


IIT Madras
ONLINE DEGREE

Statistics for Data Science -1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 4.3
Association between Two Categorical Variables - Relative Frequencies


(Refer Slide Time: 00:14)

Statistics for Data Science -1
 Association between categorical variables
 Relative frequencies



Row relative frequencies

	NO	Yes	Row Total
Female	10/44	34/44	44
Male	14/56	42/56	56
Column Total	24	76	100
	24/100	76/100	



So, we have seen how to construct a contingency table or in other words how to summarize the bivariate data where both my variables are categorical in nature as a contingency table. Now we introduce a very important concept, the concept of a relative frequency. We introduce these concepts when we looked at single categorical variable. Recall relative frequency was nothing, but your frequency by total number of observations.

Relative frequency is what we refer to when we talked about categorical variables as frequency into total number of observations. We have already introduced a notion of a relative frequency, but what do we have in a contingency table. In a contingency table for example I had female, I had male. I had 44 females and I had 56 males, I had 76 people who owned a cellphone.

I have 24 people who did not own a cellphone this was what was my data. So, we can go back to your data here. We are talking about this data here. So, I have 70 in this dataset you can see that I

had 44, 56 then I have 42 men and 34 women owned cellphone. So, 10 women and I have again 14 men did not own a cellphone. So, this is we call this is the contingency table I have.

Now this is my row total what is my row total 44 females and 56 men 24 and my 76 represent or my column totals you can see that the total of the row totals and the total of the column totals are equal and they add up to 100.

(Refer Slide Time: 2:58)


Statistics for Data Science - I
Association between categorical variables
Relative frequencies

Row relative frequencies

- ▶ What proportion of total participants own a smart phone?
- ▶ What proportion of female participants own a smart phone?

Gender	Own a smartphone		Row total
	No	Yes	
Female	10 / 44	34 / 44	44
Male	14 / 56	42 / 56	56
Column total	24 / 100	76 / 100	100

Row relative frequency: Divide each cell frequency in a row by its row total.



So, now suppose if I am interested in asking a question what is the proportion of total participants who own a phone? Now that is simple total participants are 100 of which 76 people own a phone and 24 people do not own a phone. So, 76 out of 100 people actually own a phone in other words this is the proportion of my total participants who own a phone. Similarly, 24 out of 100 people do not own a phone. So this answer to this question is easily given that the proportion of total participants who own a phone is 76 percentage.

Now let me modify this question a bit and ask that what is the proportion of female participants who own a phone. Now how do we answer this question again if you go back here you can see that there are totally 44 female participants of which 34 participants own a phone. So, the proportion of female participants who own a phone is 34 divided by 44 of the total number of female participants how many female participants we have totally 44 of them I am asking what is the proportion who own a phone 34 by 44.

Similarly, I will have 10 by 44 is the proportion of female participants who do not own a phone. Likewise, 14 by 56 is a proportion of male participants who do not own a phone and 42 by 56 is a proportion of male participants who own a phone. Now what is this 10 by 44? 34 by 44, 14 by 56 and 42 by 56? These are what we refer to as the row relative frequency. What is a row relative frequency? I divide each cell frequency by its row total so I divide this by 44, this by 44, this by 56 and this by 56. Of course for the column this total also 24 by 100 and 76 by 100. So, I know 76% of my total participants owned a phone and 24% of my total participants did not own a phone.

(Refer Slide Time: 05:57)

Statistics for Data Science - I
 Association between categorical variables
 Relative frequencies

Example 1: Row relative frequency

Gender	Own a smartphone		Row total
	No	Yes	
Female	10/44	34/44	44
Male	14/56	42/56	56
Column total	24/100	76/100	100

Gender	Own a smartphone		Row total
	No	Yes	
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

And once I do this I see that these are in percentages. 76.27% of the total female participants own a phone that is what this number gives me. 75% of total male participants own a phone, 76% of the total participants own a phone. So, what you have in these cells are what are referred to as the row relative frequencies.

(Refer Slide Time: 06:37)



Example 2: Row relative frequency

Income level	Own a smartphone		Row total
	No	Yes	
High	2/20	18/20	20
Medium	27/66	39/66	66
Low	9/14	5/14	14
Column total	38/100	62/100	100

Income level	Own a smartphone		Row Total
	No	Yes	
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100



Now let us look at the second example. In the second example I do the same thing I know the row totals were 20 for a high income group, 2 people did not own a phone so 2 by 20 which is a 10%, 18 by 20 which is 90%. So, the way I can articulate this is of the high income group or 90% of the high income group people own a phone whereas 10% of the high income group do not own a phone.

In the medium income group 40% or 41% of medium income group do not own a phone 60% own a phone or 59% own a phone and this is 41%, this is 59% own a phone whereas, in the low income group I have a whopping 64% who do not own phones and a 36% who own phones. In total I have 38 people who do not own 38% who do not own and 62% who own a phone. So, these are what we refer to as row relative frequencies.

(Refer Slide Time: 08:04)

सिद्धिर्भवति कर्मजा



Column relative frequencies

- ▶ What proportion of total participants are female?
- ▶ What proportion of smart phone owners are females?

Gender	Own a smartphone		Row total
	No	Yes	
Female	10	34	44
Male	14	42	56
Column total	24	76	100

Column relative frequency: Divide each cell frequency in a column by its column total.



Similarly to the row relative frequency I have what is called column relative frequencies. Now what are the type of questions we are expected to answer here. Let us go back to our contingency table here. So, now I want to know what is the proportion of total participants who were female? I have 100 participants the proportion is 44 by 100 so I have 44% female and 56% male participants.

Now of the people who own smart phones. So, I have 76 people who own smart phone I want to know among this what is the proportion of females among the smart phone owners. The answer to this question is total number of smart phone owners are 76 of which 34 of them are female. So, the proportion of people among the smart phone owners who are female are 34 by 76, proportion of male people who are owners are 42 by 76, proportion of female non owners are 10 by 24, proportion of male non owners are 14 by 24 these values 10 by 24, 14 by 24, 34 by 76 and 42 by 76 are what we refer to as a column relative frequency.

How do we obtain the column relative frequency? We divide each frequency by their respective column totals we get the column relative frequency.

(Refer Slide Time: 10:04)



Example 1: Column relative frequency

Gender	Own a smartphone		Row total
	No	Yes	
Female	10/24	34/76	44/100
Male	14/24	42/76	56/100
Column total	24	76	100

Gender	Own a smartphone		Row Total
	No	Yes	
Female	41.67%	44.74%	44.00%
Male	58.33%	55.26%	56.00%
Column Total	24	76	100



So, you can see the relative frequency is 41% of non owners are female, whereas 58% of non owners are male whereas when it comes to owning a cellphone 44% are female and about 55% are male. Totally 44% female and 56% are male. So, these are the column relative frequencies. Now let us look at the row and column relative frequencies for the second example which was the income versus the ownership of a smart phone.


(Refer Slide Time: 10:51)

Statistics for Data Science -I
└ Association between categorical variables
└ Relative frequencies

Example 2: Column relative frequency

	Own a smartphone		
Income level	No	Yes	Row total
High	2/38	18/62	20/100
Medium	27/38	39/62	66/100
Low	9/38	5/62	14/100
Column total	38	62	100

	Own a smartphone		
Income level	No	Yes	Row Total
High	5.26%	29.03%	20.00%
Medium	71.05%	62.90%	66.00%
Low	23.68%	8.06%	14.00%
Column Total	38	62	100



Now when you look at the column relative frequency for the second example you see that 20% high income, 66% medium income and 14% low income group. Now among the owners you can see that among the people who own a phone, 29% of the people who own a phone are from the high income, 63 are from medium and a low 8% are from the low income groups. Now when it comes to the people who do not own a phone you see only 5% are from the high income group, 71% are from the medium and 24% from the low income group. So, this is how you compute the relative frequency.

(Refer Slide Time: 11:53)



Section summary

- Concept of relative frequency: row relative frequency and column relative frequency.



We have introduced a concept of a relative frequency. In particular we have seen how to compute the row relative frequency and the column relative frequency. We now we will see how to use the concept of a row relative frequency and a column relative frequency to answer questions about association between variables.

(Refer Slide Time: 12:22)



Association between two variables

1. Contingency table - Summarizes
2.

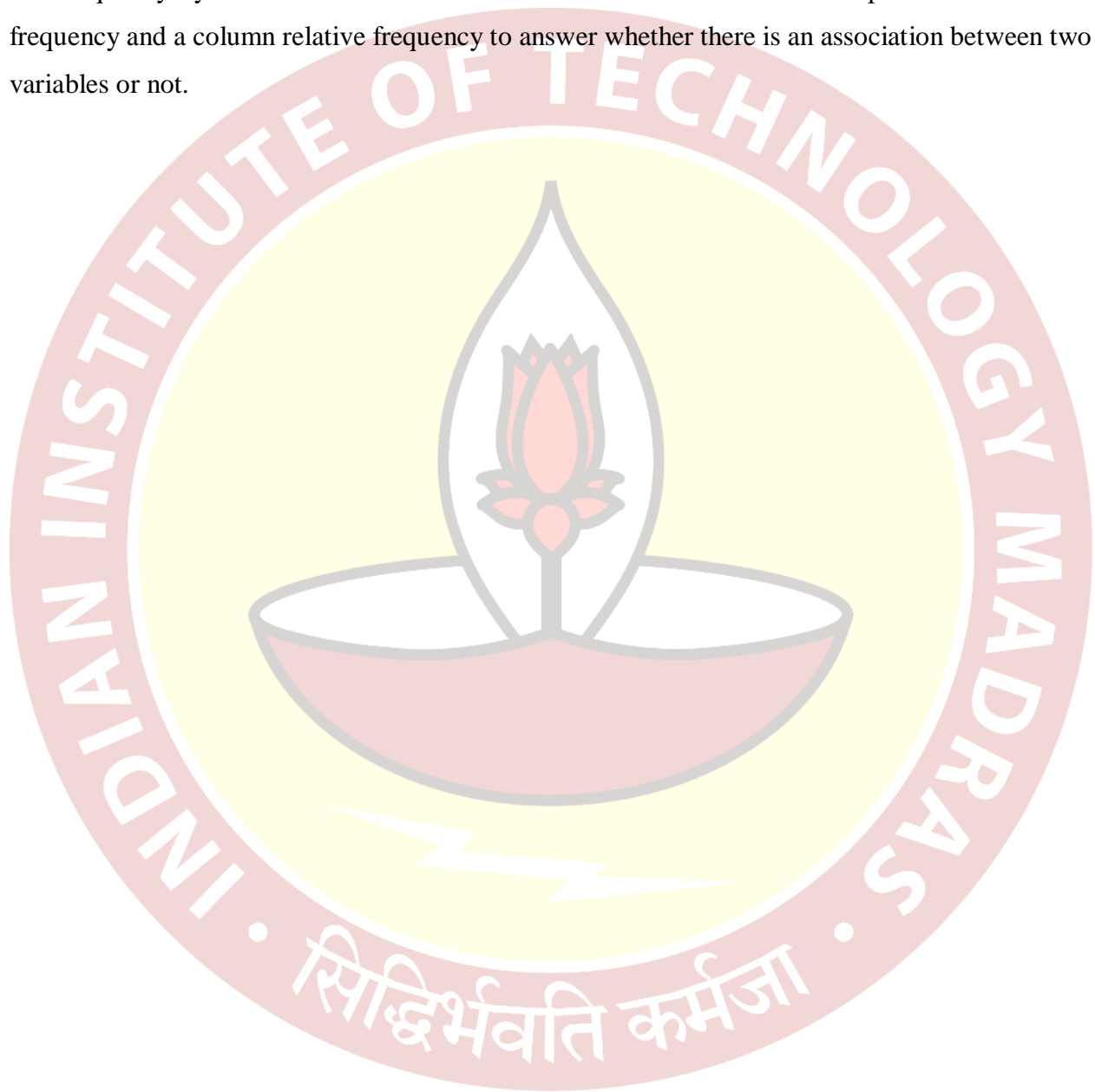
Row relative freq -	Cell freq / Row total
Col. relative freq -	cell freq / Col. total



So, now we address the questions which we started this module with or this lecture with by wanting to answer whether there is an association between two variables. We first introduce how to set up what we call a contingency table. A contingency table basically summarizes your bivariate data

then we introduced a notion of both a row relative frequency, to compute the row relative frequency you measured each or you divided each cell frequency by its row total.


And then we also introduced a notion of a column relative frequency where again I divided each cell frequency by its column total. Now we will see how to use this concept of a row relative frequency and a column relative frequency to answer whether there is an association between two variables or not.




(Refer Slide Time: 13:35)

Statistics for Data Science - I
└ Association between categorical variables
└ Association between variables

Association between two variables



- ▶ What do we mean by stating two variables are associated?
Knowing information about one variable provides information about the other variable.
- ▶ To determine if two categorical variables are associated, we use the notion of relative row frequencies and relative column frequencies described earlier.



So, what do we mean by saying or stating the two variables are associated. In other words, we want to know that whether information about one variable provides some information about another variable. So, when we are seeking to answer the question whether two variables are associated actually what we are seeking to answer is whether if I have information about a particular variable whether it actually gives me something or tells me something about the other variable. So, to determine whether two categorical variables are associated we will now show how we use the notion of relative row frequencies and relative column frequencies.

(Refer Slide Time: 14:32)



Association between two variables

- ▶ If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.
- ▶ If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.



So, let us look at our relative row or column frequencies. We already know how to compute these frequencies. If the row relative frequency or the column relative frequencies are the same for all rows. I repeat, if the row relative frequency or the column relative frequencies are the same for all rows or columns, then we say the two variables are not associated with each other.

If the row relative frequency or the column relative frequencies are different for some rows then we say that the two variables are associated with each other. So, if the row relative or the column relative are same we say they are not associated. If the row relative frequency or the column related frequency are different, then we say they are associated with each other.

(Refer Slide Time: 15:45)



Example 1: Association between two variables

- If the row relative frequencies (the column relative frequencies) are the same for all rows (columns) then we say that the two variables are not associated with each other.

Gender	Own a smartphone		Row total
	No	Yes	
Female	22.73%	77.27%	44
Male	25.00%	75.00%	56
Column total	24.00%	76.00%	100

Gender	Own a smartphone		Row Total
	No	Yes	
Female	41.67%	44.74%	44.00%
Male	58.33%	55.26%	56.00%
Column Total	24	76	100

Gender and smartphone ownership are not associated



So, let us go back to our examples to and apply this rule which compares the row relative frequency and the column relative frequency. So, let us look at this example in this example where I plotted or I tabulated the gender versus ownership of a phone you can see that when it comes to a ownership of a phone 24% of a total population did not own a phone, 76% owned a phone.

Now if I am looking at the pattern within the gender you see again 23% of the females did not own a phone and about 77% of the females owned a phone. When it comes to male again I see about 25% did not own a phone and 75% owned a phone. So, you see that the ownership pattern which is 24% not owning and 76% owning a phone is consistent with both the female subgroup and the male subgroup.

You do not see any inconsistencies that is both when you look at females also you see that about 23% are not owning and 77% are owning. In the male also we see about 25% are not owning and 75% are owning. So in general, the ownership pattern does not change depending on the gender. Let us look at the column frequencies. Again if you look at the column frequencies I had 44 % female and 56% male.

Now if you look at only owners of the 76 people again I see about about 45% are female and 55% are males which is almost the same as my total gender diversity. This same percentage among people who do not own a phone again I have about 42% females and 58% males. So, the gender

diversity also among owners of a phone and not owners of a phone is also the same that is 44% and around 44% and 56%.

So, both the row relative frequencies and the column relative frequencies are the same for all the rows and columns. Hence, I can say that both my gender and smart phone are not associated with each other which is consistent with the definition earlier. Now let us look at the second example.

(Refer Slide Time: 19:02)

Statistics for Data Science - I
 ↳ Association between categorical variables
 ↳ Association between variables


Example 2: Association between two variables

► If the row relative frequencies (the column relative frequencies) are different for some rows (some columns) then we say that the two variables are associated with each other.

Income level	Own a smartphone		Row Total
	No	Yes	
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

Income level	Own a smartphone		Row Total
	No	Yes	
High	5.26%	29.03%	20.00%
Medium	71.05%	62.90%	66.00%
Low	23.68%	8.06%	14.00%
Column Total	38	62	100

Income and smartphone ownership are associated



When I look at the second example I plot both the row relative frequencies, what are my row relative frequencies here. So, you have here what were the two variables; the income level and whether you own a phone or not. Again let us look at it in the first case I know 38% do not own a phone and 62% own a phone, but when you look at the high income group you see that 90% own a phone and 10% do not own a phone whereas, in the low income group, 65% do not own a phone and 35% own a phone.

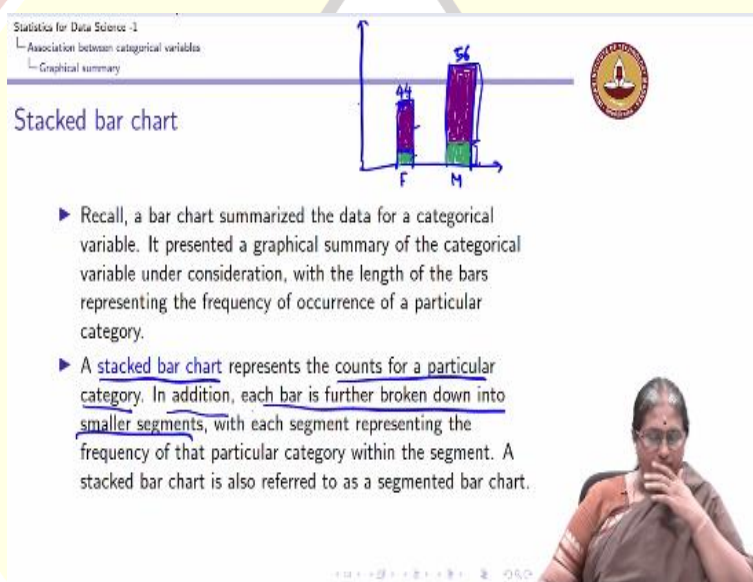
So, you can see that whether you own a phone or not is playing the percentages of ownership of a phone actually is different for the high income group and the low income group. The row relative frequencies are not the same among the categories. If you look at the column relative frequencies again I have a distribution of 20%, 66% and 14% in my high, medium and low income groups.

If I just look at the owners I have 62% who are owners of which I have about 30% who are coming from the high income group and only a 8% who come from the low income group. Among the non

owners I have a 5% who are from the high income group and a whopping 24% who are from the low income group. This is not consistent with my total distribution of my income categories.

Since the row frequencies and the columns frequencies are different among the rows and the column, I can say that the income and smart phone ownership are actually associated with each other which is very intuitive. You would expect the ownership of a phone to actually be associated with your income level whereas the ownership of a phone need not be associated with whether you are a female or a male and we have seen both the examples which actually have demonstrated this phenomena.

(Refer Slide Time: 21:50)



So, we have seen how we can use the concept of a relative frequency to decide whether two variables are associated or not. We again recall we said that if the row relative frequencies or the column relative frequencies are same for all the rows and columns we say two variables are not associated with each other. If they are same if they are not same or if they are different we say that they are associated with each other and we demonstrated this through the two examples which we have been discussing.

Now, how do I graphically show this result? So, again let us go back to our examples here see that we have a contingency table which is given here. Recall, when we wanted to summarize a single categorical variable we used what was called a bar chart. Now, I want to see how these two

variables behave with each other. So, for that what I do is I construct what is called a stacked bar chart or sometimes it is also referred to as a segmented bar chart. So, a bar chart summarizes the data for a categorical variable where the length of the bars were representing the frequency of occurrence of a particular category.

This is what a bar chart does. Now a stacked bar chart represents the counts for a particular category in addition each bar is further broken down to smaller segments. Now let us illustrate what we mean by this? Now if we are looking at a bar chart for the first example I had two categories the female category and the male category. Among the female I had 44 females and 56 males and you see that this is my bar chart.

Now what is a stacked bar chart? Now if I want the second category to be super imposed on this. What do I mean by this? Again you go back to your contingency table. You see that out of 44 in the first example I have out of 44 people I have 77.27 who own a phone that was this was 30 actually among the 44 people I have 77. So, this was 34 and this was 10 this was how I had it and this was a 14, 46.

This was not 46 this was a 14 and this was a 42. So, if I now given this bar chart I know that 77% so I have constructed this bar chart. In this I want to know what percentage of the female own a phone so what I can do in that case is I can look at a percentage let me look at another color. So, I have about 77% of this who own a phone approximately this is about 77%.

So, of this 44 I have about 77% similarly here I have the same percentage who own the phone here so this purple shaded area represents the number of people or the proportion of people within each category who own a phone. I have about 77% who own a phone and this green shaded area represents the proportion of people among females who do not own a phone and this green area here represents the proportion of males who do not own a phone.

So, the difference between a stacked bar chart and a bar chart is, when I have only the female and male category I could have. So, this entire bar represented the count of female and this entire bar represented which was 56 represented the count of female whereas a stacked bar chart in addition to the so what you can see that in addition to the count of a particular category it breaks it down into smaller segments.

So, I broken down this entire of 44 into smaller segment. Here again two segments where this segment represents the owners of or the female owners and the green segments represents the female non owners. Similarly, this segment represents the male owners and this segment represents the male non owners of the graph. So, since you have the segmented bars it is called a segmented bar chart or it is also known as a stacked bar chart.

(Refer Slide Time: 27:53)

Statistics for Data Science -1
Association between categorical variables
Graphical summary

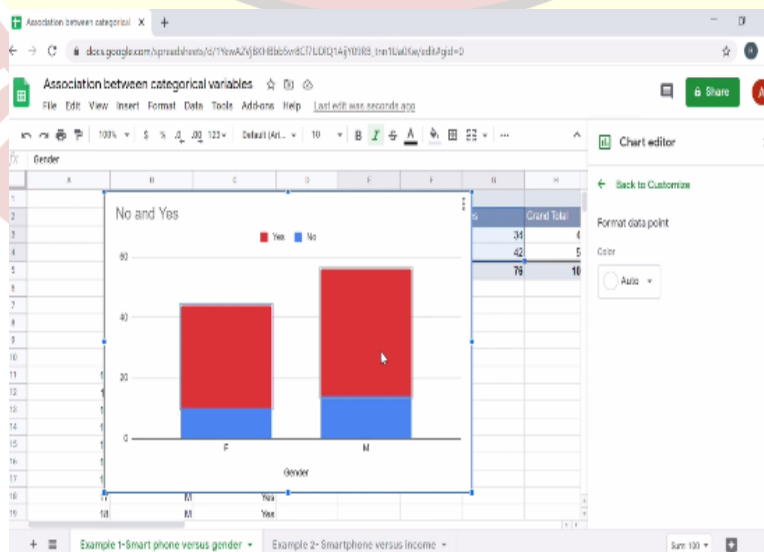
Stacked bar chart using google sheets

Step 1: Select the data you want to include in the contingency table.
Step 2: Click Insert - chart- choose stacked bar option

The slide features a watermark of the Indian Institute of Technology Madras and a small inset image of a woman in a brown sari.

How do we construct a stacked bar chart using Google sheet.

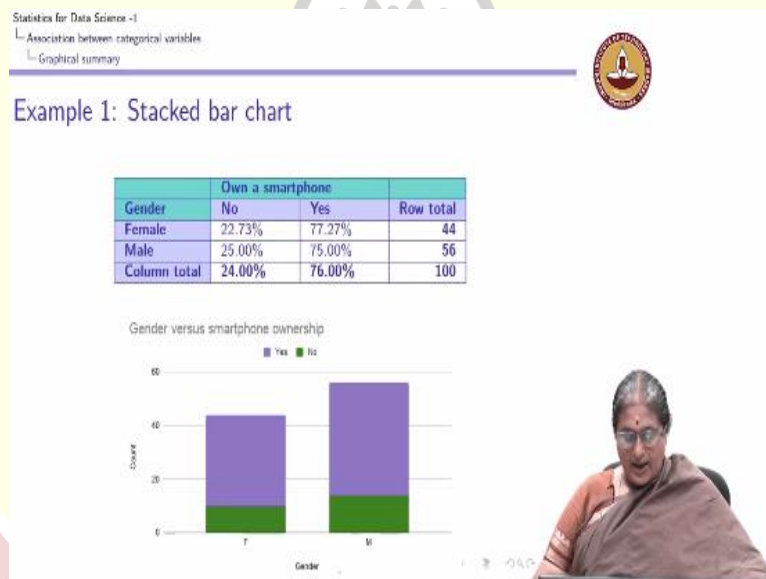
(Refer Slide Time: 27:55)



So, now we go back to the contingency table which we had already constructed. I select the data for which I have the contingency table I am selecting the gender and the no, yes I go to insert chart in an insert chart I am looking at a stacked column chart. You can see that I am looking at a stacked column chart. So, you can see that this is the stacked column chart of which you can also see that I have two genders. Here I have a female I have a male.

This is the gender I have and within the female I have 44 within the male I have 76. 77% of the female gender is actually owning a phone the red color indicates yes the blue color indicates no. Again close to 77% of male own a phone again yes. So, within the male bar I have indicated how many own a phone and how many do not own a phone.

(Refer Slide Time: 29:21)

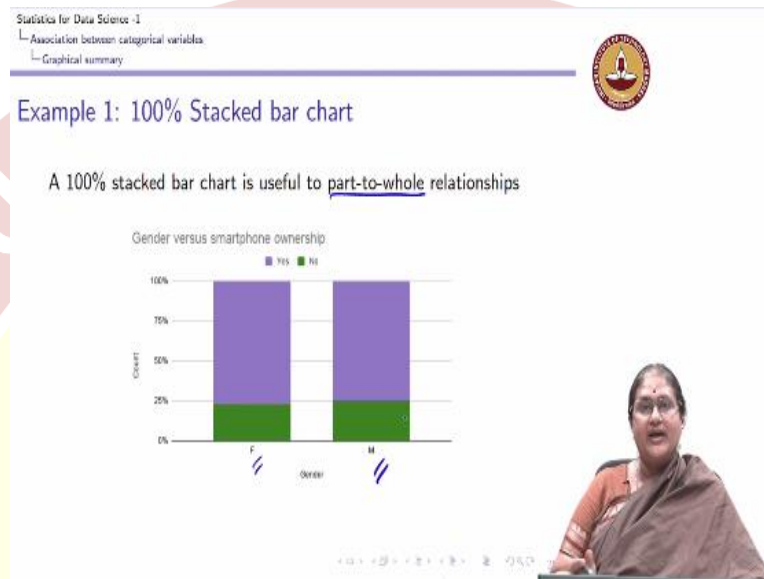


So, you can see that a stacked bar chart is a good way of summarizing the ownership in a graphical way. Now you can go back to your Google sheet and you can see that this stacked bar chart give me the actual counts, but suppose I am interested and this was what we referred to as a standard stacked bar chart. So, you can go to a chart style go to setup I have a stacked column chart.

Under stacking you can see that I have listed the standard option, but if I click on what I call a 100% stacked bar chart. What a 100% stacked bar chart gives me is you can see that of if I consider it does not give me the actual counts of a female and male, but what a 100% stacked bar chart gives me is the proportion of females who own a phone to the proportion who do not own.

And similarly the proportion of males who own a phone to proportion who do not own. Now, where is a 100% stacked bar chart useful to me. Now suppose I go to a 100% stacked bar chart. So, where is a 100% stacked bar chart useful.

(Refer Slide Time: 31:07)



You can see that when I am actually not interested in knowing the count of each category, but I am interested in knowing about a part-to-whole relationship or the proportional relationships I can use what is a 100% stacked bar chart. Here you can see it very visualize visually it is showing me that the distribution or ownership of a phone to not having a phone for the category female and male is almost the same you do not see any sizable difference between the ownership pattern and gender.

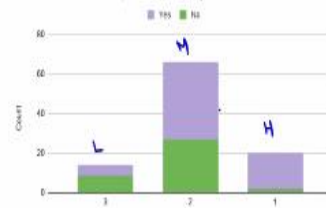
(Refer Slide Time: 31:50)



Example 2: Stacked bar chart

Income level	Own a smartphone		Row Total
	No	Yes	
High	10.00%	90.00%	20
Medium	40.91%	59.09%	66
Low	64.29%	35.71%	14
Column Total	38.00%	62.00%	100

Income versus smartphone ownership



For the second example when you look at a stacked bar chart. Now if you look at this chart I have the category 1. 1 you recall was associated with the high income group, 2 was associated with the medium income group, 3 was associated with the low income group again their order I either have to maintain a low medium high order or a high medium low order. This you have to maintain the order you cannot have a low, high, medium order because it is an ordinal data maintain the order in which you represent the variables.

Now if you look at this high income group yes is the purple, green is the no. You can see that among the high income group you have more people who own the phone to people who do not own the phone. In the low income group you have more people who do not own the phone to people who own the phone and in the medium income group it is almost you have equal number of people who own a phone to do not own a phone. Graphically this is very clear.

सिद्धिर्भवति कर्मजा

(Refer Slide Time: 33:02)



Now, how does a 100% stacked bar chart look for this example. Now if you look at a 100% stacked bar chart where I am not interested in the actual counts, but I am interested in looking at how the 100% stacked bar chart looks. You can see this very clearly in the high income group I have a lot of people who own the phone. In the low income group I have this green is higher than the purple.

Green is the number of people who do not own the phone whereas for medium I have equal number of the proportion of people who own the phone is equal to the proportion of people who do not own a phone. So, you can see when you are not interested in the actual counts, but you are interested in comparing these groups with each other to tell you a story a 100% stacked bar chart is very useful.


And this story is since you do not see a varying pattern we can reaffirm what we saw from the column and row relative frequency that income and ownership are associated with each other. Whereas, when we looked at gender versus ownership they were almost the same gender and ownership are not associated with each other.

(Refer Slide Time: 34:36)

Statistics for Data Science - I
└ Association between categorical variables
└ Graphical summary

Section summary

- Understand whether two categorical variables are associated using the concept of relative frequencies.
- Graphical summary of association using stacked bar chart.



So, at the end of this section you should know, you should be able to use the concept of relative frequency to tell whether two variables are associated with each other. You further validated through a graphical summary. This graphical summary is what we referred to as the stacked bar chart.