IIT Madras

ONLINE DEGREE

Statistics for Data Science -1

Lecture 4.3: Association between two categorical variables-Relative frequencies

Usha Mohan

Indian Institute of Technology Madras

Row relative frequencies

Row relative frequencies

▶ What proportion of total participants own a smart phone?

Row relative frequencies

▶ What proportion of total participants own a smart phone?

▶ What proportion of female participants own a smart phone?

Row relative frequencies

▶ What proportion of total participants own a smart phone?

▶ What proportion of female participants own a smart phone?

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

Row relative frequencies

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

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

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

Example 1: Row relative frequency

Example 1: Row relative frequency

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

Example 1: Row relative frequency

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

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

Example 2: Row relative frequency

	Own a smartphone		
Income level	No	Yes	Row total
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

	Own a smartphone		
Income level	No	Yes	Row Total
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

Column relative frequencies

▶ What proportion of total participants are female?

▶ What proportion of smart phone owners are females?

Column relative frequencies

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

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

Column relative frequencies

▶ What proportion of total participants are female?

▶ What proportion of smart phone owners are females?

	Own a smartphone		
Gender	**No**	**Yes**	**Row total**
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.

Example 1: Column relative frequency

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

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

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**

Section summary

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

Association between two variables

Association between two variables

► What do we mean by stating two variables are associated?

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.

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.

Association between two variables

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.

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.

Example 1: Association between two variables

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.

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.

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

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

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

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

Example 2: Association between two 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.

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.

	Own a smartphone		
Income level	No	Yes	Row Total
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

	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

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.

	Own a smartphone		
Income level	No	Yes	Row Total
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

	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

Income and smartphone ownership are associated

Stacked bar chart

Stacked bar chart

▶ Recall, a bar chart summarized the data for a categorical variable. It presented a graphical summary of the categorical variable under consideration, with the length of the bars representing the frequency of occurrence of a particular category.

Stacked bar chart

- ▶ Recall, a bar chart summarized the data for a categorical variable. It presented a graphical summary of the categorical variable under consideration, with the length of the bars representing the frequency of occurrence of a particular category.

- ▶ A stacked bar chart represents the counts for a particular category. In addition, each bar is further broken down into smaller segments, with each segment representing the frequency of that particular category within the segment. A stacked bar chart is also referred to as a segmented bar chart.

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

Example 1: Stacked bar chart

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

Gender versus smartphone ownership
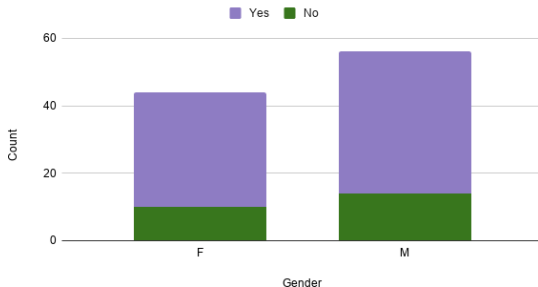
Example 1: 100% Stacked bar chart

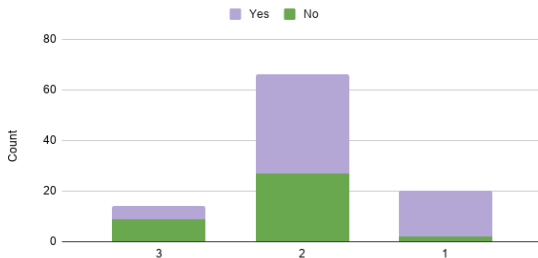A 100% stacked bar chart is useful to part-to-whole relationships

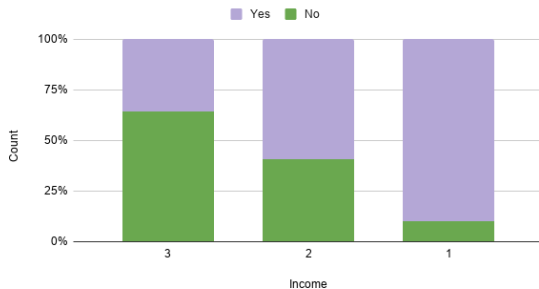Gender versus smartphone ownership
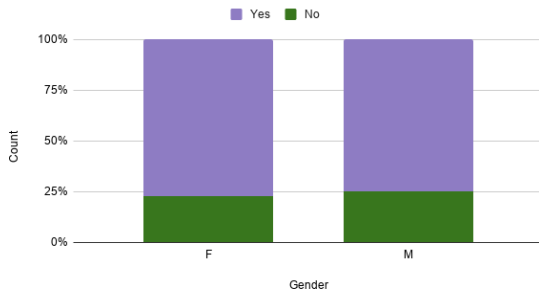
Example 2: Stacked bar chart

	Own a smartphone		
Income level	No	Yes	Row Total
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

Example 2: 100% Stacked bar chart

Section summary

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