n(\min(k,l)-1) = 100(\min(\underset{k}{\overset{1}{2}},\underset{l}{\overset{1}{2}})-1) = 100(2-1)$$

$$= 100$$

It indicates moderate association.

the value of chi-square to be here 16.66, and we had concluded that the Association is moderate. So, now

Refer Slide Time: (38:28)



Association between Two Discrete Variables

Cramer's *V* Statistics

- Value of *V* close to 0 ⇒ low association between the variables.
- Value of *V* close to 1 ⇒ high association between the variables.
- Other values indicates the moderate association between the variables.

For earlier example, $\chi^2 = 16.66$. So

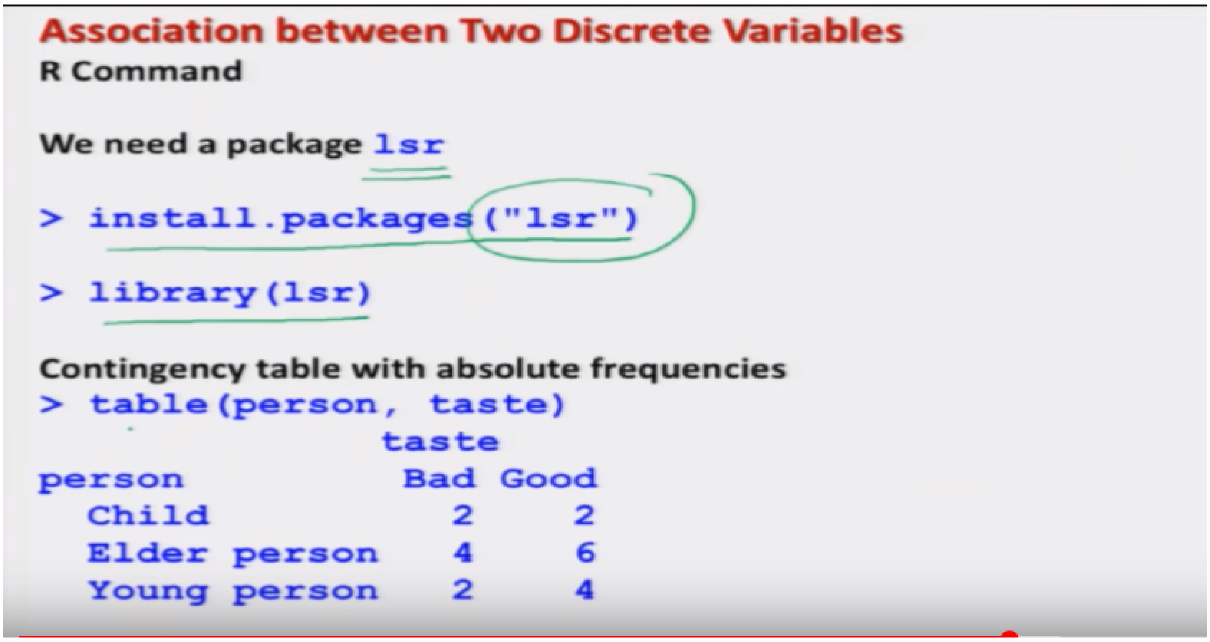
$$V = \sqrt{\frac{16.66}{100[\min(2,2)-1]}} = 0.40$$

$0 \le V \le 1$

This again shows a moderate association.

for the same example I try to compute this value. So, using that value of chi-square to be 16.66, I try to compute the value of V statistics and this comes out to be 0.40. So, that would indicate once again that ideally, V should lie between zero and one but it would lying somewhere in the middle. So, I can say that there is a moderate association

Refer Slide Time: (38:52)



Association between Two Discrete Variables

R Command

We need a package `lsr`

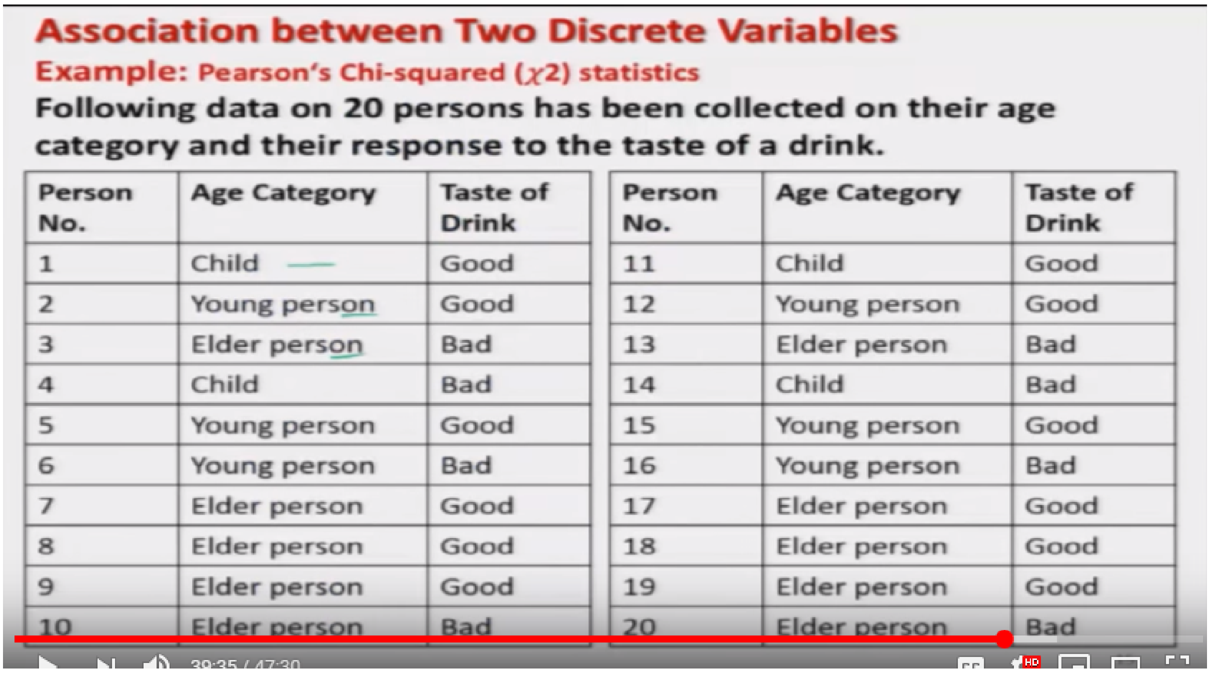
```
> install.packages("lsr")

> library(lsr)
```

Contingency table with absolute frequencies

```
> table(person, taste)
                  taste
person            Bad Good
   Child            2    2
   Elder person     4    6
   Young person     2    4
```

and similarly, if you want to compute this V statistics in the R software for that we need a special package which is called as lsr. So, first we need to install this package by using the command install dot packages and inside the argument within the double quotes you have to write lsr and once you do it the package will be installed I am not showing you here because I discuss it in the starting lectures and after that you need to note the library by using the command library inside the argument lsr, and if you see what we had obtained earlier in this data set.

Refer Slide Time: (39:35)



Association between Two Discrete Variables

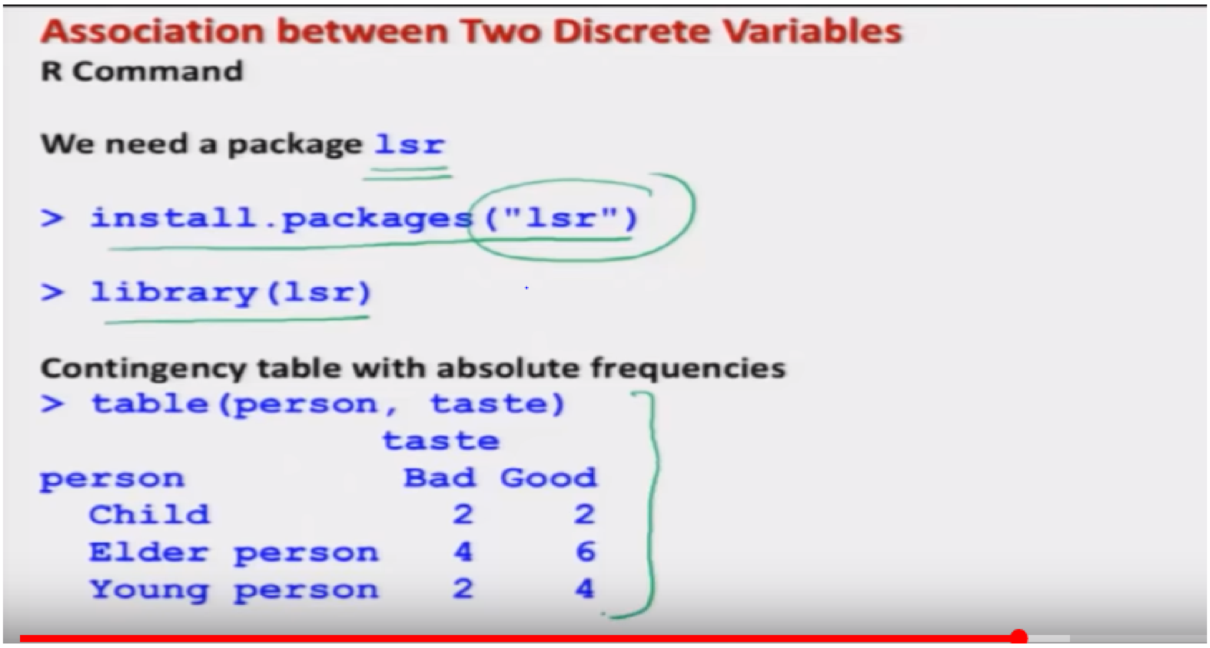
Example: Pearson's Chi-squared (χ2) statistics

Following data on 20 persons has been collected on their age category and their response to the taste of a drink.

Person No.	Age Category	Taste of Drink	Person No.	Age Category	Taste of Drink
1	Child	Good	11	Child	Good
2	Young person	Good	12	Young person	Good
3	Elder person	Bad	13	Elder person	Bad
4	Child	Bad	14	Child	Bad
5	Young person	Good	15	Young person	Good
6	Young person	Bad	16	Young person	Bad
7	Elder person	Good	17	Elder person	Good
8	Elder person	Good	18	Elder person	Good
9	Elder person	Good	19	Elder person	Good
10	Elder person	Bad	20	Elder person	Bad

This example where we had 20 percent who responded for the taste of drink.

Refer Slide Time: (39:39)



Association between Two Discrete Variables

R Command

We need a package lsr

```
> install.packages("lsr")

> library(lsr)
```

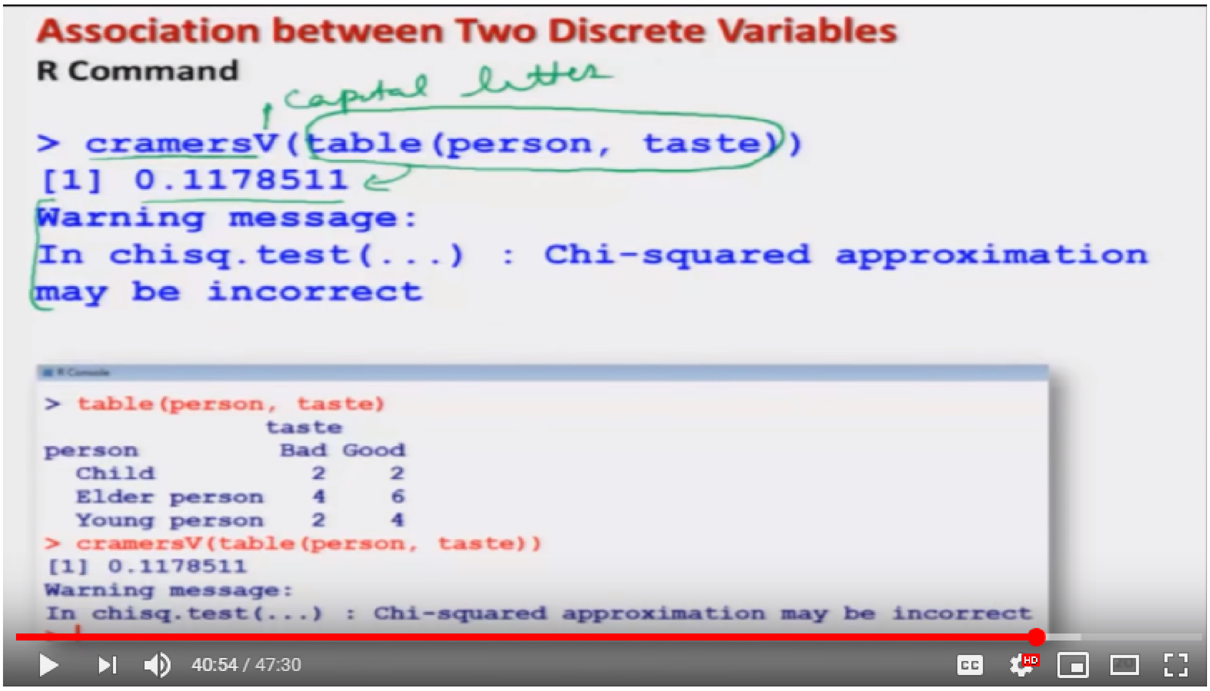
Contingency table with absolute frequencies

```
> table(person, taste)
                   taste
person             Bad Good
   Child             2    2
   Elder person      4    6
   Young person      2    4
```

So, I am going to calculate this value for this thing. So, remember one thing I am taking here two examples one I am trying to do manually and and another I am trying to do on the basis of R software.

Right? So, this is an example which we had just done on the R software. So, we had obtained here this type of contingency table and

Refer Slide Time: (40:02)



Association between Two Discrete Variables

R Command

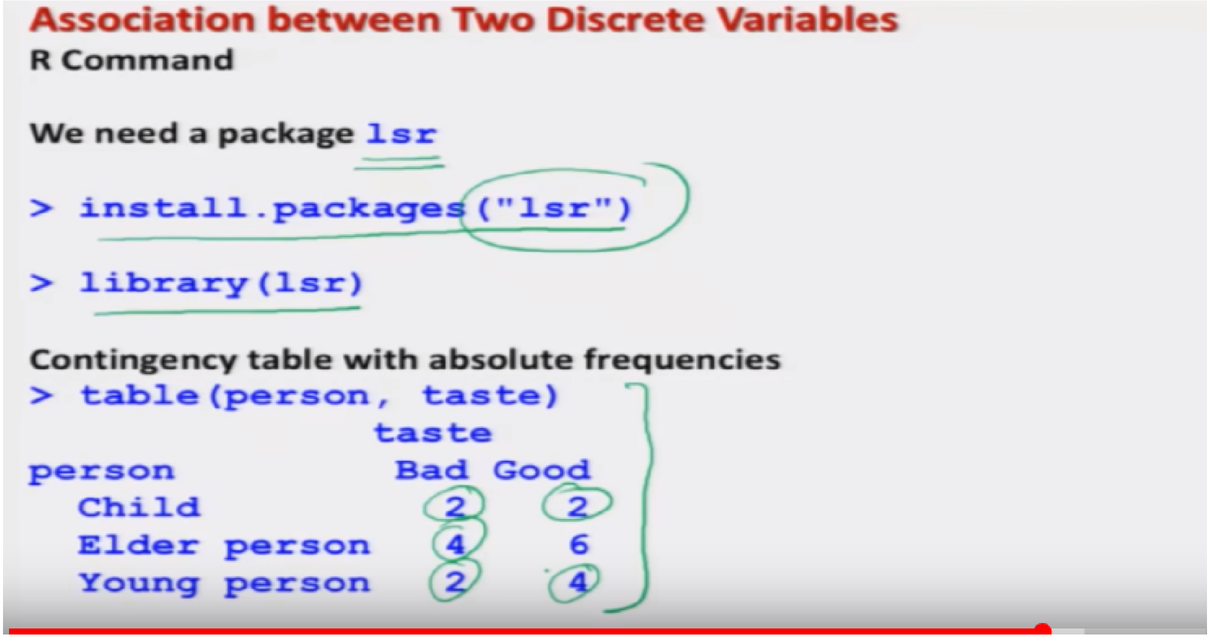
capital letter

```
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation
may be incorrect
```

```
> table(person, taste)
                taste
person           Bad Good
  Child            2    2
  Elder person     4    6
  Young person     2    4
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation may be incorrect
```

Now I would try to compute the value of cramers V statistics. So, once again I would like to show you here two things what is the command? and what is the interpretation? So, you can see here the command here is Cramers V Cramers and V is in capital letters that you have to remember, and then I have to give the contingency table for which I would like to compute the value of Cramer's V statistics and once I do it this gives me the value 0.11 . Right? and yeah you once again you will see here the warning message this warning message is coming out because of the same reason that this is based on the chi-square statistic and the total number of frequencies in every cell

Refer Slide Time: (40:57)



Association between Two Discrete Variables

R Command

We need a package `lsr`

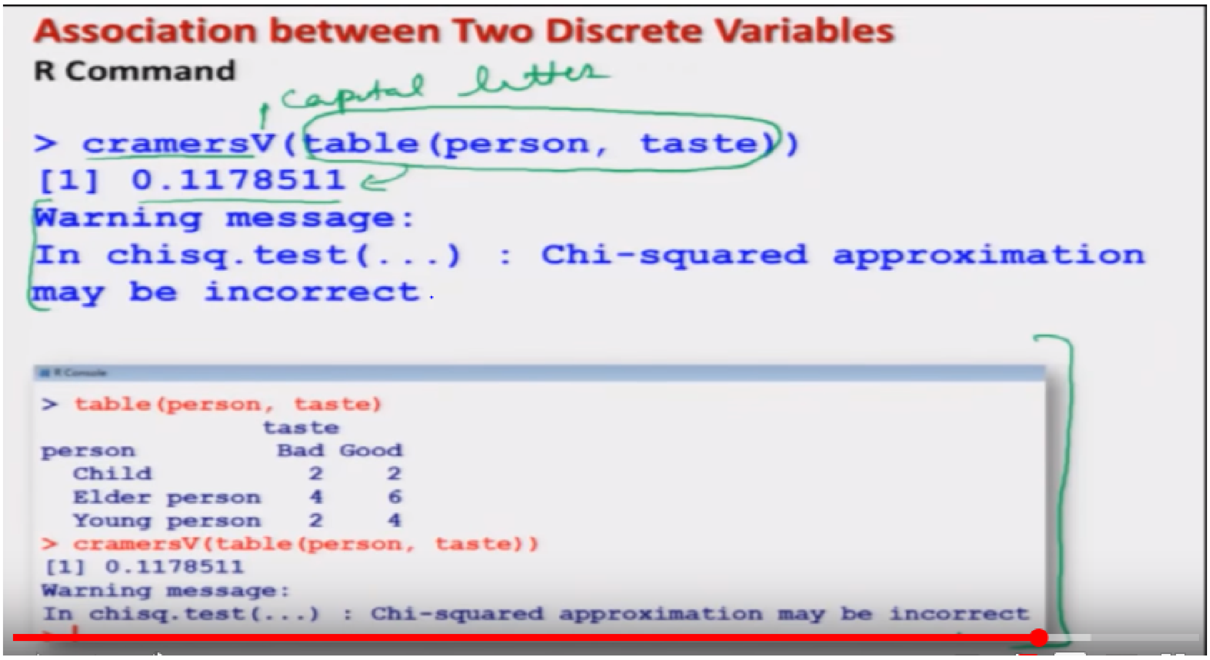
```
> install.packages("lsr")

> library(lsr)
```

Contingency table with absolute frequencies

```
> table(person, taste)
                   taste
person             Bad Good
  Child              2    2
  Elder person       4    6
  Young person       2    4
```

you can see here they are say, smaller than five like as two, two, four and so on. Right?

Refer Slide Time: (41:03)



Association between Two Discrete Variables

R Command

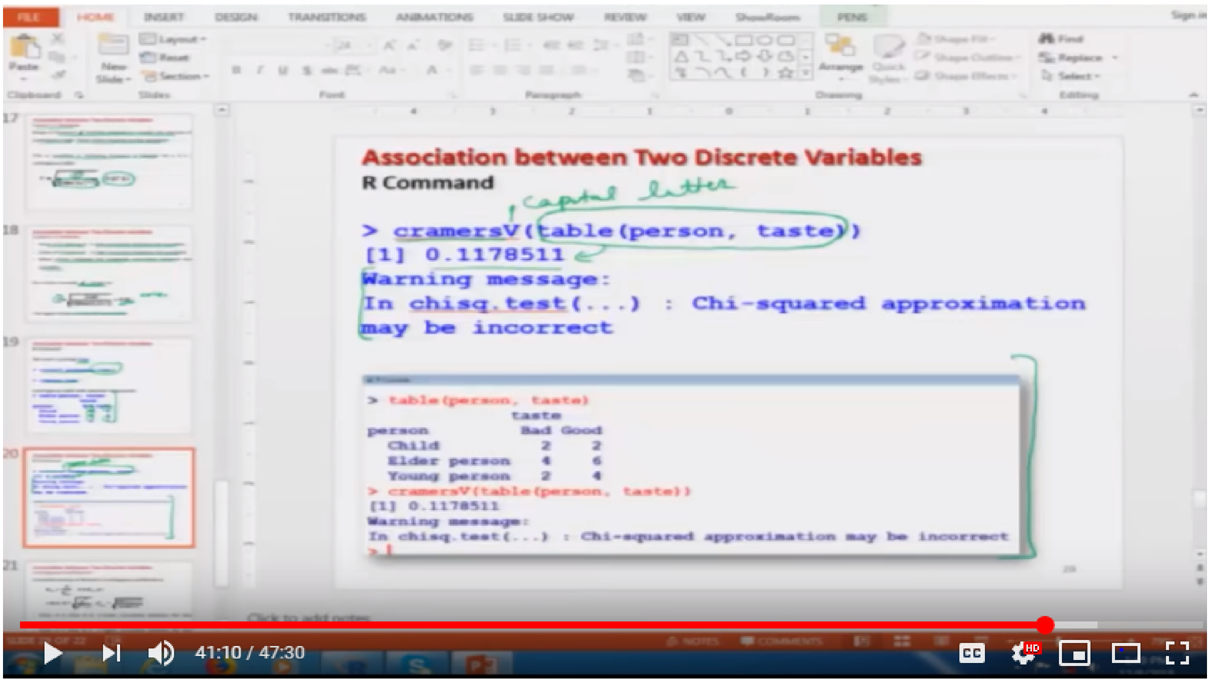
capital letter

```
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation
may be incorrect.
```

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation may be incorrect
```

So, this is how you can do it? and this is the screen shot but I would like to show you it on the R console also.

Refer Slide Time: (41:10)



Association between Two Discrete Variables

R Command

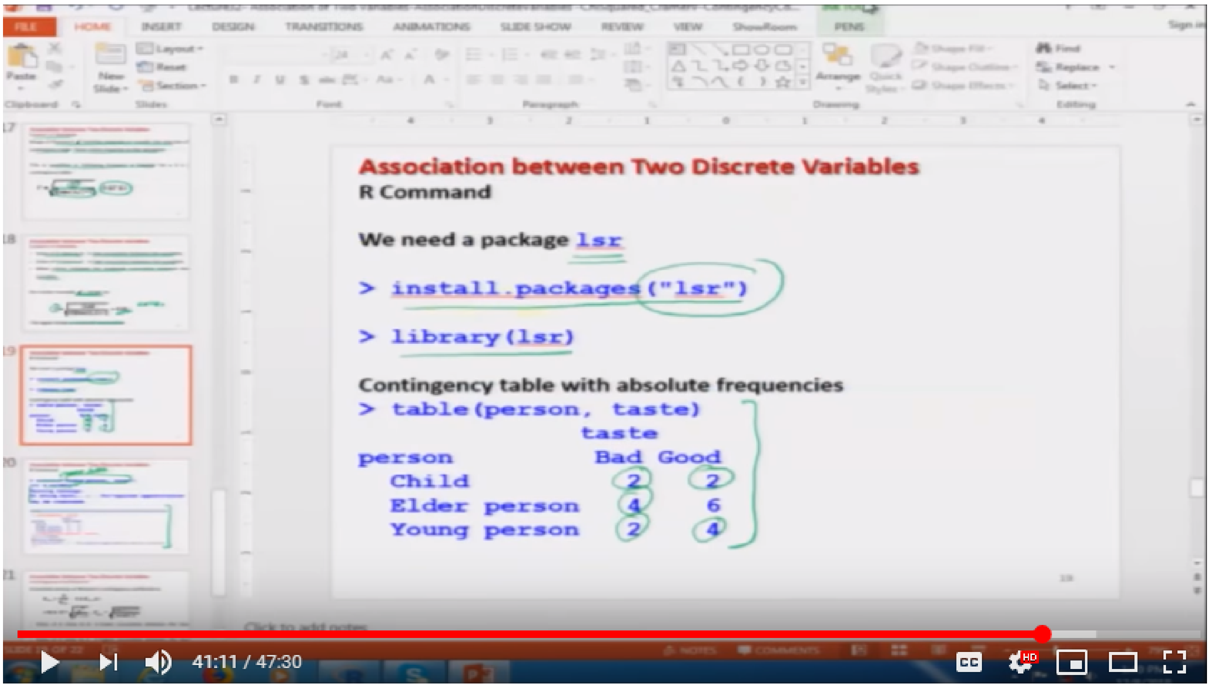
capital letter

```
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation
may be incorrect
```

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation may be incorrect
> 
```

So, first I try to note the library.

Refer Slide Time: (41:11)



Association between Two Discrete Variables

R Command

We need a package lsr

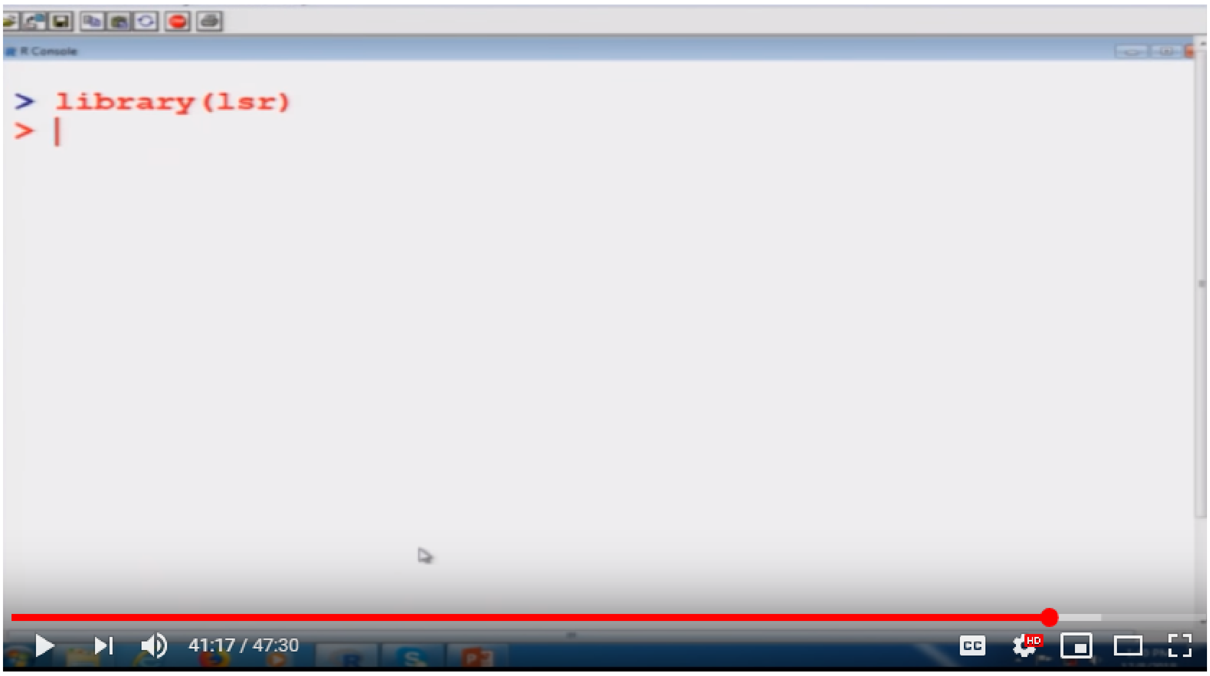
```
> install.packages("lsr")

> library(lsr)
```

Contingency table with absolute frequencies

```
> table(person, taste)
                  taste
person            Bad  Good
   Child           2    2
   Elder person    4    6
   Young person    2    4
```

I already have installed this package on my computer.

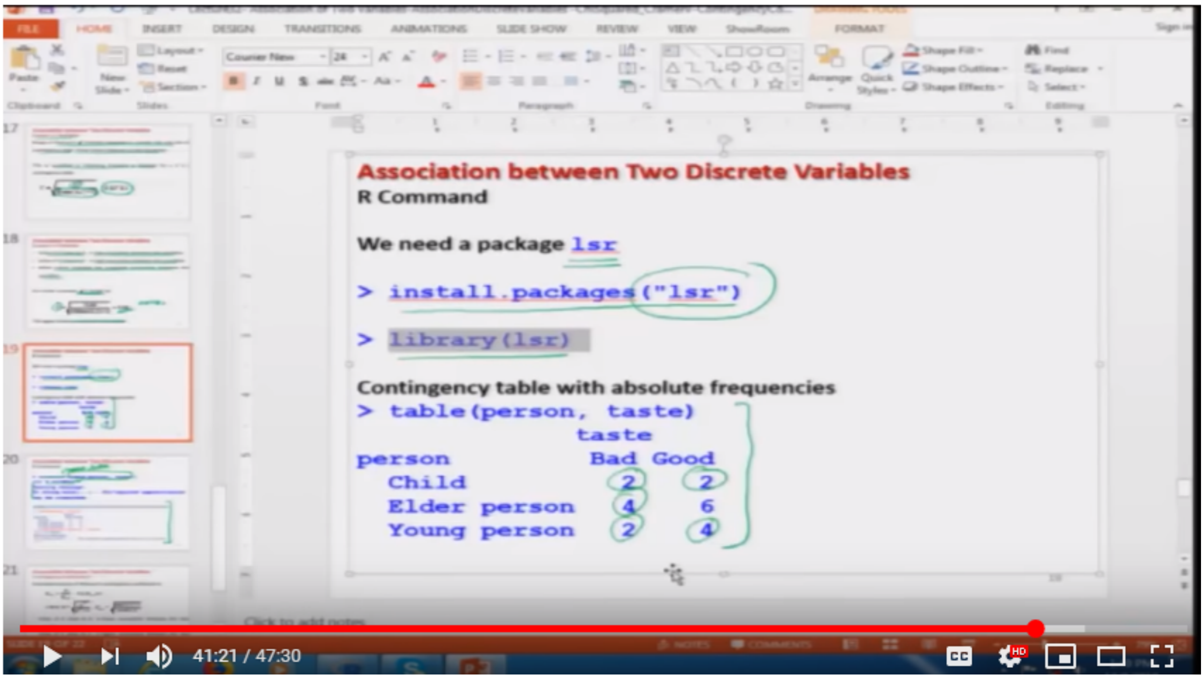
Refer Slide Time: (41:15)



```
> library(lsr)
> |
```

So, I simply have to load it

Refer Slide Time: (41:20)



Association between Two Discrete Variables

R Command

We need a package lsr

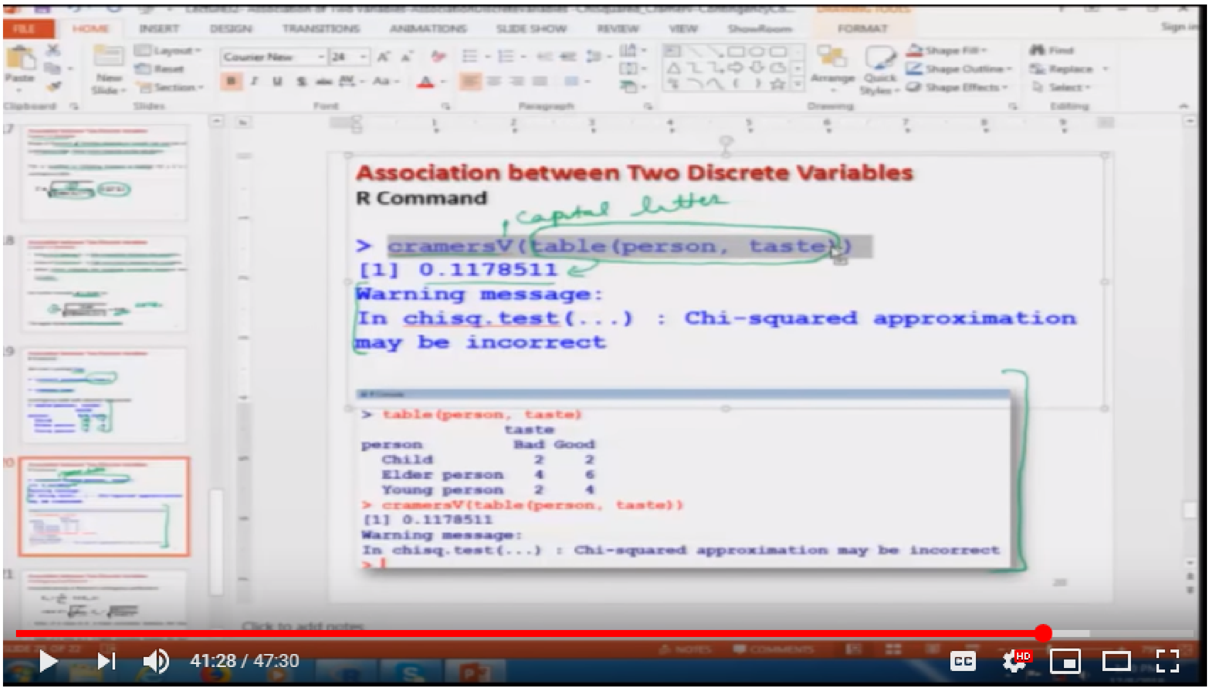
```
> install.packages("lsr")

> library(lsr)
```

Contingency table with absolute frequencies

```
> table(person, taste)
                  taste
person            Bad  Good
   Child           2    2
   Elder person    4    6
   Young person    2    4
```

and after that

Refer Slide Time: (41:22)



Association between Two Discrete Variables

R Command

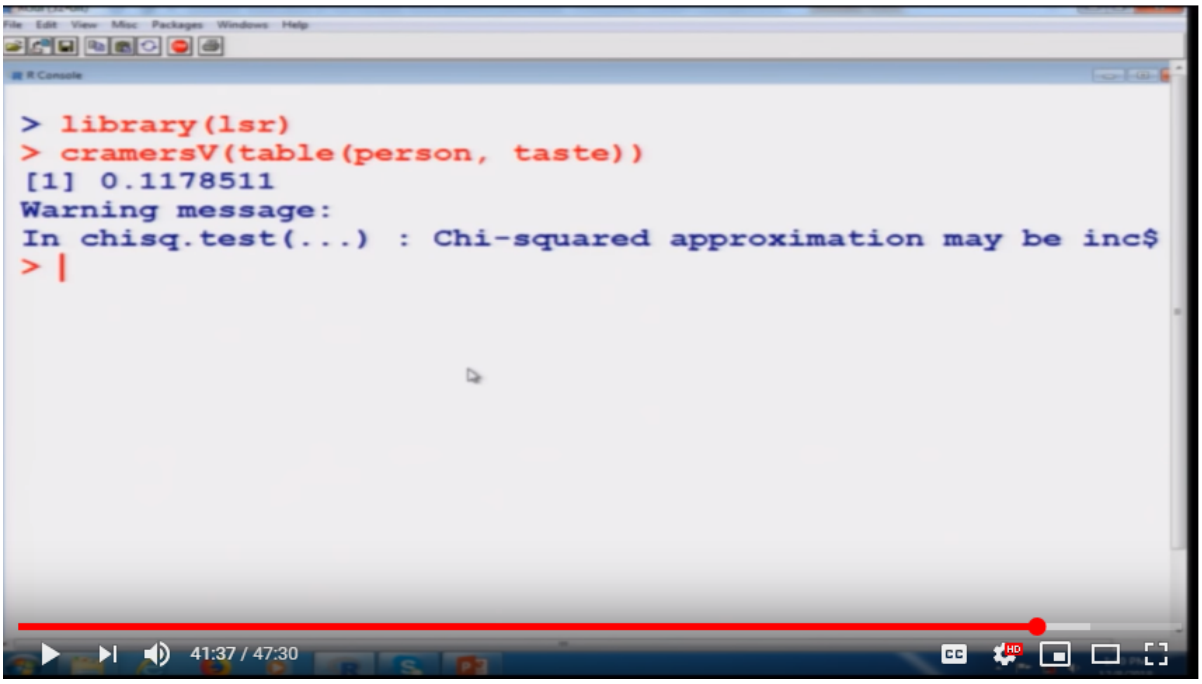
capital letter

```
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation
may be incorrect
```

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation may be incorrect
> |
```

I will try to use the command here Cramers V and inside the arguments

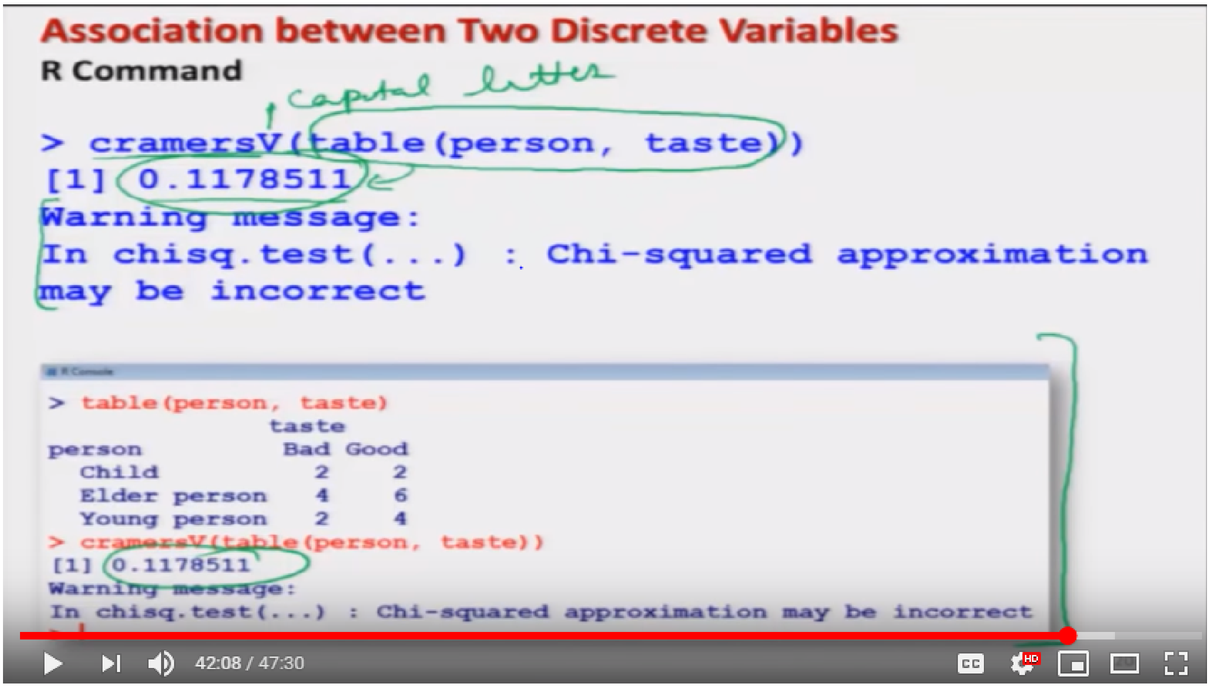
Refer Slide Time: (41:32)



```
> library(lsr)
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation may be inc$
> |
```

I have to give the contingency table of for which we want to compute it. So, you can see here this is coming out to be like this. Right?

Refer Slide Time: (41:43)



Association between Two Discrete Variables

R Command

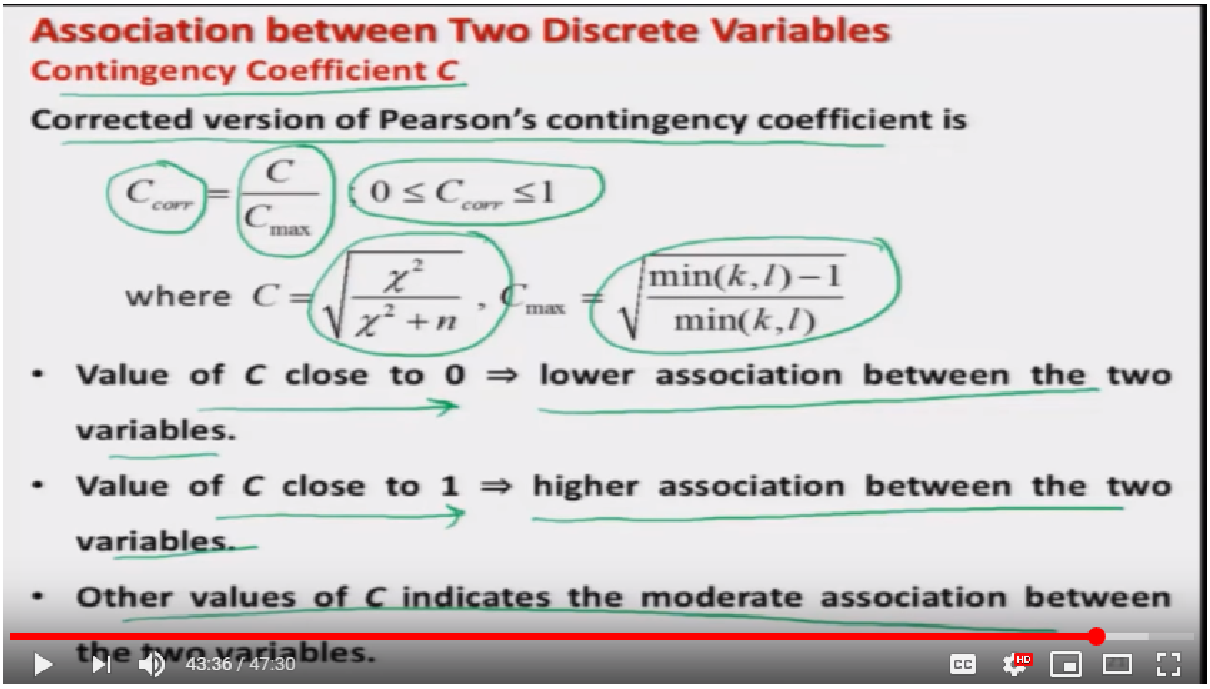
capital letter

```
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation
may be incorrect
```

```
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
> cramersV(table(person, taste))
[1] 0.1178511
Warning message:
In chisq.test(...) : Chi-squared approximation may be incorrect
```

Okay. So, you can see it is not difficult at all and it is more easy to interpret because all the values are going to lie between zero and one. So, by looking at the value of see here 0.11, one can conclude here or here that the association is there, but it is quite low. Right? So, the taste is not much depending on the age.

Refer Slide Time: (42:09)



Association between Two Discrete Variables

Contingency Coefficient C

Corrected version of Pearson's contingency coefficient is

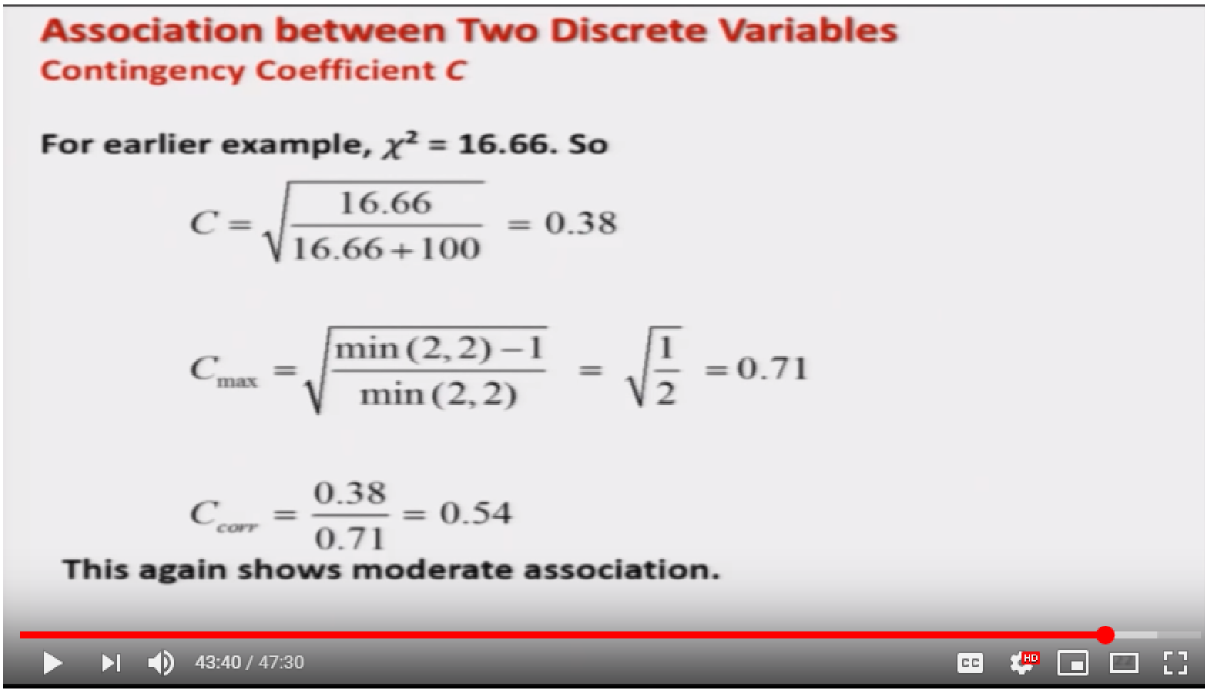
$$C_{corr} = \frac{C}{C_{\max}} \qquad 0 \le C_{corr} \le 1$$

$$\text{where } C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \quad C_{\max} = \sqrt{\frac{\min(k,l)-1}{\min(k,l)}}$$

- Value of *C* close to 0 ⇒ lower association between the two variables.
- Value of *C* close to 1 ⇒ higher association between the two variables.
- Other values of *C* indicates the moderate association between the two variables.

Now, after this there is another coefficient which is used to measure the degree of association in such a case and this is actually the called as contingency coefficient. And this is simply the corrected version of the Person's contingency coefficient. which is based on the chi-square statistics once again and this contingency co-efficiency which I am denoting here as C, C or Double R which is an abbreviation for corrected this is defined as C upon C max, where C is given by square root of Chi square divided by Chi square plus n and C max is given by this quantity that is the square root of minimum of k, l minus one divided by minimum of k, l and this is statistics also has an advantage that it is lying between zero and one so it is more convenient to take conclusions or statically inference using the value which lies between zero & one. So, this also have similar interpretations like it's value of C, if it is close to zero that would indicate a lower association between the two variables, in case if the value of C is close to one this is going to indicate a higher association between the two variables and other values of C between zero and one they would try to negatable indicate the degree of association between the two variables.

Refer Slide Time: (43:36)



Association between Two Discrete Variables

Contingency Coefficient *C*

For earlier example, $\chi^2 = 16.66$. So

$$C = \sqrt{\frac{16.66}{16.66+100}} = 0.38$$

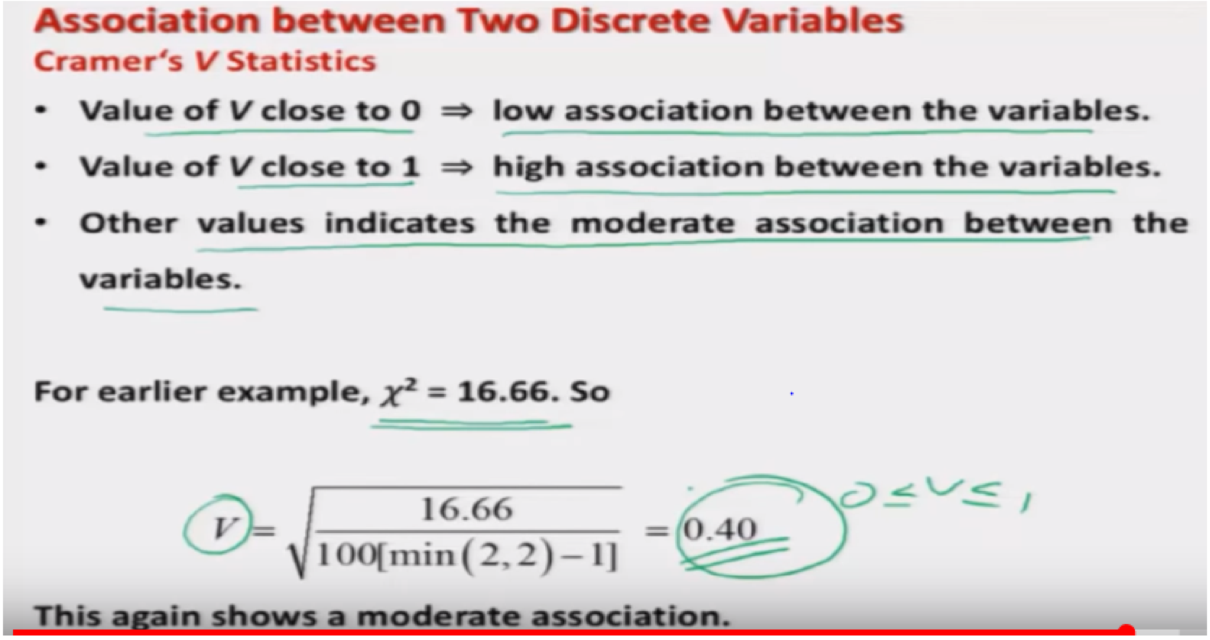
$$C_{\max} = \sqrt{\frac{\min(2,2)-1}{\min(2,2)}} = \sqrt{\frac{1}{2}} = 0.71$$

$$C_{corr} = \frac{0.38}{0.71} = 0.54$$

This again shows moderate association.

And now, in case if I try to take this example just for your remembrance.

Refer Slide Time: (43:42)



Association between Two Discrete Variables

Cramer's *V* Statistics

- Value of *V* close to 0 ⇒ low association between the variables.
- Value of *V* close to 1 ⇒ high association between the variables.
- Other values indicates the moderate association between the variables.

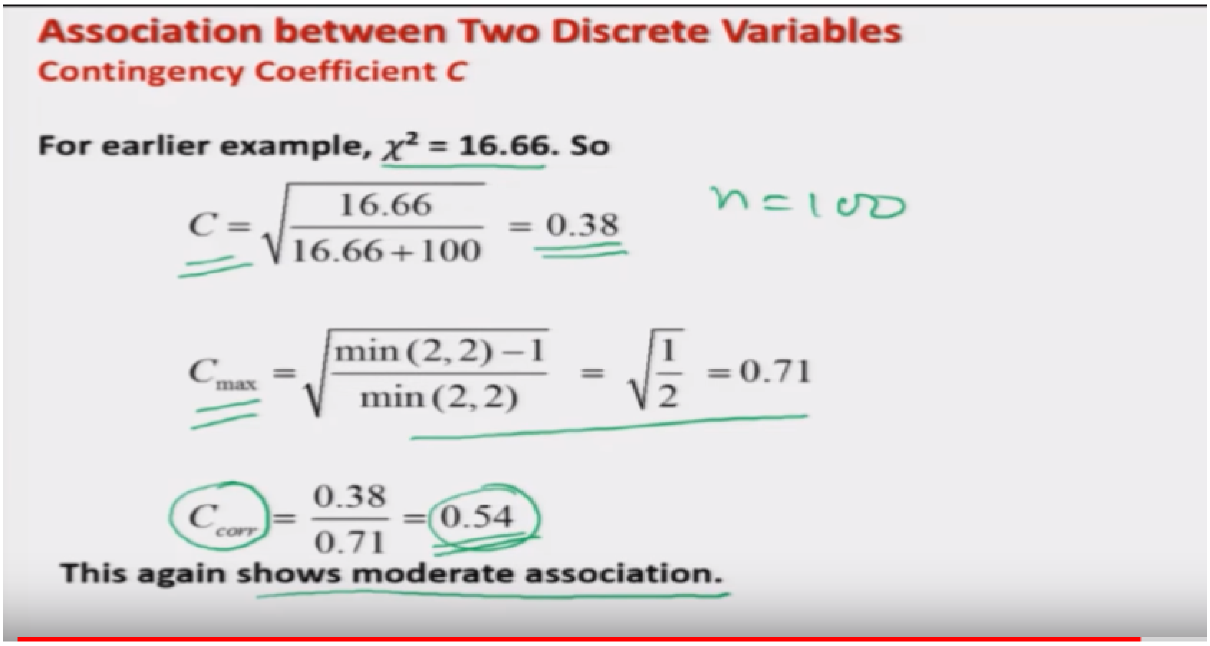
For earlier example, $\chi^2 = 16.66$. So

$$V = \sqrt{\frac{16.66}{100[\min(2,2)-1]}} = 0.40 \qquad 0 \le V \le 1$$

This again shows a moderate association.

This students versus tuition where we have used we have found the value of here Chi Square to be 16.66, now for the same thing I would like to find out this contingency coefficient. So, you can see here from this value I was concluding that there is a moderate Association then, I use the cramers V statistics this is also indicating the moderate Association and now let me see what happens in the case of this contingency coefficient.

Refer Slide Time: (44:09)



Association between Two Discrete Variables

Contingency Coefficient *C*

For earlier example, $\underline{\chi^2 = 16.66}$. So

$$\underline{\underline{C}} = \sqrt{\frac{16.66}{16.66+100}} = \underline{\underline{0.38}} \qquad n = 100$$

$$\underline{\underline{C_{\max}}} = \underline{\sqrt{\frac{\min(2,2)-1}{\min(2,2)}} = \sqrt{\frac{1}{2}} = 0.71}$$

$$\boxed{C_{corr}} = \frac{0.38}{0.71} = \boxed{0.54}$$

This again shows moderate association.

So, if I try to compute the value of C based on the values of chi-square equal to 16.66, then the value of C is coming out to be 0.38, obviously n is here hundred and the value of C max is coming out to be here 0.71, and finally the value of contingency coefficient C is coming out to be 0.54. So, since this value is lying between zero and one. So, this value 0.54 is lying somewhere in the middle of zero and one so once again, I can say that this is indicating a moderate association between the two variables. Now, we have considered different types of things, different types of measures to find out the degree of association between the two variables where the variables are in the form of counting data, well, you can see here that different statistics they are giving us different values and they have got different interpretations but then obviously, I believe, I personally believe that, if there is an association present inside the data, then ideally all the statistics should indicate the same thing there can be a small difference for example, Cramers V statistics is close to 0.46 whereas this contingency coefficient is giving 0.56 and so on. But definitely they are indicating, yes, they that there is a moderate

Association. So, this is how we go. Now, definitely the question comes how to decide whether this is a really low moderate or a strong? for that you have to use your own judgment and this type of power to judge that you can very easily develop by practice. So, I would request you please try to take some more example and try to practice it. Now, I would like to stop with this topic of measuring the association between different types of variable either they are continuous, ranked observation or discounting observation, and you also have learnt that how to compute them on the basis of the given software? So, the main thing is this if you understand the concept? it is easy to compute them but, the main thing come how to interpret them. So, the main objective which you have to emphasize here is this how to choose the right tool and how to compute it correctly and then how to take the correct statistical inferences out of that so you practice and learn this technique, develop this technique and I will see you in the next lecture till then good bye.