Displaying and Describing Categorical Data

John Karuitha

4/22/2021

Introduction

- We have already defined categorical data.
- Categorical data falls into two categories: ordinal vs nominal.
- Again, we have seen that to summarise categorical data, we use frequency tables or contigency tables.
- We beriefly review frequency tables.

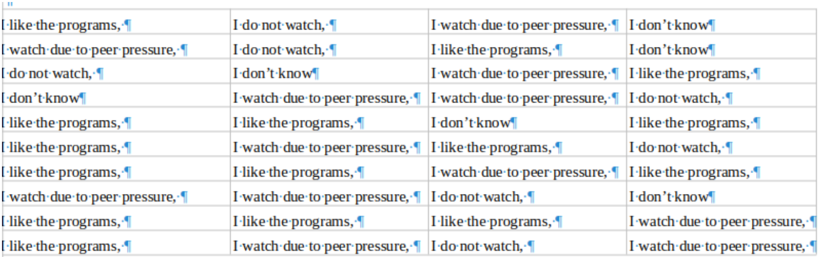
Frequency Tables

- A frequency table records the counts for each of the categories of the variable.
- These are the pure frequency tables
- Some tables report percentages. These are RELATIVE frequency tables.
- Many tables also report both counts and percentages.

Frequency Tables

- Example:
- We asked 40 people whether they watch soap operas on TV because they are interested in the program itself or due to peer pressure.
- There are four answers: I like the programs, I watch die to peer pressure, I do not watch, I don't know. ;
- The results of the poll are as follows; See the data in the file

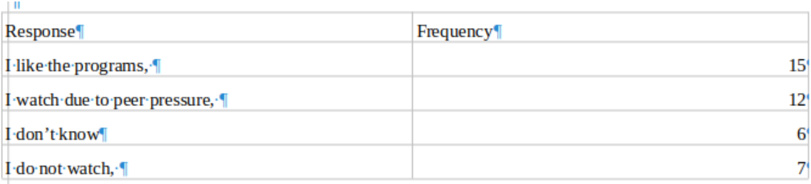
Frequency Tables



I like the programs,	I do not watch,	I watch due to peer pressure,	I don’t know
I watch due to peer pressure,	I do not watch,	I like the programs,	I don’t know
I do not watch,	I don’t know	I watch due to peer pressure,	I like the programs,
I don’t know	I watch due to peer pressure,	I watch due to peer pressure,	I do not watch,
I like the programs,	I like the programs,	I don’t know	I like the programs,
I like the programs,	I watch due to peer pressure,	I like the programs,	I do not watch,
I like the programs,	I like the programs,	I watch due to peer pressure,	I like the programs,
I watch due to peer pressure,	I watch due to peer pressure,	I do not watch,	I don’t know
I like the programs,	I like the programs,	I like the programs,	I watch due to peer pressure,
I like the programs,	I watch due to peer pressure,	I do not watch,	I watch due to peer pressure,

Figure 1: data for frequency tables

Frequency Tables

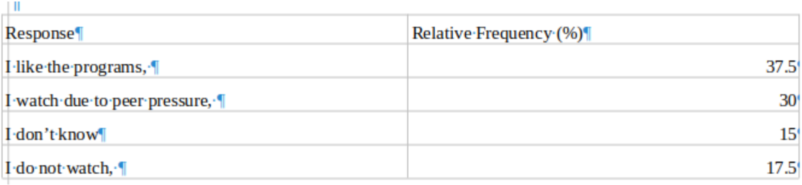


Response	Frequency
I like the programs,	15
I watch due to peer pressure,	12
I don't know	6
I do not watch,	7

Figure 2: freuency table itself

Frequency Tables: Relative frequency tables

- We convert the counts to percentages to get relative frequency tables



Response	Relative Frequency (%)
I like the programs,	37.5
I watch due to peer pressure,	30
I don’t know	15
I do not watch,	17.5

Figure 3: Relative frequency tables

Visualizing categorical data

- Two commonly used visualization tools for categorical data are
  - Pie Charts.
  - Bar graphs
- Pie charts are less favored given that they use area to represent data.
- The human mind finds it hard to interpret areas (angles).
- The bar chart is easier for the human mind because its a matter of comparing heights. It is linear.

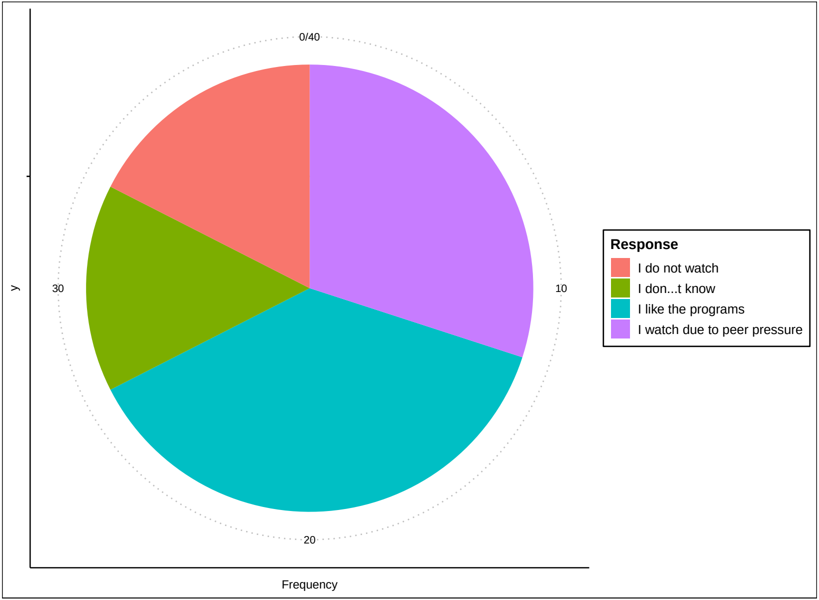
Visualizing categorical data: The area principle

- The best data displays observe a fundamental principle of graphing data called the area principle.
-
  - The area principle states that the area occupied by a part of the graph should correspond to the magnitude of the value it represents.
- That is why, in doing a bar graph, make sure the bars have the same widths. The comparison should only be on height.

Visualizing categorical data: The Pie Chart

Response	Frequency	Relative_frequency
I like the programs	15	37.5
I watch due to peer pressure	12	30.0
I don't know	6	15.0
I do not watch	7	17.5

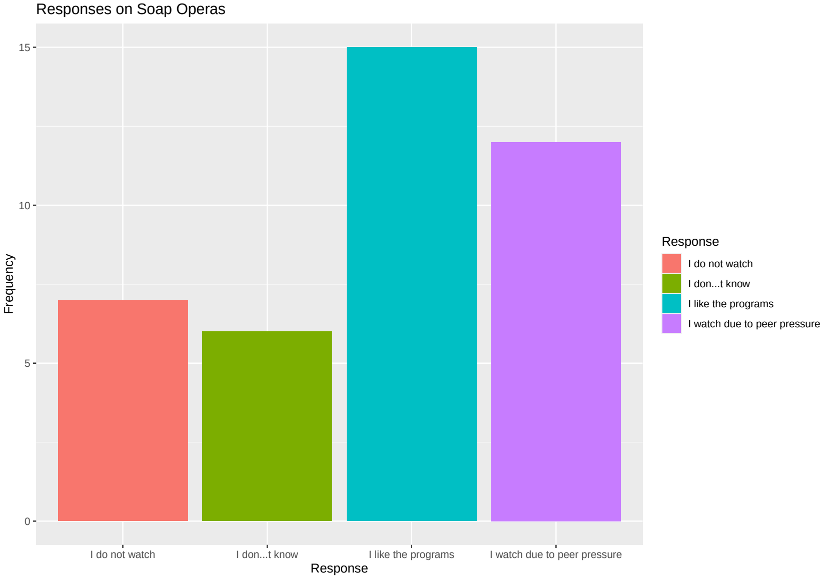
Visualizing categorical data: The Pie Chart



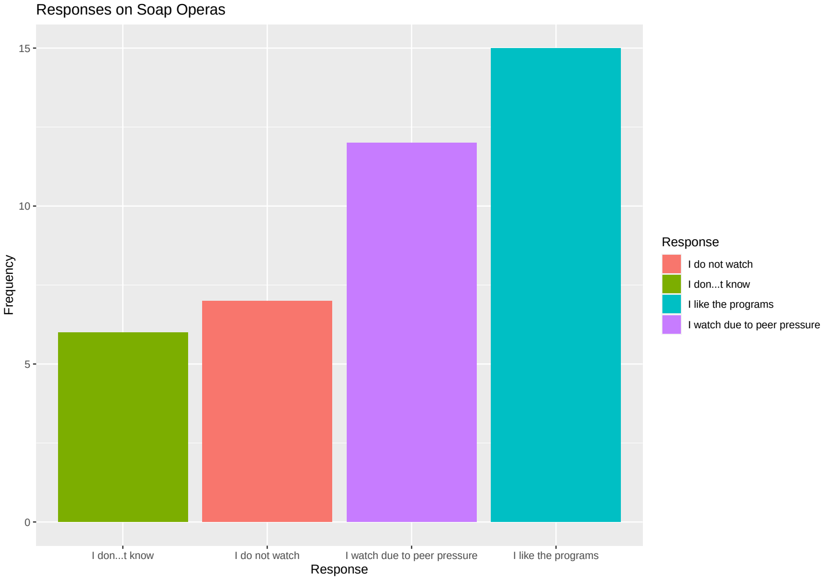
Visualizing categorical data: The bar graph

- The x - axis has the categories while the y-axis has the values.
- As noted, due to the area principle, let the width of the bars be the same so people can compare the heights only.
- Note that it is easier to interpret the bar chart as compared to the pie chart.
- When you have many categories, interpreting the pie chart gets even harder.
- For bar charts its better to arrange the bars in asceding or descending order of height. See examples.

Visualizing categorical data: The bar graph



Visualizing categorical data: The bar graph (Looks better with order)



Visualizing categorical data: The bar graph

- Visualizing and summarising data is perhaps one of the most important but under-estimated skill in statistics and data analysis.
- Before subjecting a dataset to a battery of statistical tests, do the following.
  - Draw a chart.
  - Draw a chart.
  - Draw a chart.
  - Summarise the data - mean, median, mode, SD, Variance, Quartiles, Extreme values, IQR.

Visualizing categorical data: Exercise

- The following dataset shows the responses of individuals in Kenya regarding whether they are generally happy or not.
- Draw a relative frequency table.
- Draw a pie chart from the relative frequncy table.
- Draw a bar chart with % on y -axix and responses on the x-axis, arranging the reponses in ascending order of relative frequency (%)

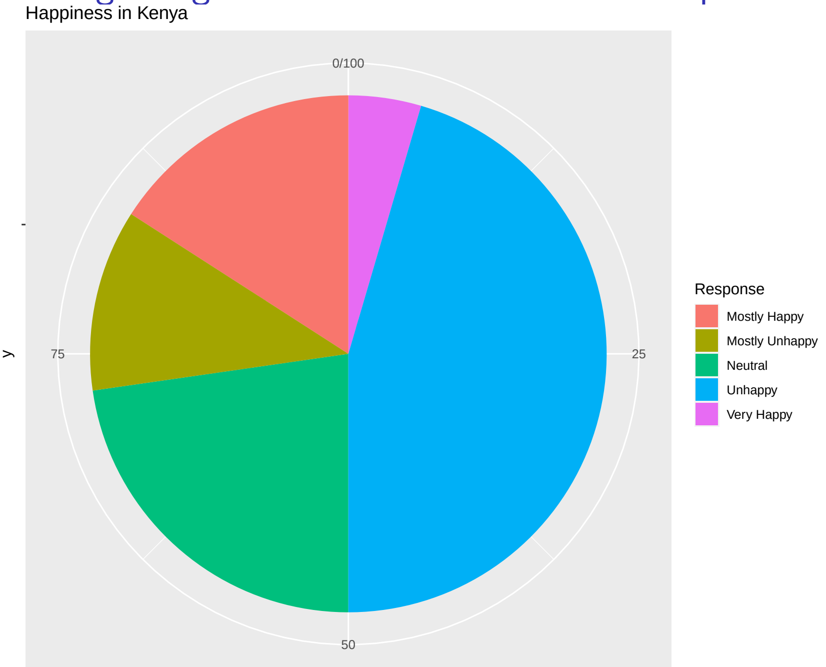
Visualizing categorical data: Exercise

Response	Frequency
Very Happy	200
Mostly Happy	700
Neutral	1000
Mostly Unhappy	500
Unhappy	2000

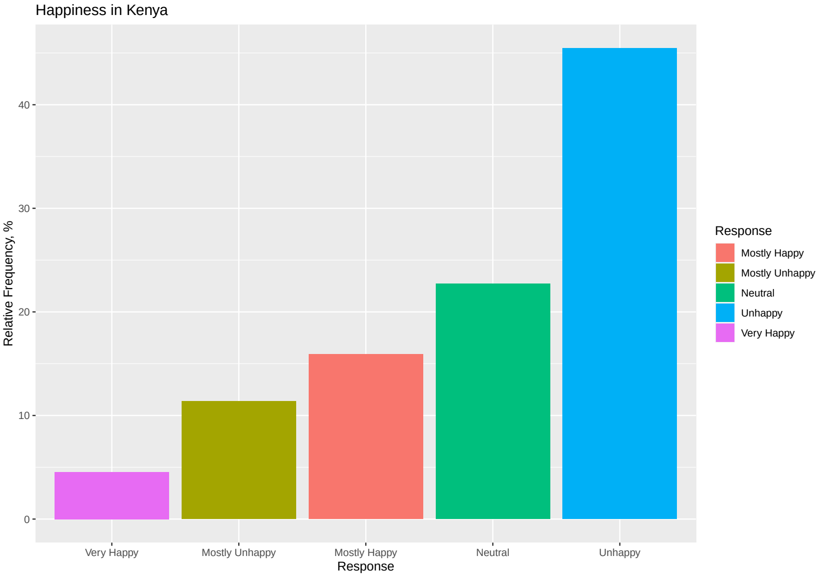
Visualizing categorical data: Exercise solutions- relative frequency table

Response	Frequency	relative_freq
Very Happy	200	4.545454
Mostly Happy	700	15.909091
Neutral	1000	22.727273
Mostly Unhappy	500	11.363636
Unhappy	2000	45.454546

Visualizing categorical data: Exercise solutions- pie chart



Visualizing categorical data: Exercise solutions- pie chart



Exploring Two Categorical Variables: Contingency Tables

- Sometimes you have two categorical data that you want to summarise together.
- In this case you may use a special type of frequency table called the contigency table.
- The contigency table may present counts or proportions.
- The one with proportions is a relative contigency table.

Exploring Two Categorical Variables: Contingency Tables

- In the previous example, we add a variable for the sex of the respondents, Female or Male.
- See the data in excel.

Exploring Two Categorical Variables: Contingency Tables

- Here are the first 10 rows of the dataset.

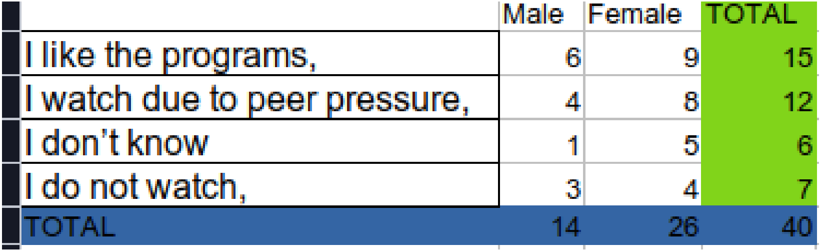
Comment	Sex
I like the programs,	Male
I watch due to peer pressure,	Male
I do not watch,	Male
I don't know	Male
I like the programs,	Male
I like the programs,	Male
I like the programs,	Male
I watch due to peer pressure,	Male
I like the programs,	Male
I like the programs,	Male

Exploring Two Categorical Variables: Contingency Tables

- A contigency table will break down the data by both variables, comment and sex.
- For instance, how many men said they do not watch the programs.
- How many women watch the programs out of peer pressure. and so on.
- Again, the summaries can be in the form of counts or percentages.

Exploring Two Categorical Variables: Contingency Tables

- Here we go



	Male	Female	TOTAL
I like the programs,	6	9	15
I watch due to peer pressure,	4	8	12
I don't know	1	5	6
I do not watch,	3	4	7
TOTAL	14	26	40

Figure 4: my contigency table

- The percentages can either be horizontal, by response or vertical, by sex.

Exploring Two Categorical Variables: Contingency Tables

Table 5: Responses

	Female	Male
I do not watch,	0.5714286	0.4285714
I don't know	0.8333333	0.1666667
I like the programs,	0.6000000	0.4000000
I watch due to peer pressure,	0.6666667	0.3333333

Table 6: Responses

	Female	Male
I do not watch,	0.1538462	0.2142857
I don't know	0.1923077	0.0714286
I like the programs,	0.3461538	0.4285714
I watch due to peer pressure,	0.3076923	0.2857143