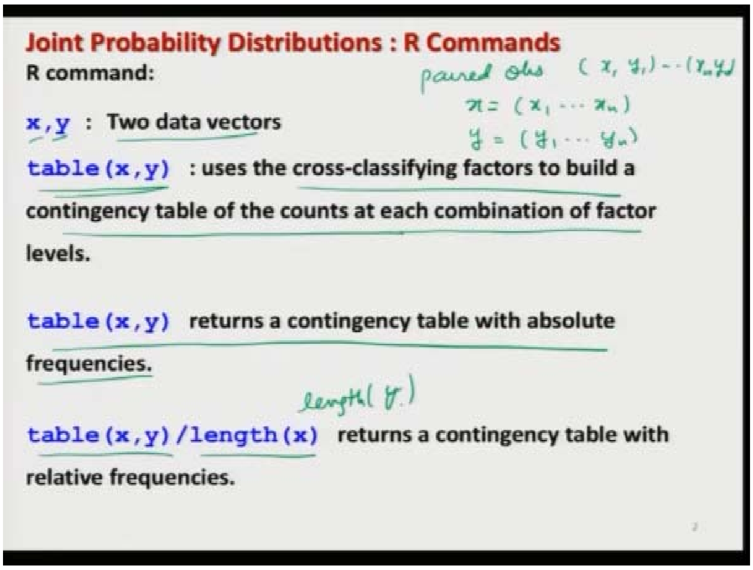
Essentials of Data Science with R Software- 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur
Lecture 44
Bivariate Probability Distribution in R software

Hello friends, welcome to the course Essential of Data Science with R Software- 1, in which we are trying to understand the basic concepts of probability theory and statistical inference. So, you can recall that in the last lecture we had understood various concepts related to the bivariate probability mass function and we have taken an example and then we computed the joint probability, marginal probabilities, conditional probabilities etcetera.

And, yes, that was a pretty long lecture because it takes a longer time to understand the concept, once you understand then it becomes very easy to implement it, that is the rule of the nature. So, on the same lines following the same rules of the nature this lecture is going to be very brief, very short, why? We are going to do the same thing, but we are going to implement it them in the R software and you will see that it is very easy to compute such distributions and probabilities in the R software. So, let us begin our lecture and try to compute the joint probability distribution, joint probability mass function their respective probabilities and distributions in the R software.

(Refer Slide Time: 1:34)



Joint Probability Distributions : R Commands

R command:

paired obs $(x_1, y_1) \cdots (x_n, y_n)$

$x = (x_1 \cdots x_n)$

$y = (y_1 \cdots y_n)$

`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.

length(y)

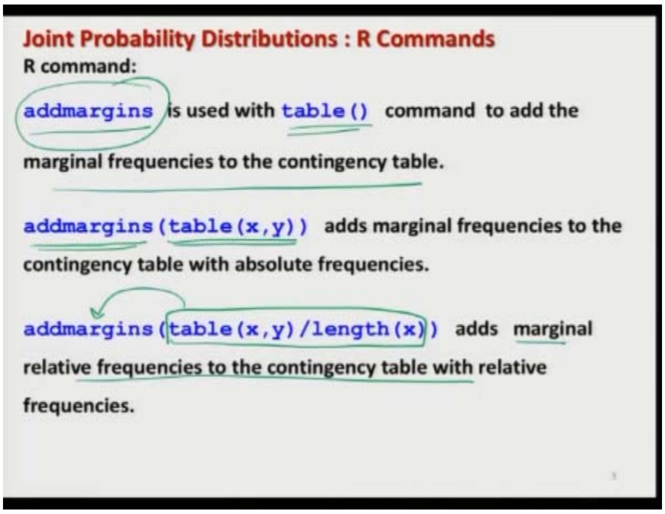
`table(x,y)/length(x)` returns a contingency table with relative frequencies.

We begin, so in R software if there are two random variables, they are, suppose we are trying to indicate them by x and y. So, these are the small x and small y they are indicating the two data vectors on x and y. Remember, one thing x and y they will be the paired observations, something like say (x_1, y_1),…, up to here (x_n, y_n), so what we have done we have collected all the observations on x x_1, x_2,…, x_n in one data vector and all observations of y on one data vector y_1, y_2,…, y_n and we are calling them as x and y.

Now, the command table(x , y), this will create a table which will be cross classifying the factors and it will generate a contingency table and this is going to count that for each of the combinations of the factor levels. The same way as you have considered the example of mathematics, biology with respect to the gender male and female. So, the type of counting what we did manually it will conduct that type of counting automatically.

So, this command actual table x, y will return a contingency table with absolute frequencies. Now, what you want? You want to have the relative frequency because they are going to represent the probabilities, so we try to divide this command table x, y by length of x. So, obviously the number of observations in x and y data vector they are going to be the same, so here actually you can use either length of x or length of y, whatever you want they are they are going to be the same but do not use the length of x comma y otherwise the number of observation will become double.

(Refer Slide Time: 3:31)



Joint Probability Distributions : R Commands

R command:

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

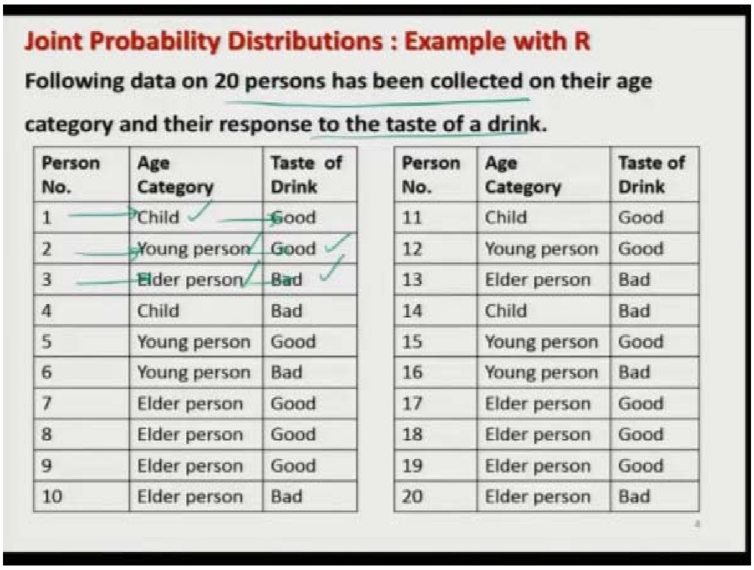
addmargins(table(x,y)) adds marginal frequencies to the contingency table with absolute frequencies.

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

And after this we try to use here a command addmargins and they are used along with that table command and this is going to add the marginal frequencies to the contingency table, you can see that even the name is conveying the same addmargins. So, that means add marginal frequencies and in case if I try to use the earlier command table x, y and if I try to use the command addmargins on this table(x , y) then it is going to add the marginal frequencies in terms of absolute frequencies.

And if I try to consider here the relative frequencies like a table(x , y) divided by length of x and then I try to add the command addmargins over it, then it is going to add the marginal relative frequencies for the contingency table which will have the probabilities in terms of or the frequencies in terms of relative frequencies.

(Refer Slide Time: 4:38)



Joint Probability Distributions : Example with R

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

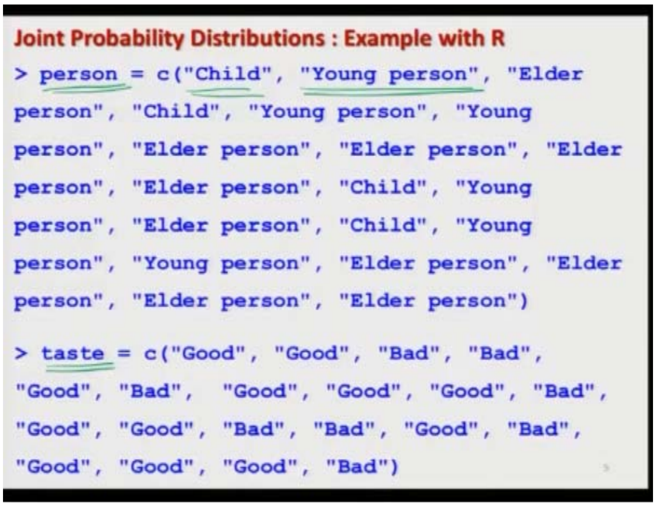
So, let us try to take a very simple example to understand these things and then you will realize that they are not difficult. So, here in this example, I have collected the data on 20 persons, and I have collected the data with respect to their age category and the response to the taste of a drink. You know that all these drinks they are very much age dependent. For example, the cold rings they are more preferable among the youngster then the older people and say here like a tea, coffee etc. they are more popular among the elders compared to the youngest and so on.

So, now in this example, means I am trying to take here three categories child, young person and elder person. Remember one thing that I have taken the example in the last lecture only on the bivariate data but I had explained you that these things can be extended without any problem to more than two categories also, I had given you very brief introduction and that was intentional because I believe that you are going to make such computations only on the software, so I will try to show you here in this software that instead of taking a variable with only two categories, I am trying to take here the variables into three categories.

So, one variable will have three categories and another variable will have only two categories, so that will give you an idea that how are you going to create such contingency table for any number of categories or say for any number of variables whatever you want to say.

So, now you can see here I have here three categories- child, young person and elder person and the categories for the taste of the drink are only two- good or bad, so now you can read the data person number one is a child and who said that the drink is good and taste, the second person is a young person and that person also said the taste of the drink is good, the third person is then elder person who said that the drink is bad and so on we have collected the 20 observation, well you can ask me why I have taken only 20 observation because I have only this much space in my slide to show you clearly, if I try to make your hundred observations, I cannot explain you clearly.

(Refer Slide Time: 7:07)



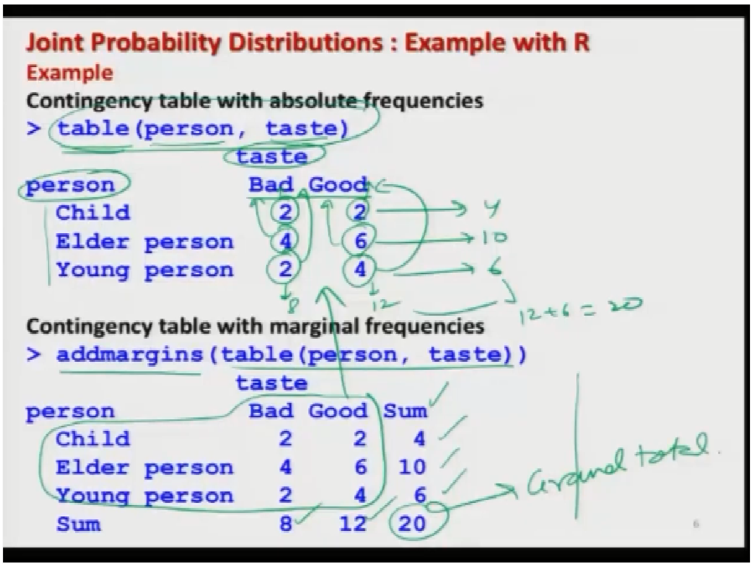
Joint Probability Distributions : Example with R

```
> 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")
```

Now, I try to create two data vectors for this sample of data, one let me call this as a person in which I try to collect the data whether the person is child, young or elder and second one is here is taste. So, the taste is going to be good or bad and so on and you can see here that I am trying to give this data inside the double quote because these are some character strings. So, I try to give this data, well I expect that when you are trying to get some data from some outside for the analysis this data will already be available in the tabular format.

(Refer Slide Time: 7:48)



Joint Probability Distributions : Example with R

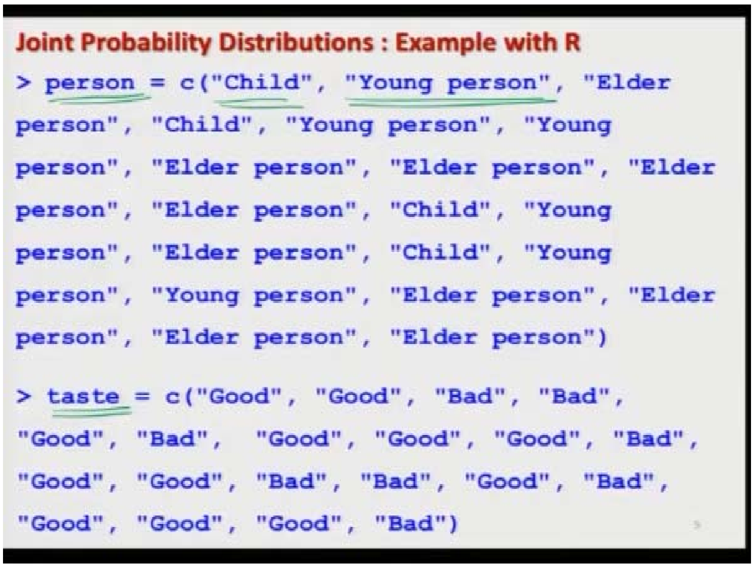
Example

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
```



Joint Probability Distributions : Example with R

```
> 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")
```

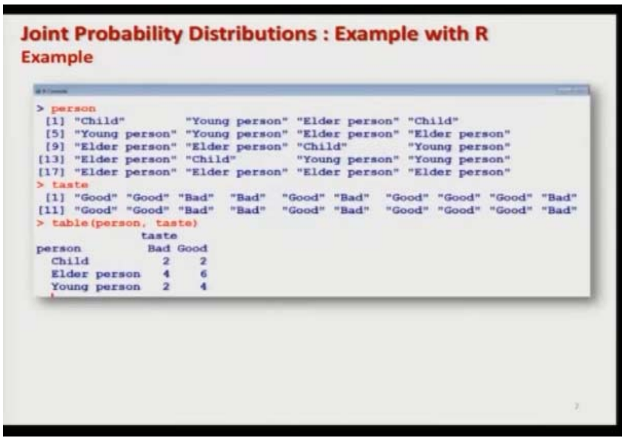
Now, I try to create here a contingency table with the absolute frequencies, I simply try to take here the command table(person, taste) and now you can see what is happening. Here the data that you have given it is like a child, young person, elder person, good and bad but this R has automatically converted them and counted them that what are the total number of persons in a particular category.

So, the R automatically has computed the absolute frequencies. For example, it is saying that there are two variables here one is here taste and another here is person and for taste there are two categories bad and good and in the person there are three categories child, elder person, young person and the total number of people who are trying to give the taste as good or bad is described in this contingency table. For example, there are two children who are saying the taste is bad, there are two children who are saying that the taste is good.

Similarly, there are four elder person who are saying the taste to be bad, there are six person who are saying that the taste was good and there are two young persons who are saying the taste is bad and there are four young persons who are saying the taste is good right. So, this is the continuous table, now you want to find out here the marginal frequencies. So, what you have to do? You can do it here manually, this will come here 2 plus 2 is equal to 4, 4 plus 6 equals to 10, 2 plus 4 is equal to 6 and then if you try to go with the column wise.

So, column by sum is 2 plus 4 plus 2 which is 8 and 2 plus 6 plus 4 this is 12, and if you try to take the grand total this will be 12 plus 6 is equal to 20 and now the same thing can be obtained by the command here addmargins on this same command table, percentage and you can see here the outcome, you can see here this outcome is the same as earlier which was there and you can see here another row and columns have been added here as a sum and you can see here this is 4, 10, 6 and then here 8, 12 and this is here the grand total of the total numbers in the rows and columns, so this is here 20, so you can see the same table can be obtained so easily in the R software.

(Refer Slide Time: 10:22)



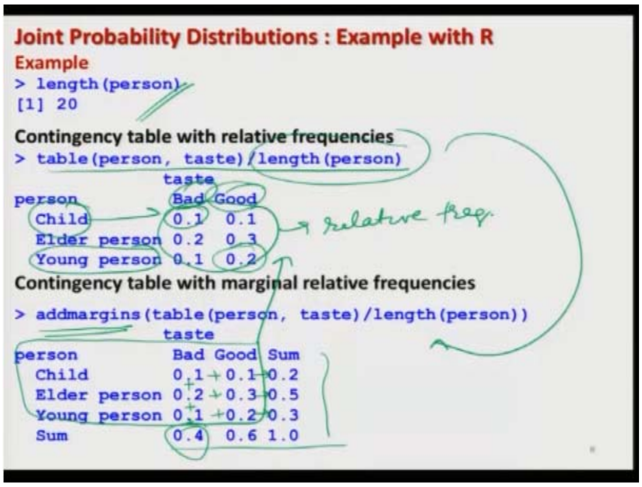
Joint Probability Distributions : Example with R

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
```

And this is here the outcome on the screenshot of the outcome on the R console.

(Refer Slide Time: 10:32)



Joint Probability Distributions : Example with R

Example

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

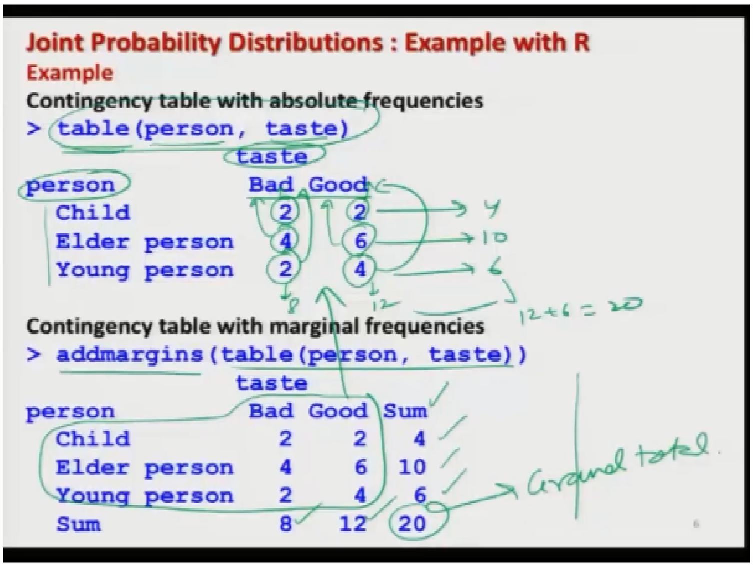
Contingency table with relative frequencies

```
> 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
```

relative freq.

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
```



Joint Probability Distributions : Example with R

Example

Contingency table with absolute frequencies

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

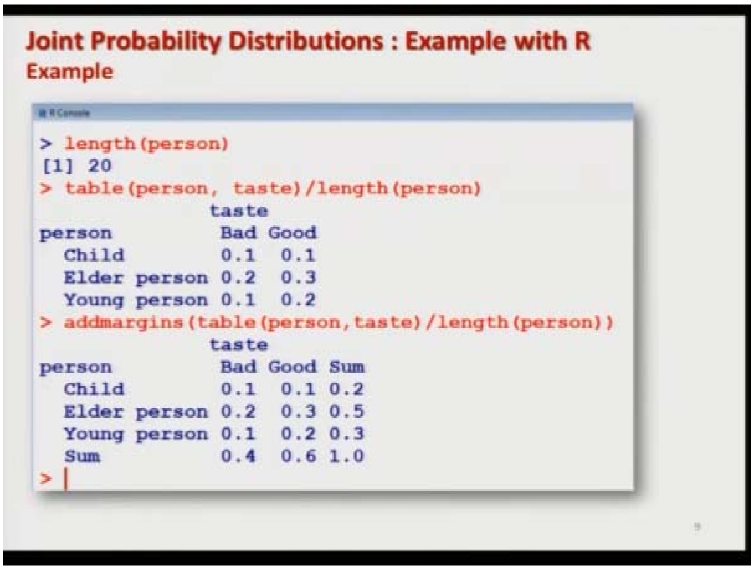
Contingency table with marginal frequencies

```
> 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
```

But before going into the R console let me try to show you these calculations with respect to the relative frequencies also so you can find out here the length of person is coming out to be 20 and now if you try to find out the contingency table with respect to the relative frequency, so you have to use the same command, you simply have to divide it by length of person. So, every observation which you have obtained here every value of absolute frequency that is here that is going to be divided by 20 and you can obtain here these values, so these are the values of relative frequencies and the interpretation is exactly in the same way as we did earlier.

The relative frequency of the number of children who are saying that the taste is bad is 0.1, the number of young persons who are saying the taste is good this relative frequency is 0.2 and so on. Now, in case if you want to find out here the marginal relative frequencies what you have to do is you have to simply use the command addmargins in the same command which you have used here and then it will give, you can see here this part is the same as earlier and this is here the new row and column which has been added as sum and you can see here this is 0.1 plus 0.1 is 0.2, 0.2 plus 0.3 is 0.5, 0.1 plus 0.2 is 0.3 and the 0.1 plus 0.2 plus 0.1 this is 0.4 and so on. So, you can see here it is not difficult at all.

(Refer Slide Time: 12:06)



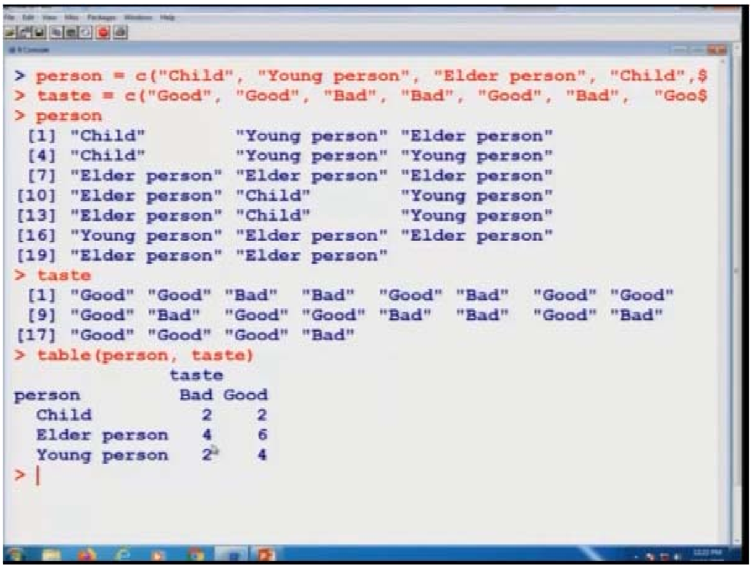
Joint Probability Distributions : Example with R

Example

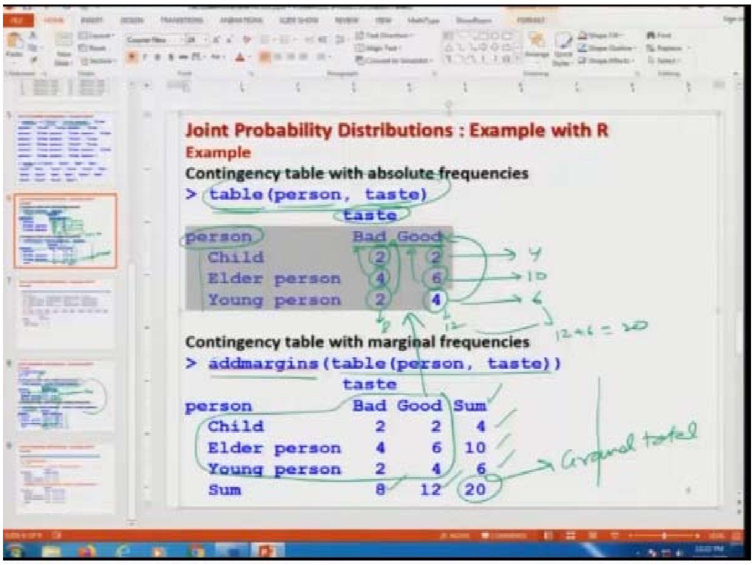
```
> length(person)
[1] 20
> table(person, taste)/length(person)
              taste
person         Bad Good
  Child        0.1  0.1
  Elder person 0.2  0.3
  Young person 0.1  0.2
> addmargins(table(person,taste)/length(person))
              taste
person         Bad Good Sum
  Child        0.1  0.1 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 if you try to see this is the screenshot of the same operation which I have just shown you.

(Refer Slide Time: 12:15)



```
> person = c("Child", "Young person", "Elder person", "Child",$
> taste = c("Good", "Good", "Bad", "Bad", "Good", "Bad",  "Goo$
> 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" "Good"
 [9] "Good" "Bad"  "Good" "Good" "Bad"  "Bad"  "Good" "Bad"
[17] "Good" "Good" "Good" "Bad"
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
> |
```



Joint Probability Distributions : Example with R

Example

Contingency table with absolute frequencies

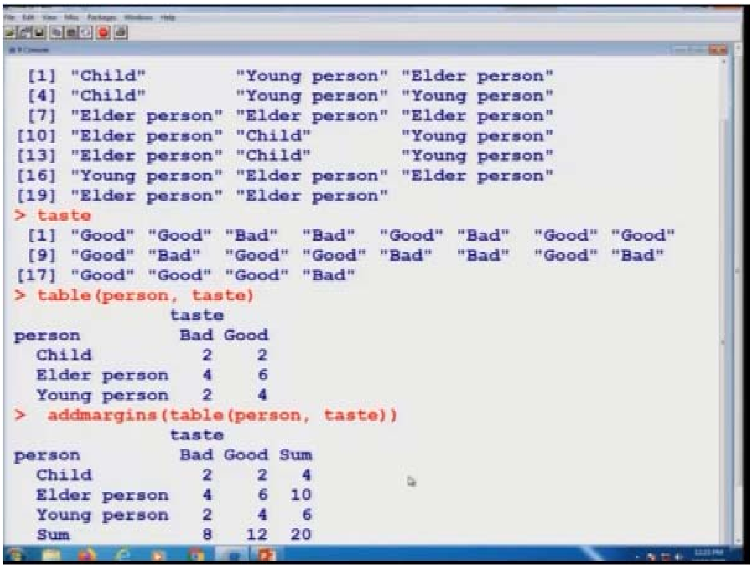
```
> table(person, taste)
                taste
person          Bad Good
  Child           2    2
  Elder person    4    6
  Young person    2    4
```

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
```

Now, I will try to show you these operations on the R console. So, let me try to first insert this data, so far I will simply copy it to save the time, you can see here this is here like this, and this data on the here taste you can see here this I am entering here as a taste, and now I try to execute here this command here table, person, taste. So, you can see here this is your here data person and here taste and the table is coming out to be here like this, and so you can see here this is the same outcome which you can see here from here to here you can see.

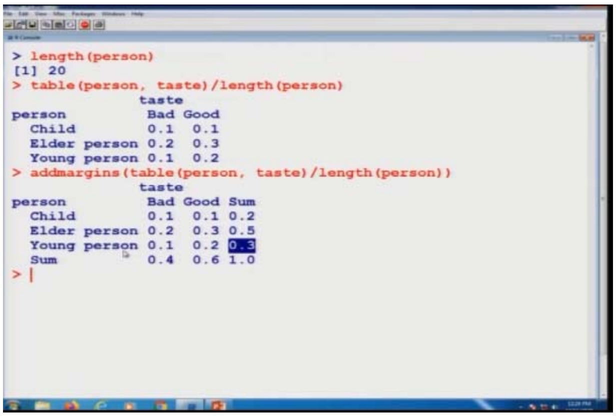
(Refer Slide Time: 13:13)



```
 [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" "Good"
 [9] "Good" "Bad"  "Good" "Good" "Bad"  "Bad"  "Good" "Bad"
[17] "Good" "Good" "Good" "Bad"
> table(person, taste)
              taste
person         Bad Good
  Child          2    2
  Elder person   4    6
  Young person   2    4
>  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 in case if you want to find out the marginal frequencies in terms of absolute frequencies. So, I can simply see here ctrl d and you can see here now in this outcome this sum has been added 4, 10, 6 and 8 and 12. So, you can see here it is not difficult at all.

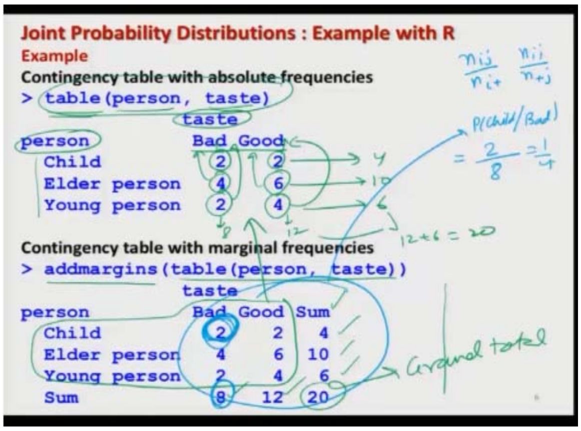
(Refer Slide Time: 13:36)



```
> 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
> |
```

Now, in case if you want to find out the same thing in terms of relative frequencies, so you can see here that the length of person data vector this will come out to be 20 here and if you try to find out the contingency table in terms of relative frequency, this is just the outcome of the table which is now divided by the length of the data vector and if you want to compute here the marginal relative frequency, so you can see here this is coming out of here like this, only the row and columns are added representing the row sums and column sums.

(Refer Slide Time: 14:20)



Joint Probability Distributions : Example with R

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
```

$\frac{n_{ij}}{n_{i+}}$ $\frac{n_{ij}}{n_{+j}}$
P(Child/Bad)
$= \frac{2}{8} = \frac{1}{4}$
4
10
6
8
12
12+6 = 20
Grand total

And now if I come back here and if you want to find out here the conditional probability you are here do you think that is it going to be a very difficult task? The conditional probability was defined here as for example, if I try to take here the example of this absolute frequency this was defined as a n_{ij} upon n_{i+} or this was n_{ij} plus here n_{+j} and so on, so if you try to see here, if you try to look into this thing, this will try to give you here the conditional probability of observing a child which is trying to say that the taste is bad.

So, this will become here this value 2 and divided by here the total number of people who said that the taste is bad this is given by here number 8, so this will become here 2 upon 8 1 upon 4, you can see here that this is very easy to compute in a real data set. So, now we come to an end to this lecture and as I promised you that this lecture is going to be very quick because that we already have covered most of the concepts in the last lecture, so in this lecture you had to only understand that how are you going to implement it, but it was more important for me to first explain you what are you trying to compute and how are you going to interpret it.

So, now you see handling the bivariate data in real life is not difficult, the only part is how are you going to make interpretation and believe me there can be very interesting observations which can be taken out from such joint probability distribution and those conclusions are very helpful in real life, whenever somebody wants to know about the process, they are really going to help us.

So, I will say now request you that try to take some data and try to practice it and try to make different types of conclusions. For example, if you want, I can extend this lecture to even two hours of duration just by giving you different types of conclusion, different types of interpretation and believe me these things are very useful when you are going to work in a bigger data sets in a data science topics.

Now, before I conclude, let me try to tell you here one thing more, you had seen that when we considered the discrete and continuous random variables, we had understood their concepts and when we understood their basic fundamentals we had observed one thing, that when we are trying to deal with the discrete random variable we have to use the summation to find out different types of probabilities and in the case of continuous we had used the integration to find out different types of probabilities.

So, now in case if I ask you to extend these concepts of the bivariate distribution to a continuous random variable case, then do you think that is it going to be difficult? Certainly not number one, number two when I come to the real-life examples, when you are trying to get that data, data is always in the form of some discrete numbers. So, whether you have a variable which is discrete or continuous your data is going to finally have a numerical value to indicate the process.

So, now even if you are trying to consider the continuous random variables and you are trying to define different types of joint probability distributions your data is going to be ultimately in the form of a numbers. So, whatever you have learnt here, we have learnt it under the topic of discrete random variable, but the same methodology is going to be used even when you are trying to get the numbers for a continuous random variable.

So, in the next turn I will try to simply extend or translate the concept which I had learnt in the earlier chapter when we discussed the discrete random variable to a continuous random variable setup, so it is very important that before you try to attend the next lecture, both this lecture, the current lecture and the earlier lecture should be well revised, so I request you to please do it and then come to the next lecture till then good bye.