Chapter Two Part One: Displaying and Describing One Categorical Variable

Recall from Chapter One that categorical variables are variables that take on _____ as _____.

There are three main ways we extract information about categorical variables:

- **Counts(frequencies)**:

- **Proportions**:

- **Percentages**:

When we encounter a categorical variable in *the wild*, it is often useful to

1.

2.

This process is known as describing the _____ of the categorical variable. We can summarize the distribution of a single categorical variable by using data visualizations:

1. Frequency/Relative Frequency Table
2. Bar Chart
3. Pie Chart ☹

Frequency Table/Relative Frequency Table

Frequency Table:

Relative Frequency Table:

Example

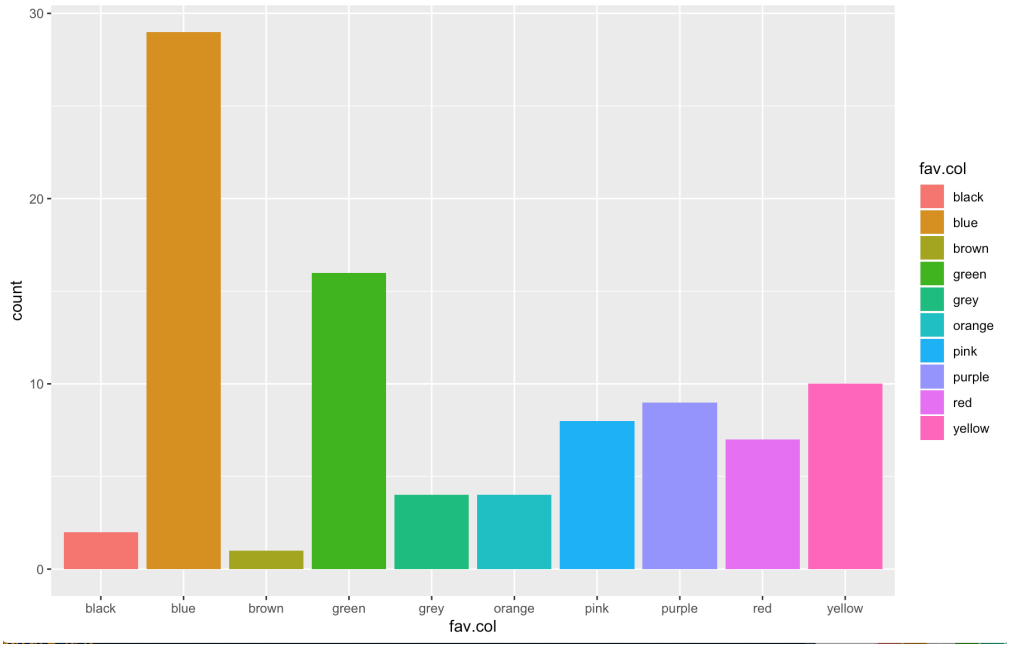
Here is the distribution of favourite colours for this STAT 101 class! Overwhelmingly blue is a favourite colour.

Favourite Colours	Frequency	Relative Frequency
Red	7	0.0778
Orange	4	0.0444
Yellow	10	0.111
Green	16	0.178
Blue	29	0.322
Purple	9	0.100
Pink	8	0.0889
Brown	1	0.0111
Grey	4	0.0444
Black	2	0.0222

Bar Chart

Important Characteristics

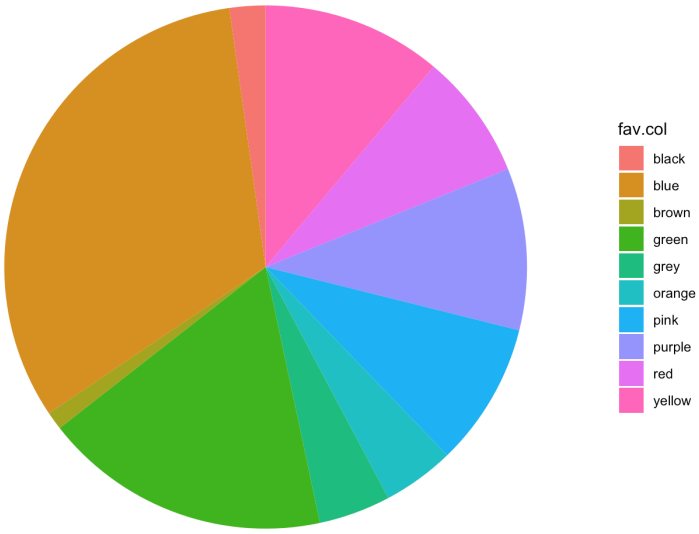
- Order of bars does not matter
- There are spaces between the bars
- Do not need to have all categories displayed
- does not necessarily show the shape of the distribution



Pie Chart ☹

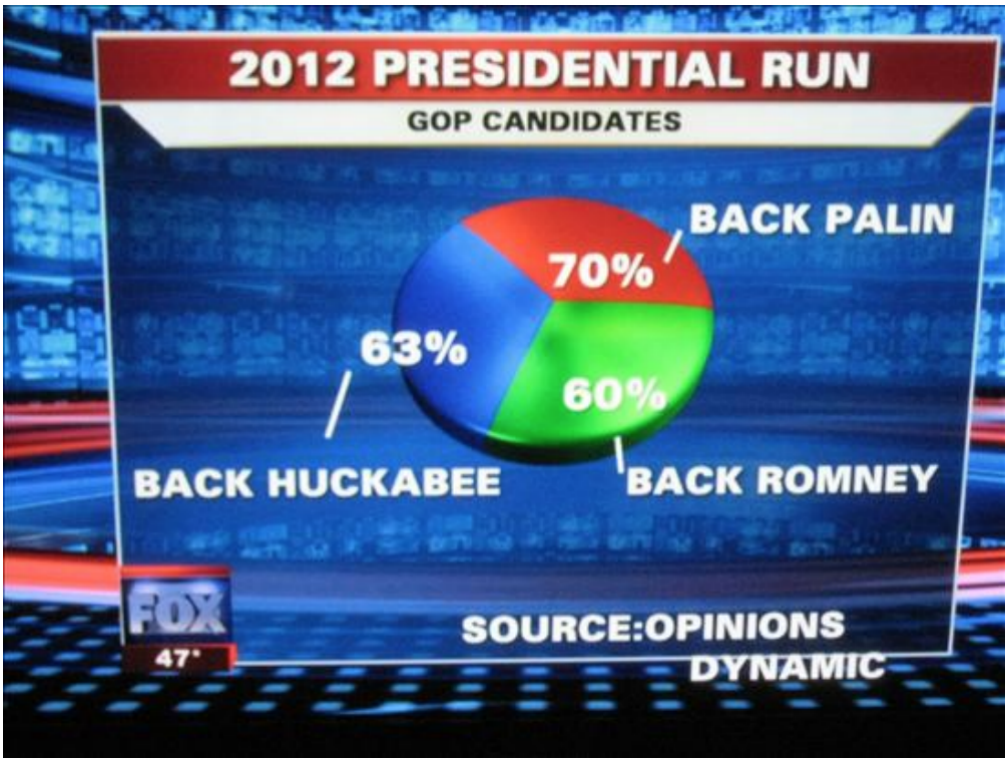
Important Characteristics

- Compare the size of the pie slices
- Often can provide percentages on the pie chart
- Percentages always add up to 100



Problems

- Only limited number of categories can really be captured
- Sometimes colours are hard to discern
- Not a super convincing datavisualization method



Chapter Two Part Two: Describing a Relationship between Two Categorical Variables

Response vs. Explanatory Variables

Often we are interested in studying how one variable can impact another variable. We call the variable of interest the _____. We call the variable being used to "explain" the differences in the distribution of the response variable the _____.

Example *Suppose we ask a random survey of Americans the following questions:*

- *What political party are you a member of?*
- *Who did you vote for in the last presidential election?*

What might the response and explanatory variable be in this case?

When our explanatory variable and response variable are both _____, we describe their relationship in terms of an _____.

- **Association:**

- **Independence:**

Example *Suppose we take a simple random sample of 100 eligible voters and ask them their political party and who they voted for. Based on the following frequency and table do you think there is an association between these two variables? (Hint: calculate relative frequencies to get a better idea.)*

Political Party	Candidate A	Candidate B	Count
Democrats	44	6	50
Republicans	19	31	50
Count	63	37	100

Would there still be an association if equal proportions of Democrats and Republicans voted for Candidate A over Candidate B?

Displaying Two Categorical Variables

There are two main ways we can display the relationship between two categorical variables:

- Contingency Tables
- Mosaic Plots

Contingency Tables

A contingency table is the cross-classification of observations according to the categories of two categorical variables.

A contingency table is built in the following way:

1.
2.
3.

Example *Below is a contingency table of 2068 students from STAT 101.*

	Eye Color					
Sex	Blue	Brown	Green	Hazel	Other	Total
Female	370	352	198	187	18	1125
Male	359	290	110	160	24	943
Total	729	642	308	347	42	2068

What proportion of students have blue eyes?

What proportion of students have blue or brown eyes?

What proportion of students have brown eyes and are female?

What percentage of female students have brown eyes?

Marginal Distributions

A _____ looks at counts or proportions for each variable separately and ignores the other variable.

	Eye Color					
Sex	Blue	Brown	Green	Hazel	Other	Total
Female	370	352	198	187	18	1125
Male	359	290	110	160	24	943
Total	729	642	308	347	42	2068

Notice how the marginal distribution for Eye Colour is in the _____ of the table.

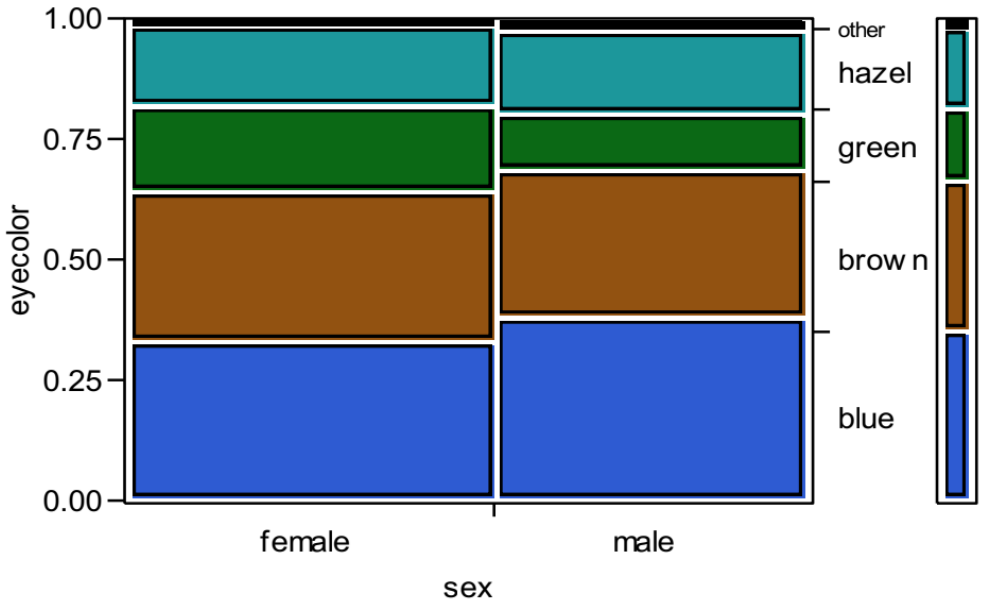
Conditional Distribution

A _____ looks at the counts or proportions for one variable contingent upon (conditioned on, given) a particular category of another variable. So for example, what might the distribution of eye colours look like, given that we are only interested in female students?

	Eye Color					
Sex	Blue	Brown	Green	Hazel	Other	Total
Female	370	352	198	187	18	1125
Male	359	290	110	160	24	943
Total	729	642	308	347	42	2068

Mosaic Plots

A mosaic plot is pretty much the same thing as a segmented bar plot. It visualizes the conditional distribution of the data and a summary of the marginal distribution. The _____ display the _____ of the variable on the vertical axis, conditioned on the value of the variable on the horizontal axis.



Differences in conditional distributions can indicate an association so Mosaic plots are great for establishing association.

- **Association:**

- **No Association:**

Do you think there is an association between eye colour and sex based on the mosaic plot above?

Chapter Two Summary

1. Ways of extracting information from categorical variables (frequency tables)
2. Displaying Categorial Variables (bar charts, pie charts ☹)
3. Response vs. Explanatory variables
4. Displaying Two Categorical Variables (contingency tables, mosaic plots)