


IIT Madras
ONLINE DEGREE

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 4.2

Association between Two Categorical Variables - Introduction


(Refer Slide Time: 00:14)

Statistics for Data Science -1



Review


1. What is statistics?
 - ▶ Descriptive statistics, inferential statistics.
2. Understand how data are collected.
 - ▶ Identify variables and cases (observations) in a data set
3. Types of data-
 - ▶ Classify data as categorical or numerical data.
 - ▶ Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
 - ▶ Creating frequency tables, understanding relative frequency
 - ▶ Creating pie charts and bar charts
 - ▶ Descriptive measures of Mode and Median
5. Describing numerical data
 - ▶ Creating frequency tables: single valued and grouped data.
 - ▶ Measures of central tendency: Mean, Median, and Mode
 - ▶ Measures of dispersion: Range, Variance, Standard deviation
 - ▶ Percentiles, Quartiles, Interquartile range.



Now what is the next thing?


(Refer Slide Time: 0:17)

Statistics for Data Science -1



Learning objectives

1. Use of two-way contingency tables to understand association between two categorical variables.
2. Understand association between numerical variables through scatter plots; compute and interpret correlation.
3. Understand relationship between a categorical and numerical variable.



So far we have focused on understanding only about summarizing a single variable. But most of the time we are interested in understanding whether two variables are associated with each other. When I talk about association I am not mentioning about causality. Association is not always causality. We are not talking about causality here, but you are just asking questions about association between variables.

So, in this module the learning objectives are first we start by understanding association between two categorical variables. Here, we will introduce the notion of contingency tables and how we use contingency tables to understand the association between two categorical variables. Then we move forward to understand how two numerical variables are associated with each other. Here we talk about scatter plots. The nature of this plot and how we measure the association between two numerical variables.

Though the focus of this module is mainly to understand association between two categorical and association between numerical variables, we also spend some time to understand how you will talk about a relationship or an association between a categorical and a numerical variable. So, these are the learning objectives of this week.

(Refer Slide Time: 02:01)

Statistics for Data Science - I

- Association between categorical variables
- Introduction

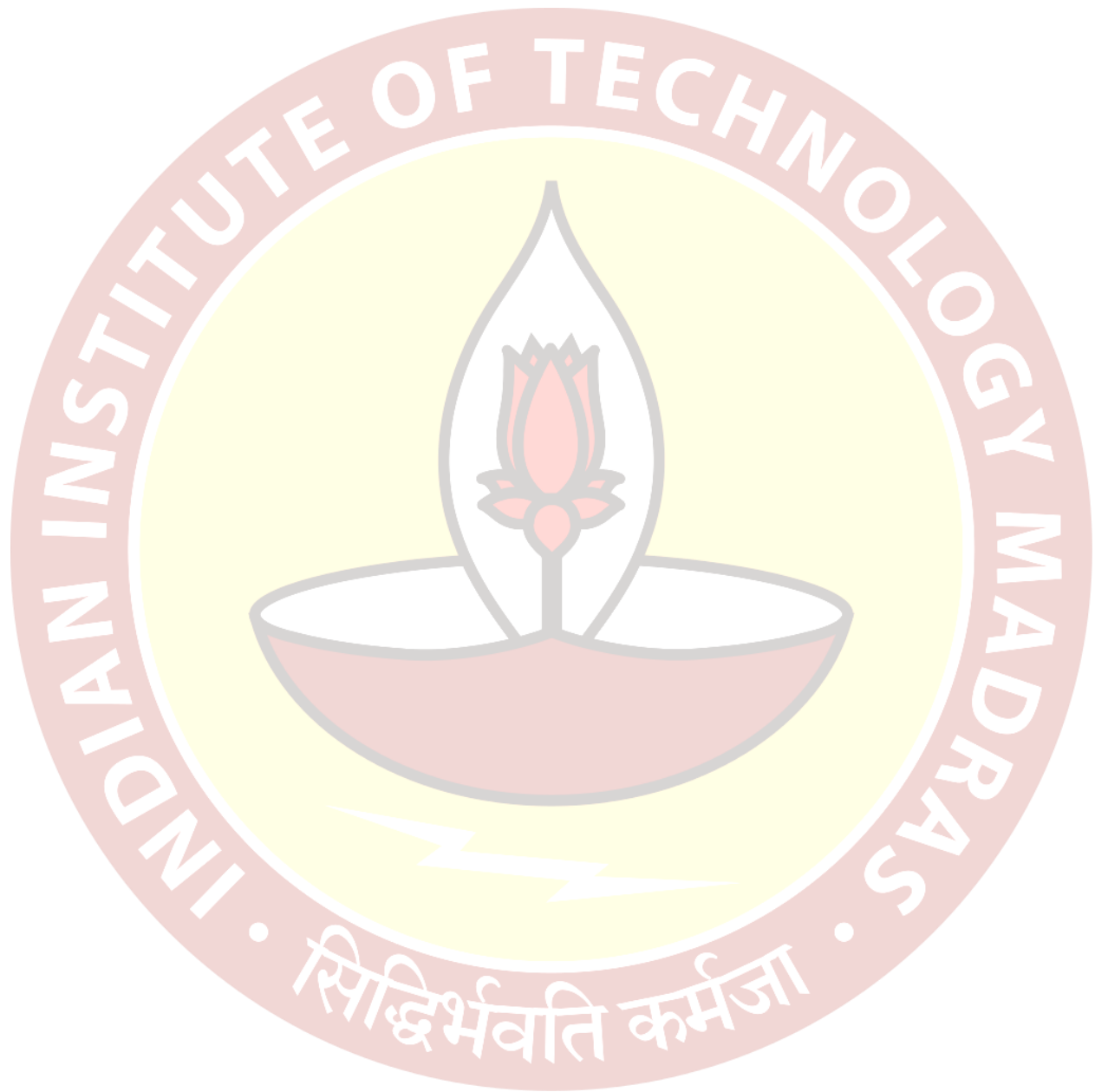
Introduction

- ▶ To understand the association between two categorical variables.
- ▶ Learn how to construct two-way contingency table.
- ▶ Learn concept of relative row/column frequencies and how to use them to determine whether there is an association between the categorical variables.

6/26

So, we start with the association between categorical variables. So, what is the main objective of this section? So, here we are going to understand about how to construct what we refer to as a two

way contingency table. We will introduce the concept of relative frequencies and use how you can use this concept of relative frequency to determine whether there is an association between two categorical variables or not.



(Refer Slide Time: 02:37)

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



Example 1: Gender versus use of smartphone

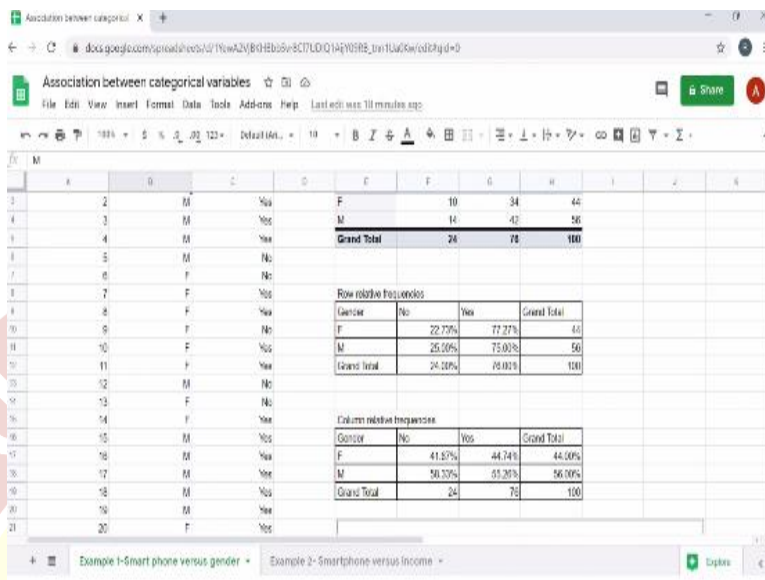
- ▶ A market research firm is interested in finding out whether ownership of a smartphone is associated with gender of a student. In other words, they want to find out whether more females own a smartphone while compared to males, or whether owning a smartphone is independent of gender.
- ▶ To answer this question, a group of 100 college going children were surveyed about whether they owned a smart phone or not.
- ▶ The categorical variables in this example are
 - ▶ Gender: Male, Female (2 categories)- Nominal variable
 - ▶ Own a smartphone: Yes, No (2 categories)- Nominal variable



So, let us start with an example. Now if you look at this example I have a market research firm which is interested in finding out whether ownership of a smart phone is associated with gender of a student. In other words, the company is interested in knowing whether more females own a smart phone while compared to males or whether owning a smart phone is independent of the gender of a person. This is the main question.

So, how do we go about answering this question? Immediately you see that when I talk about this I have actually discussed two variables. The first variable is gender and the second variable is ownership of a smart phone. So, how have we captured this gender is again a categorical variable. It has two categories. I am assuming male and female. Ownership the way I have captured the ownership of a smart phone is again through a categorical variable. Here, it is a binary if you own a phone you say yes, if you do not own a phone you say no.

(Refer Slide Time: 04:03)



The screenshot shows a Google Sheet with a dataset of 100 students. The first two columns are 'Gender' (M for Male, F for Female) and 'Smartphone' (Yes/No). The last column shows the 'Grand Total' for each row, which is 100. The sheet also includes a pivot table summarizing the data by gender and smartphone ownership.

Gender	Smartphone	Grand Total
F	No	44
F	Yes	56
M	No	24
M	Yes	76
Grand Total		100

Row relative frequencies:

Gender	No	Yes	Grand Total
F	22.73%	77.27%	44
M	25.00%	75.00%	56
Grand Total	24.00%	76.00%	100

Column relative frequencies:

Gender	No	Yes	Grand Total
F	41.87%	44.74%	44.00%
M	58.13%	55.26%	56.00%
Grand Total	24	76	100

So, let us look at the data. So, if you look at the data you can see that this is a data. What is a data? I have the data which has been collected where the variables which I am talking about are basically you can see gender and whether they own a smart phone or not, that is the data that is collected and this data is actually collected for 100 university students. So, what is the data?

The data that is collected is a group of 100 college going children were surveyed about whether they own a smart phone or not. So, the data collected is for each student what was captured, gender and whether they own a phone or not. So yes, no, and gender was captured. So, for example, person 1, I ask a gender it could be a male. This person if they owned a cellphone it would have been yes. Person 2, male does not own a phone.

Person 3 could have been a female does not own a phone. Person 4 could have been a female owns a phone, person n, and this is the way for 100 students we collect this data. So, the two variables here are gender. The second variable here is whether they have a smart phone or not. Yes, if they have a phone no, if they do not have a phone. So, this is the data we have collected. The total number of observations are 100 and this is how we have collected the data.

So, what are the categorical variables in this example? The first categorical variable is gender. They are two categories and we know that gender is a nominal variable. Then next categorical variable is whether you own a smart phone or no. The values this variable take are yes and no. The

values gender take are male and female. Again I have two categories, again it is a nominal variable because there is no order in this variable.

Hence, you can see that I need to know that what is the kind of my variable and what is the scale of measurement. Here I have both the categorical variables with nominal scales of measurement.

(Refer Slide Time: 6:53)

Statistics for Data Science - I
Association between categorical variables
Contingency tables

Example 1: Gender versus use of smartphone-summarize data

- ▶ We have the following summary statistics
 1. There are 44 female and 56 male students
 2. 76 students owned a smartphone, 24 did not own.
 3. 34 female students owned a smartphone, 42 male students owned a smartphone.
- ▶ The data given in the example can be organized using a two-way table, referred to as a contingency table.

	Own	Not Own	Row Total
Female	34	10	44 ✓
Male	42	14	56 ✓
Col. Total	76 ✓	24 ✓	100

So, now let us look at the data once you have the data, this is the data which I am talking about. So, here you have 100 students and from each of these observations I copy their gender and whether they own a smart phone or not. So, what is the summary statistics I have from my data? The summary statistics I have from my data are I have 44 female students and 56 male students. Remember we are asking 100 students the question a for every student we record what is their gender and we record whether they own a smart phone or not. This is our survey.

Further, 76 students owned a smart phone and 24 did not own. So, if you look at it what we discussed in the previous weeks was, first if I look at gender as a categorical variable how to summarize this gender. I know 44 female and 56 male we saw that we could summarize it using a bar chart. Again owning a smart phone or not. Ownership yes no, I have 76 yes and 24 no. Again there are two values this ownership takes, yes and no and both of them here $44 + 56$ add up to a 100. $76 + 24$ add up to a 100.

Now I want to know the association between these two variables. So, I have another data which is useful to me which says that 34 female students owned a smart phone and 42 owned a smart phone. So, this is the data which is given to us. So, the first question we ask is how do I summarize this data. So, given this data which is in the form of this table which I have here, the question is how do I summarize this data? I have two variables, the first variable I can write it as a gender.

Now there are two values this gender takes I write it as a female, I write it as a male. So this is my variable. The other variable is ownership. Now this is a yes or this is a no this is a yes and I have grand totals. So, if I have the grand row total, I have it here, I have a column total here. So, you can see that there were 44 female that is what is given here, 56 male that is what is given here.

Now when I come to ownership 24 did not own, so my no is 24, 76 owned. So, we can see that this total add up to 100 student. This is the first thing. I am not looking at filling in the inside table, but these are the 44 females, 56 males in my dataset, 24 not owning, 76 owning. Now further the so this is what I have tabulated. So, this is 34 because 34 female owned a smart phone. Similarly, 42 male owned a smart phone.

Now how many did not own that is easy. 10 which is $44 - 34$ and here I have 14 which is $56 - 42$. We can see $10 + 14 = 24$, $34 + 42 = 76$ and we can also check that $10 + 34 = 44$ and $14 + 42 = 56$. So, summarizing this data or the data given here is referred to as a two way table more popularly referred to as a contingency table. How do we construct a contingency table?

To construct a contingency table we look at the first variable. The levels of the first variable in this the first variable is gender and it has two values female and male. So, suppose this variable takes 3 values. So the level 1, level 2, level 3 or level m of the first variable goes into my rows. I look at my second variable. Suppose there are n values of the second variable. I have level 1, level 2, level n , I will have n columns.

And what goes into the i, j th column here is the number of observations of the i th variable and the j th variable together. For example 34 is a number of female students who own a phone. So, this is how we construct what is a contingency table.

(Refer Slide Time: 12:30)

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

Example 1: Gender versus use of smartphone-summarize data


► We have the following summary statistics

1. There are 44 female and 56 male students
2. 76 students owned a smartphone, 24 did not own.
3. 34 female students owned a smartphone, 42 male students owned a smartphone.

► The data given in the example can be organized using a two-way table, referred to as a contingency table.

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

Nominal



And you can see that this is how we have summarized the contingency table. Now in this example, both gender and ownership of smart phone both of them were what we referred to as a nominal variable. There was no order in this variable. So, if I had constructed my contingency table by looking at the following, that is male, female, no, yes it would not have made a difference because the information given is the same.

The order did not matter whether I had female and male or yes or no it would not have made a difference. So, the order in which you are stating your variables in the contingency table will not matter when both my variables are nominal in nature.

(Refer Slide Time: 13:39)



Contingency table using google sheets

- Step 1 Choose the columns of the variables for which you seek an association.
- Step 2 Go to Data-click on Pivot table option
- Step 3 Click on create option in the pivot table- it will open the pivot table editor:
 - 3.1 Under the Rows tab, click on the first categorical variable.
 - 3.2 Under the columns tab, click on the second categorical variable.
 - 3.3 Under the values tab, click on either of the variables and then click on the COUNTA tab under "summarize by" tab.



Now we will discuss how to create these contingency table using Google sheet. Go to the data, so this is the data which I have here. You can see that this data has 100 observations on both the variables. The variable gender is in my B column, the variable smart phone is in my C column. So, how do I create the contingency table? I choose my or I highlight my data, how many observations do we have?

We have 100 observations, I highlight my data so I would go and choose the columns. the variables I seek association now are gender and ownership. I choose it then I go to the data tab and click on what is called pivot table. That is step 2, go to data and click on the pivot table option. Now in the pivot table, create pivot table. It opens the pivot table editor in the existing sheet.

I am going to so I go to data I click on pivot table in the existing sheet. I am just going to give a location to my pivot table. The location I give it here and now I go to Create. Under the rows tab add the first categorical variable which is gender. Under the columns tab, click on the second categorical variable which is ownership of a smart phone that is 3.2 step. Under the values tab so I go back here, under the values tab I click on either of the variables.

So, I clicked on gender here and I am asking it to summarize by count A and you can see that I have what I got here is my precisely this is the categorical or the contingency table which we just did a few minutes before. So, you can see that there are 34 females who own a phone, 42 males

who own a phone, 10 females do not own a phone, 14 males do not own a phone, 42 females and 56 males in my dataset and 24 do not own a phone whereas 76 own a phone.

This is what I get from my dataset. So, this is the pivot table in your Google sheet which gives you a contingency table in Google sheet. So, if you saw in the earlier example we had two nominal variables. Now what would happen if I have a ordinal variable?



(Refer Slide Time: 17:02)



Example 2: Income versus use of smartphone

- ▶ A market research firm is interested in finding out whether ownership of a smartphone is associated with income of an individual. In other words, they want to find out whether income is associated with ownership of a smartphone.
- ▶ To answer this question, a group of 100 randomly picked individuals were surveyed about whether they owned a smart phone or not.
- ▶ The categorical variables in this example are
 - ▶ Income: Low, Medium, High (3 categories) - Ordinal variable
 - ▶ Own a smartphone: Yes/No (2 categories) - Nominal variable



Now, let us look at another example here. Now in the earlier thing I saw whether gender was associated with ownership of a phone, I summarize that using a contingency table. Now I am going to look at whether income actually is associated with ownership of a phone. So, again we have the same market research phone which is interested in finding out about whether ownership of a phone again this ownership of a smart phone is my variable here is associated with income of an individual.

Now the income variable is again how do we record this income variable. In this example, we have a market research firm which is interested in finding out whether ownership ownership of a smart phone is associated with income. So, what are the two variables here? The first variable is ownership. Again this was the variable we considered in our earlier example, but now instead of gender I am considering income. Now how is this income recorded?

How is this income recorded, the income is recorded as high, medium or low. So, we have categorized this income into 3 categories and what are the values of this income variable. It is a high, medium and low. So, this is a categorical variable where I am not actually calculating or I have not recorded the actual income, but I have actually categorized these 100 people again into whether they come from a high income group or a medium income group or a low income group.

And for each of these person we are asking whether their income is a high income and whether they own a smart phone or not. This is how I have recorded my data. So, here if you look at this case now what are my variable? Again the categorical variable are income which is low, medium and high and the second categorical variable whether you own a smart phone or not.

This variable whether you own it or not has two categories, the yes category and the no category. It is a nominal variable whereas the income which has three categories, the low, medium and high is an ordinal variable because there is an order in low, medium, and high because low income is lesser than medium income which is lesser than high income. So, recall when you are summarizing two nominal variables.

We said the order in which they appear in the table is not of any relevance. However, when you have an ordinal variable it is good to maintain the order. What do we mean by this?

(Refer Slide Time: 20:25)

	Income	No	Yes	Grand Total
High		2	18	20
Low		9	5	14
Medium		27	39	66
Grand Total		38	62	100

	Income (Coded)	No	Yes	Grand Total
1		2	18	20
2		27	39	66
3		9	5	14

Suppose I again continue with the same way. I choose income and own a smart phone. I choose these 100 observations, I go to my data, I click on pivot table in my existing sheet I am going to create my pivot table here. Again under rows I add income, under columns I add own a smart phone and under values I am just going to add on a income, you can see that this is my contingency table which I have here, but what you notice in this contingency table is first I have a high which is highlighted here.

So, let us go and look at only the contingency table. So, you can see in the contingency table, if you are looking at the order of the income, you have a high income, a low income and a medium income. Whereas the actual order is either high, medium, low or low medium, high. That is the order in which the variable appears. So, you do not want to see a jumbled order of a variable where there is an order of the variable. One way to overcome this is to have a order, have a high, medium and low variable. I have just coded this variable as 1, 2, 3 where 1 represents a high income, 2 represents a medium income, 3 represents a low income.

So, now if I am looking at a contingency table between these two variables again I choose the 100 observations which I need. I choose the 100 observations. I again go to data, pivot table in the existing sheet I am going to go and create a data a pivot table here. I will just click on this. I am going to create it. Rows I again add the income, columns I add whether they add have a smart phone, values I am going to do the count again.

So, now you can see and you can actually compare these two tables, the first table my high, low and medium did not have an order whereas in the second table, 1 represents a high income group, 2 represents a medium income group, and 3 represents a low income group. So, you can see that the order is preserved in the contingency table. So, whenever you have an ordinal variable it is recommended that the order is preserved in your contingency table.


(Refer Slide Time: 23:58)

Statistics for Data Science -I
 Association between categorical variables
 Contingency tables

Example 2: Contingency table

- ▶ We have the following summary statistics
 1. There are 20 High income, 66 medium income, and 14 low income participants.
 2. 62 participants owned a smartphone, 38 did not own.
 3. 18 High income participants owned a smartphone, 39 Medium income participants owned a smartphone, and 5 Low income participants owned a smartphone.
- ▶ The contingency table corresponding to the data is given below.

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



So, what finally how does my contingency table look for this. I have this data and using this data you can see that the corresponding table, corresponding to this data is I have a high, medium, low the order is preserved in the income whether they own a smart phone or not is recorded. I have 20 people who are high income, 66 who are medium, 14 who are low income. Out of these 162 own a phone, 38 do not own a phone.

Among the high income 20 people, 18 own a phone, 2 do not own a phone. Among the 14, low income group 9 do not have a phone, 5 have a phone. Among the 66, I have 27 who do not own a phone and 39 who own a phone.

(Refer Slide Time: 25:03)

Statistics for Data Science - I

- Association between categorical variables
- Contingency tables

Section summary

- Organize bivariate categorical data into a two-way table-contingency table.
- If data is ordinal, maintain order of the variable in the table

Var 1

	Var 2
	1
	2
11/14	

So, at the end of this subsection you should know how to organize bivariate categorical data into a two way table. Record the first variable and its level in one. These could be the rows. Record the second variable here, this is variable 1, variable 2, and a cell here, the i th level and the j th level tell how many of variable 1 and variable j th level of variable 2, variable 1 i th level and variable 2 j th level are there in this particular cell and this is what is referred to as a contingency table. A word of caution, if the data is ordinal maintain the order of the variable in the table.