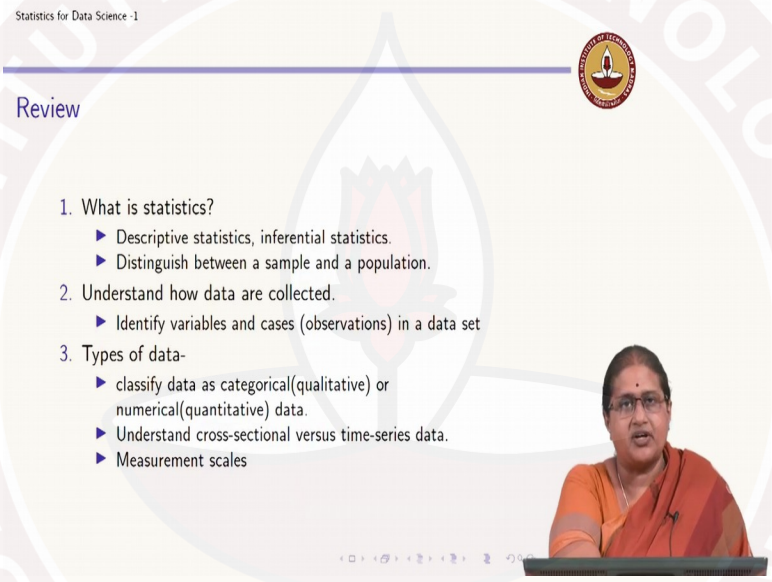

# IIT Madras

## ONLINE DEGREE

**Statistics for Data Science - 1**
**Prof. Usha Mohan**
**Department of Management Studies**
**Indian Institute of Technology, Madras**
**Lecture No. 2.1**
**Describing Categorical Data – Frequency distributions**

In the last lecture, you were introduced to basically the two important branches of statistics, which are descriptive statistics and inferential statistics. You also understood what is the difference between a sample and population.

(Refer Slide Time: 00:32)



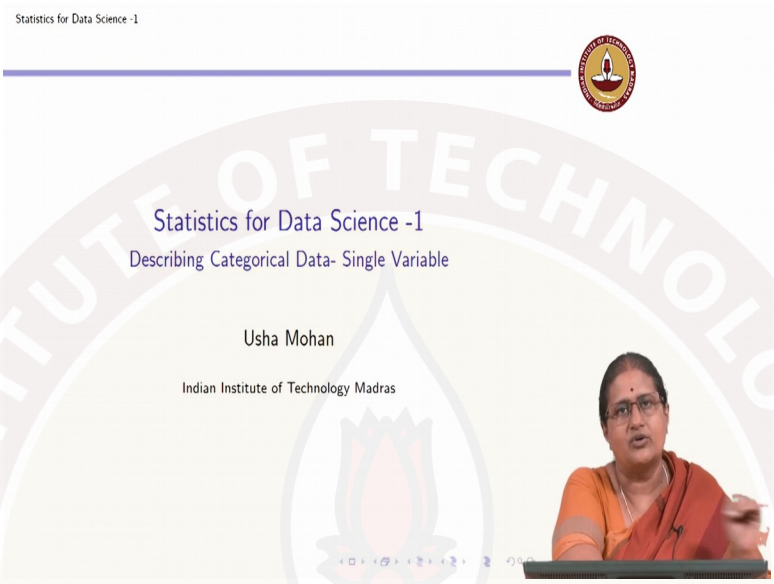

But we restricted our introduction to what is a sample and population for what is required for this particular course. Of course, in the advanced courses you will learn more about sample and population. Then we went on to identify and understand what is a data. Again for this purpose of this course we are restricting ourselves to only structured data.

In the structured data, I have data in the form of a table where the variables are recorded in a column and the observations are cases which are represented by the rows in a data set. We looked at the types of data, broadly you classify data as categorical data and numerical data. We also understood what is the difference between cross-sectional data and time series data.

And finally, you should be knowing by this time, what are the measurement scales by that I mean that when you have categorical data, you should know how to distinguish between whether

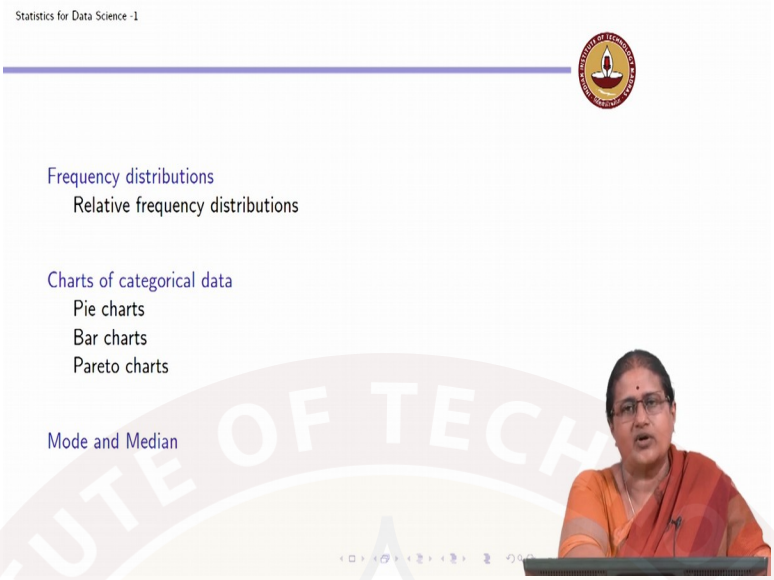
it is a nominal variable or an ordinal variable and when it comes to numerical data, you should know whether it is an interval variable or ratio variable. This is what you should be knowing at this point of time.

(Refer Slide Time: 01:50)



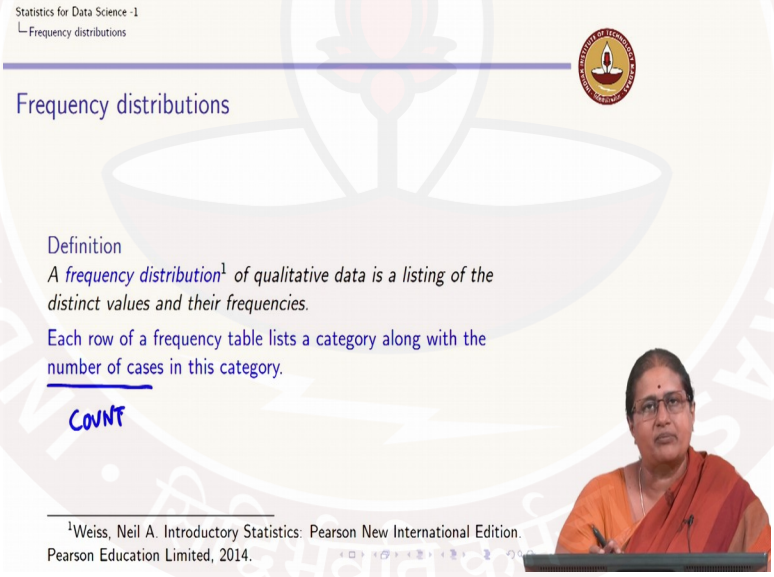

Moving forward, you are going to understand how to describe categorical data. Now, again in this module you are going to understand first, we start with describing categorical data for a single variable, and then we will look at measures of association when they have more than one variable. So, today what we are going to do is, we are going to understand how to describe categorical data.

(Refer Slide Time: 2:13)



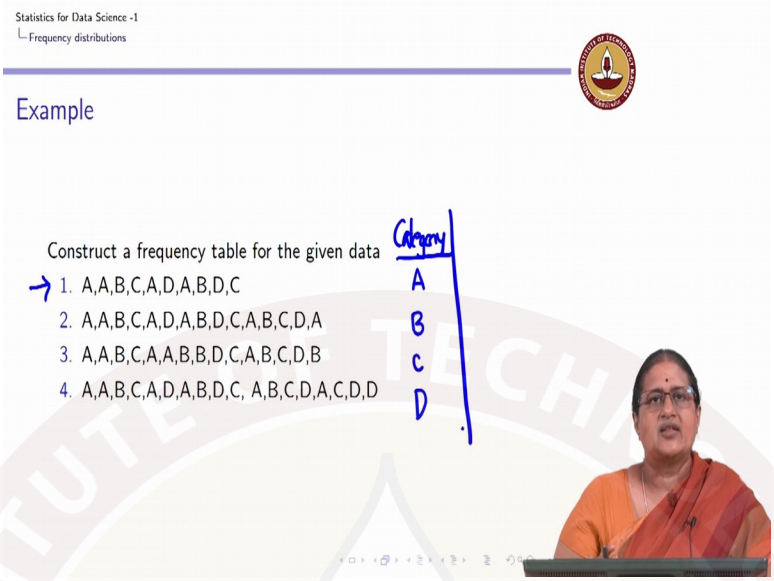

So, we start with what we understand as a frequency distribution.

(Refer Slide Time: 02:20)



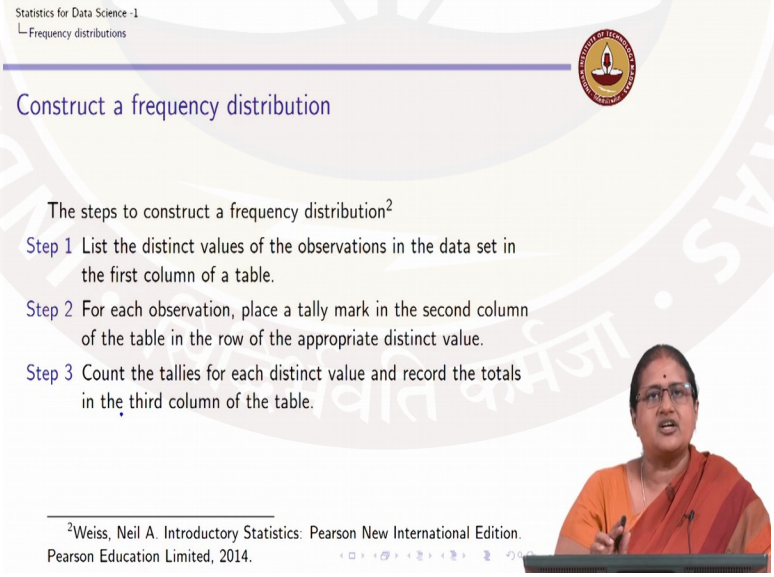

Now, the definition of frequency definition is a listing of distinct values and their frequencies. What do we mean by frequency? Frequency is nothing but the count, nothing but count. And by distinct values, you mean what are the distinct values the categorical variable actually takes. This is what we mean by that. Now, each row of a frequency table lists a category along with the number of cases or count of cases. The minute I say the number of cases, I am just, I imply how many of that particular case is there in that particular category.

(Refer Slide Time: 03:14)



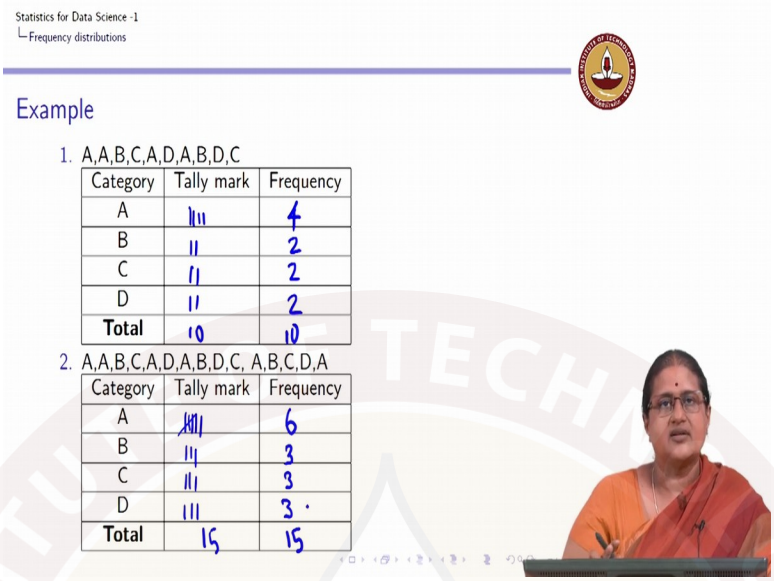

So, for example, let's look at constructing frequency table for this simple given data. The category is just alphabets. I have the alphabets, I have 4 categories, I can term them as A, B, C, and D.

(Refer Slide Time: 3:34)




So, how do I construct a frequency distribution? List the distinct values of observation. Here, what are the distinct values of my observation? In the first example, my distinct values of observation are A, B, C, and D. I list them and I write it as a category. This is the first thing I do. That is my step one. Step one is to list the distinct values and this is my first column.

(Refer Slide Time: 4:06)




So you can see that in the first column, I listed the distinct values which are A, B, C, and D. For each observation, place a tally mark in the second column. So for A, I place a tally mark, which I am marking here, the second observation again is A, I again have a tally mark, third observation is B, fourth observation is a C, fifth observation is again A, sixth observation is a D, seventh observation is an A, then I have a B, then I have a D, I have a C.

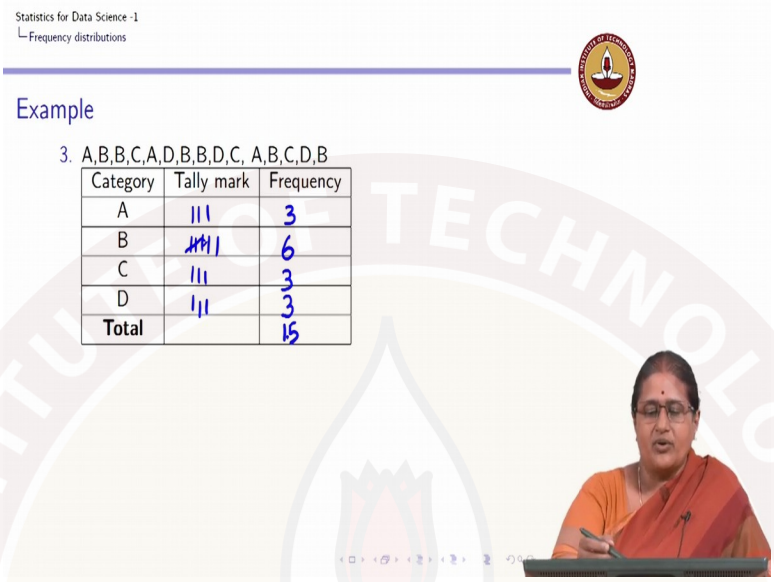
So, these are the tally marks which I am going to have. So, that is step two. For each observation, place a tally mark and then count the number of tallies. So, if I count the number of tallies, the frequency, or the count of A is 4, count of B is 2, count of C is 2, and count of D is 2. This is what I refer to as a frequency distribution, where the distinct values are given in column 1, tally marks in column 2, and the count in column 3.

Now let's go and look at this example. If you look at this example, again I have a tally mark. I will do it quietly, faster at this time, A, A, B, C, A, D, A, B, D, C. Now this A, I just cross it out, because this is my fifth value. Whenever I have more than 4, I am having a tally mark. I am crossing it out as the fifth value, then I have a B, C, D, and A. So, this 5 + 1 gives me a value of 6, this is a 3, this is a 3, this is a 3.

So, I have a total of 15 observations in this case, where this is the distribution, category A occurs six times, B occurs 3 times, C occurs three times, and D occurs three times. So, if you look at

this data where A had a tally mark of 4, B had 2, C had 2, and D had 2 with a total 10. This was a 4, 2, 2, 2. This is a 6, 3, 3, 3.

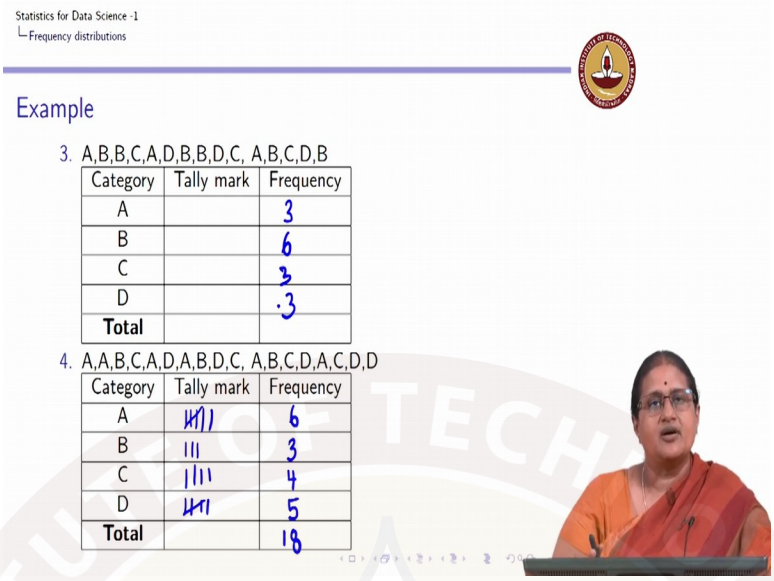
(Refer Slide Time: 06:40)




Now, let's look at this here again. The same data of 15 points, but now I have a A, I have a B, B, I have a C, I have a A, I have a D. I have a B, B, D, C, a A, a B, a C, a D, a B. So, this has A is appearing 3 times, B is appearing 6 times, C is appearing 3, and D is appearing 3. I have again a total of 15 observations. So, if you look at compare this example with the earlier example, you see that the only difference between example 2 and example 3 is, A appear 6 times here and B appears three times, C and D appear 3 times each.

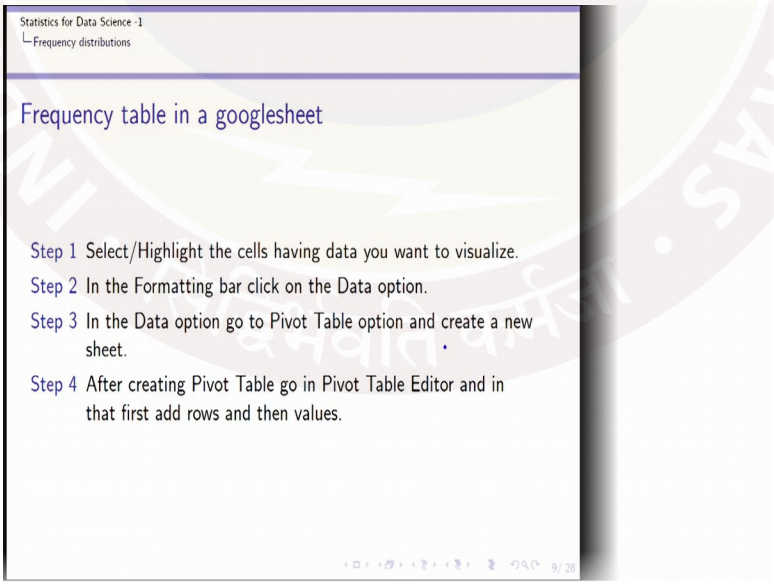
In this example, it has flipped between A and B with A appearing thrice, B appearing 6 times, and C and D appearing 3 times each. Both of them have 15 observations, and this is what is given.

(Refer Slide Time: 7:48)



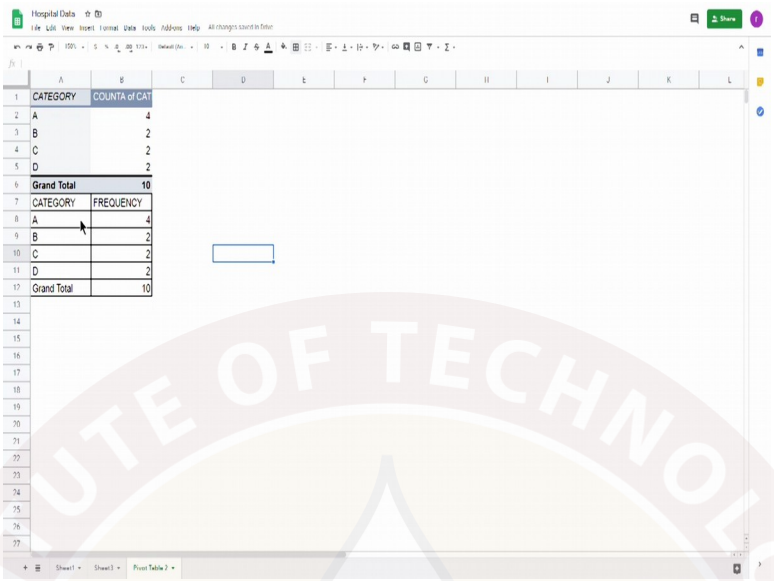

Now, let's look at a final example where I have so many observations. Again if I do a tally mark A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D, D. So if I look at this, I have a 6, I have a 3, I have a 4, I have a 5. So, I have 18 observations here. Now, if you look at this, I have a tally mark here. I had a 3, I had a 6, I had a, then I had 1, 2, 3, I had a 3. This is what I have here. So, you can see that we can construct different frequency distributions.

(Refer Slide Time: 8:48)




How do I do a frequency table in a Google Sheet? So, let's look at how to do a frequency table in a Google Sheet. I look at the same example I have taken.

(Refer Slide Time: 9:00)




So this is the Google Sheet I have. I am going to construct, I am going to add a sheet here. In that sheet, I am going to type first the category name. So, I have a category here. I am going to write down whatever data I have here, the data, I am going to list on the data. I have a A, I have A, B, C, A, D, A, B, D, C. So, you can see that this is the first example which we look, I have listed down the data here.

So, you go back to the step one, select highlight the cells you have, what are the cells I have, I just have these cells, I am highlighting these cells. That is the first step. Now you look at the second step in the formatting bar. Click on the data option. So, I go to the formatting bar, I click on the data option. Then, in the data option, go to the pivot table option. Go, first I highlight my data, I go to the data option, I have what I call a pivot table option.

Now, this pivot table I specify the range. You see that the range is specified A one to A eleven with the cell A one specifying what is the category. Here I have just given the name category, I could give any name to this category variable. I could give an alphabet or I could just tell it is some group, anything, but this I am just specifying a certain category, asking it to create a pivot table.

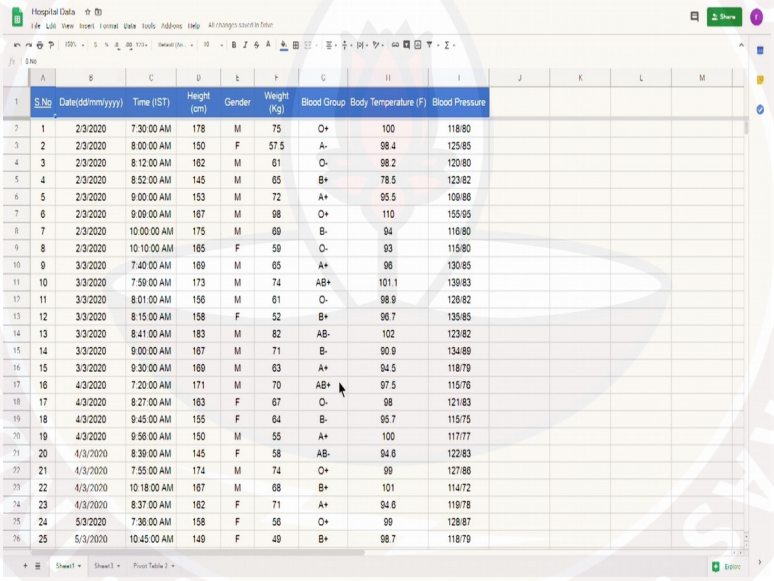
Now, let us go to the final step. After creating the pivot table in the pivot table editor, what is the pivot table editor, you have the pivot table editor which appears on the right-hand side. There you add rows. In the row, I just add a category. What are the different categories? I have

category A, B, C, and D. And in the values, I'm going to add what are the values that category has which is 4, 2, 2, 2.

So, one way to create a table is, I can just copy this and I paste the values. I can give a category table here with frequency here and I can see that this is nothing but the table we have just created. So, this is one way to create a frequency table in your Google Sheet.

So, once you have your frequency table, your frequency table, so this is precisely what we did for the first example 4, 2, 2, 10. You can see that is what our Google Sheet gives, A frequency 4, B frequency 2, C frequency 2, and D frequency 2 with a grand total of 10. This is what we have here, the first frequency table which you have created on a Google Sheet.

(Refer Slide Time: 12:49)



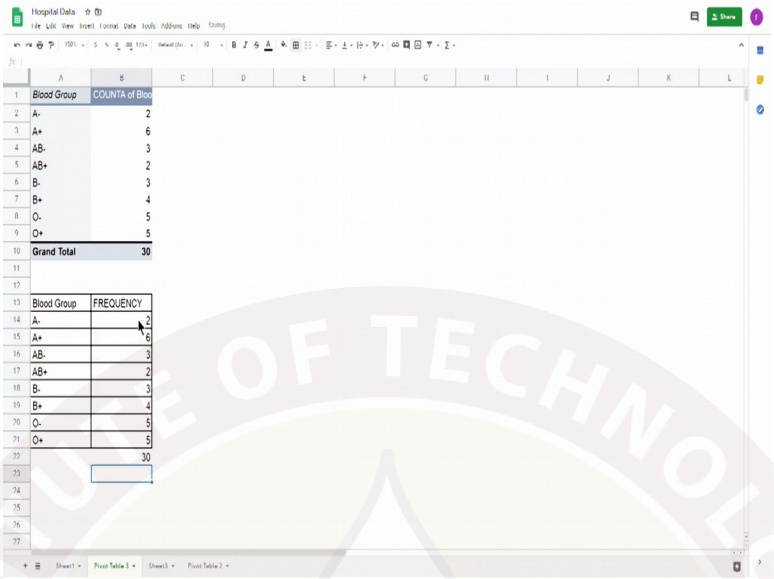
| S.No | Date(dd/mm/yyyy) | Time (IST) | Height (cm) | Gender | Weight (Kg) | Blood Group | Body Temperature (F) | Blood Pressure |
|---|---|---|---|---|---|---|---|---|
| 1 | 2/3/2020 | 7:30:00 AM | 178 | M | 75 | O+ | 100 | 118/80 |
| 2 | 2/3/2020 | 8:00:00 AM | 150 | F | 57.5 | A- | 98.4 | 125/85 |
| 3 | 2/3/2020 | 8:12:00 AM | 162 | M | 61 | O- | 98.2 | 120/80 |
| 4 | 2/3/2020 | 8:52:00 AM | 145 | M | 65 | B+ | 78.5 | 123/82 |
| 5 | 2/3/2020 | 9:00:00 AM | 153 | M | 72 | A+ | 95.5 | 109/86 |
| 6 | 2/3/2020 | 9:09:00 AM | 167 | M | 98 | O+ | 110 | 155/95 |
| 7 | 2/3/2020 | 10:00:00 AM | 175 | M | 69 | B- | 94 | 116/80 |
| 8 | 2/3/2020 | 10:10:00 AM | 165 | F | 59 | O- | 93 | 115/80 |
| 9 | 3/3/2020 | 7:40:00 AM | 169 | M | 65 | A+ | 96 | 130/85 |
| 10 | 3/3/2020 | 7:59:00 AM | 173 | M | 74 | AB+ | 101.1 | 139/83 |
| 11 | 3/3/2020 | 8:01:00 AM | 156 | M | 61 | O- | 98.9 | 126/82 |
| 12 | 3/3/2020 | 8:15:00 AM | 158 | F | 52 | B+ | 96.7 | 135/85 |
| 13 | 3/3/2020 | 8:41:00 AM | 183 | M | 82 | AB- | 102 | 123/82 |
| 14 | 3/3/2020 | 9:00:00 AM | 167 | M | 71 | B- | 90.9 | 134/89 |
| 15 | 3/3/2020 | 9:30:00 AM | 169 | M | 63 | A+ | 94.5 | 118/79 |
| 16 | 4/3/2020 | 7:20:00 AM | 171 | M | 70 | AB+ | 97.5 | 115/76 |
| 17 | 4/3/2020 | 8:27:00 AM | 163 | F | 67 | O- | 98 | 121/83 |
| 18 | 4/3/2020 | 9:45:00 AM | 155 | F | 64 | B- | 95.7 | 115/75 |
| 19 | 4/3/2020 | 9:56:00 AM | 150 | M | 55 | A+ | 100 | 117/77 |
| 20 | 4/3/2020 | 8:39:00 AM | 145 | F | 58 | AB- | 94.6 | 122/83 |
| 21 | 4/3/2020 | 7:55:00 AM | 174 | M | 74 | O+ | 99 | 127/86 |
| 22 | 4/3/2020 | 10:18:00 AM | 167 | M | 68 | B+ | 101 | 114/72 |
| 23 | 4/3/2020 | 8:37:00 AM | 162 | F | 71 | A+ | 94.6 | 119/78 |
| 24 | 5/3/2020 | 7:36:00 AM | 158 | F | 56 | O+ | 99 | 128/87 |
| 25 | 5/3/2020 | 10:45:00 AM | 149 | F | 49 | B+ | 98.7 | 118/79 |

Now again, I can create the same table for any given data. If you recall, this is the blood group data, hospital data which we discussed in our earlier class. Suppose I want to know the distribution of a particular categorical variable here, I can look at two categorical variables here. One is gender and other is blood group. When I look at blood group, this is what I have here. This is the categorical variable.

So, I go back to my pivot table step. So, what do I do, I remember, in first I select the cells, I click on the data option, I go to the pivot table option and I go to the pivot table editor. So, we

are going to do the same thing here. I select this data, I click on data option. Here I click on the pivot table option.

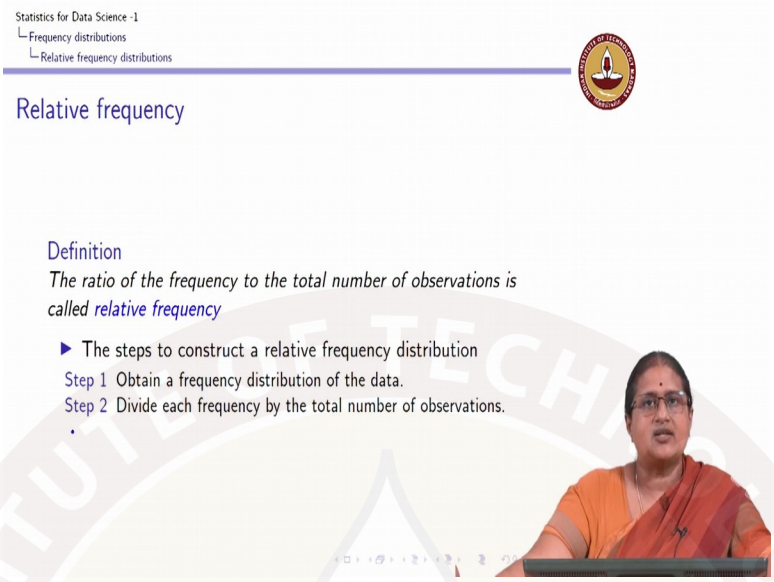
(Refer Slide Time: 13:47




I create a new sheet. Once I create a new sheet, I go to the pivot table editor. I add the rows. Now you can see the name of the variable, now here is a blood group. That is what I want to know the frequency distribution of the variable blood group. So, I click on blood group here, and I go to values and I add the values and you can see that this is the frequency distribution of the blood group. I can just copy, I can paste the values. And I can just put here, this is the blood group, and this is the frequency.

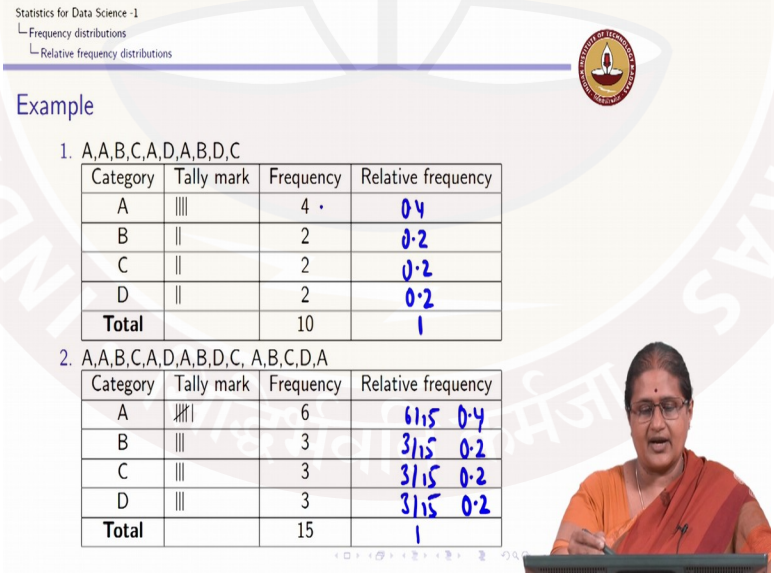
And you can see that this is the frequency distribution of table of my blood group, which I get in Google Sheets. If you look at the sum, you can see the sum of all of those study and that is precisely I have 30 observations, this is blood group with a frequency of 30 people. So, this is how we construct frequency tables both manually that is through first principles of using tally marks and this is through using a Google Sheet. Frequency table gives the count of each variable, each categorical variable.

(Refer Slide Time: 15:43)




There is another thing which is very useful and that is called relative frequency. What relative frequency captures is the ratio of the frequency to the total number of observations.

(Refer Slide Time: 15:59)



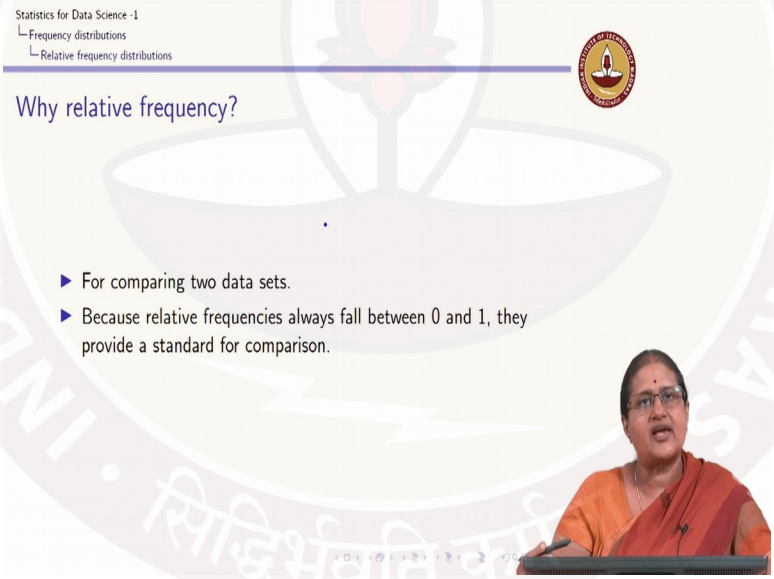

So, we already have constructed a frequency table, we already saw that there are 4 A, 2 B, 2 C, and 2 D. The ratio of the frequency, so, 4 is the frequency of A, total number of observation is 10. The ratio that is 4/10, which is 0.4, gives me the relative frequency of A in this table. So,

relative frequency distribution, I just divide each frequency by the total number of observations, and I get a 0.4, 0.2, 0.2, 0.2. And the sum total of a relative frequency should always add up to 1.

Now, if you look at this data, this was a 0.4, 0.2, 0.2, and 0.2 and adding up to 1. Here you look at it, this is going to be 6/15, this is going to be 3/15, this is going to be a 3/15, this is going to be a 3/15. And you can see that all of them add up to 15/15, which is 1. You can again see that this is a 0.4, this is 0.2, this is 0.2, this is 0.2.

Now, what I want you to see is, the frequency of these two distributions are different. Here, I had an A which is 4, 2, 2, 2. Here the frequency was 6, 3, 3, 3. Whereas when you look at the relative frequency of this dataset and this dataset, you can find that the relative frequency of A in this dataset, the relative frequency of each one of the categories or each one of the variables A, B, C, and D is the same as the relative frequency of A, B, C and D in this dataset.

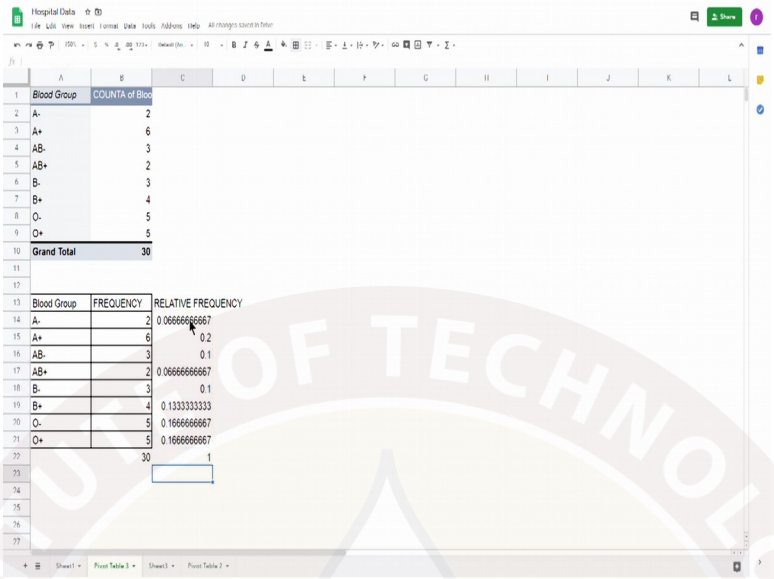
(Refer Slide Time: 18:07)




Now, the reason for why do we need relative frequency, as I have demonstrated here even though there is a difference between these two datasets in the count, you can see the relative frequency is pretty much the same. Here, I had totally 10 observations, here I had totally 15 observations, but both of them have the same relative frequency, the frequency is different.

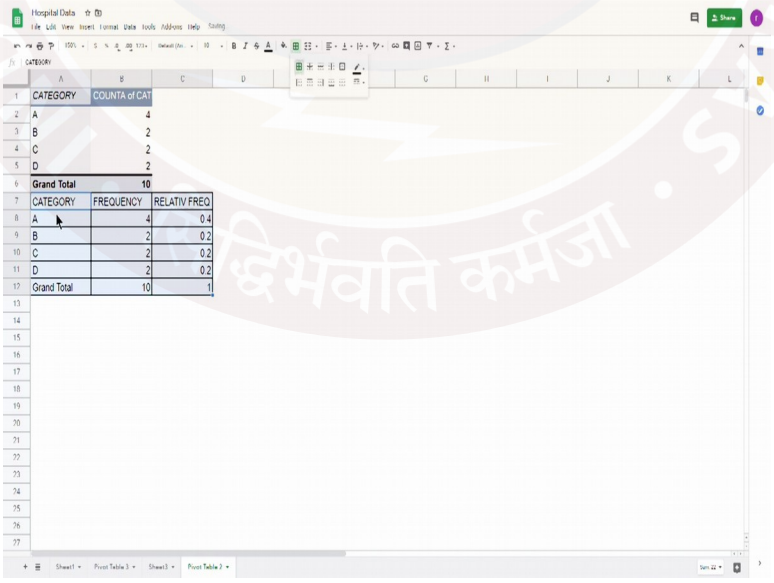
So, what relative frequency helps us, is to compare two datasets. And because relative frequency always is between 0 and 1, it is a good standard for comparison. Hence, we always prefer to have a relative frequency table.

(Refer Slide Time: 19:06)



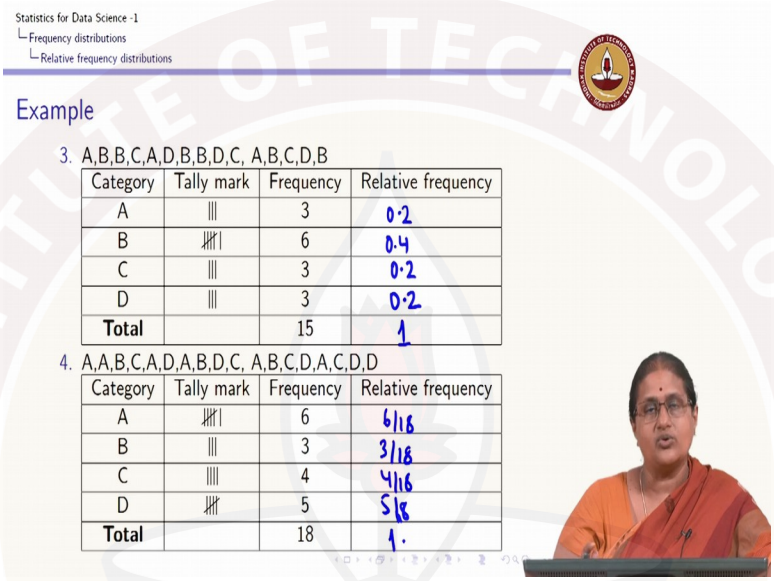

How do we create a relative frequency table in Google Sheet? In Google Sheet, we already have a frequency table. I create what is called a relative frequency column, here, and I know that relative frequency is nothing but the frequency divided by the total. That is how I define it. And I can just drag this down. And you can see that, if I look at the sum of these values, it adds up to 1 with each of these frequencies giving me the relative frequency. This is for the blood group.

(Refer Slide Time: 20:03)



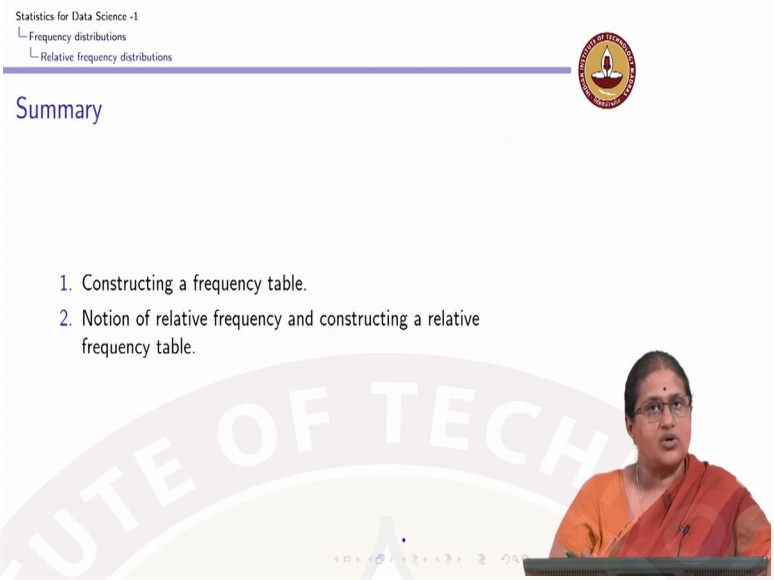
I can do the same thing for this pivot table. I can do the relative frequency, which is equal to just a 0.4, 0.2, 0.2, 0.2, and the sum of all relative frequencies would always add up to 1. So, now I have this, which is going to give me, this is not the blood group, this is category. So, I have for the first example, I have with me what I call the frequency and the relative frequency listed along with the category variable.

(Refer Slide Time: 20:56)




So, I have two more examples. In the earlier examples, so, this is going to be 3/15, which is a 1 by 5, so I have a 0.2, 6 /15, which is 2/5, which is 0.4. I have again 0.2, I have 0.2. So, for each one of them, you can see that it adds up to 1. I leave this as an exercise, but you can see that this is going to be 6 /18, this is going to be 3/18, this is going to be 4/18, and 5/18. You can see that this adds up 5+4, 9, 9+ 9, 18. So this adds up again to 1.

(Refer Slide Time: 21:35)




So in summary, what we have learned in this portion is how to construct a frequency table? What is the notion of relative frequency? And how do you construct a relative frequency table?