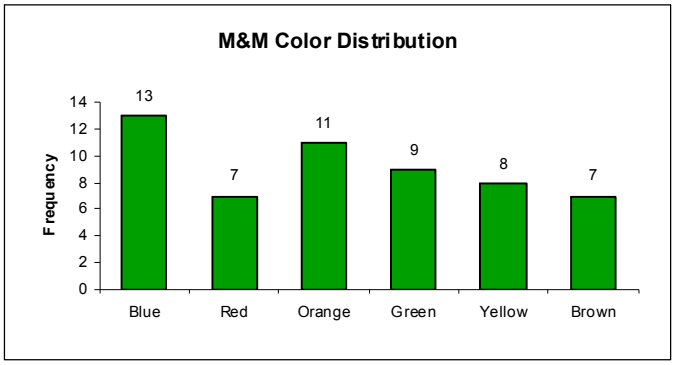
Notes: Displaying and Describing Categorical Data

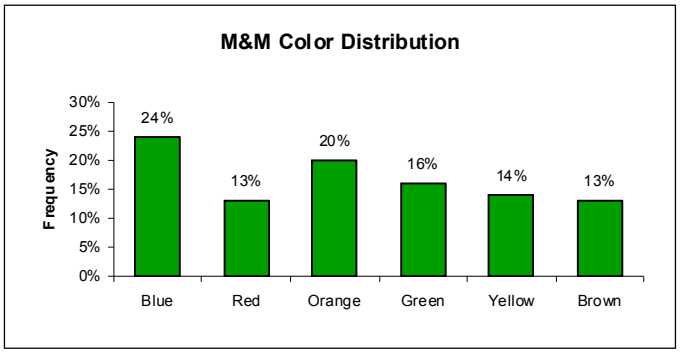
Frequency tables are often used to organize categorical data. Frequency tables display the category names and the counts of the number of data values in each category. Relative frequency tables also display the category names, but they give the percentages rather than the counts for each category.

Color	Freq.	Rel. Freq.	Percent
Blue	13	0.236	24%
Red	7	0.127	13%
Orange	11	0.200	20%
Green	9	0.164	16%
Yellow	8	0.145	14%
Brown	7	0.127	13%
TOTAL	**55**	**1.000**	**100%**

A bar chart is often used to display categorical data. The height of each bar represents the count for each category. Bars are displayed next to each other for easy comparison. When constructing a bar chart, note that the bars do not touch one another. Categorical variables usually cannot be ordered in a meaningful way; therefore the order in which the bars are displayed is often meaningless.



A relative frequency bar chart displays the proportion of counts for each category.



The sum of the relative frequencies is 100%.

A pie chart is another type of display used to show categorical data. Pie charts show parts of a whole. Pie charts are often difficult to construct by hand.

A contingency table shows two categorical variables together. The margins give the frequency distributions for each of the variables, also called the marginal distribution.

Examine the class data about gender and political view – liberal, moderate, conservative.

	Liberal	Moderate	Conservative	TOTAL
Male				
Female				
TOTAL				

- What percent of the class are girls with liberal political views?
- What percent of the liberals are girls?
- What percent of the girls are liberals?
- What is the marginal distribution of gender?
- What is the marginal distribution of political views?

A conditional distribution shows the distribution of one variable for only the individuals who satisfy some condition on another variable.

The conditional distribution of political preference, conditional on being male:

	Liberal	Moderate	Conservative	TOTAL
Male				

The conditional distribution of political preference, conditional on being female:

	Liberal	Moderate	Conservative	TOTAL
Female				

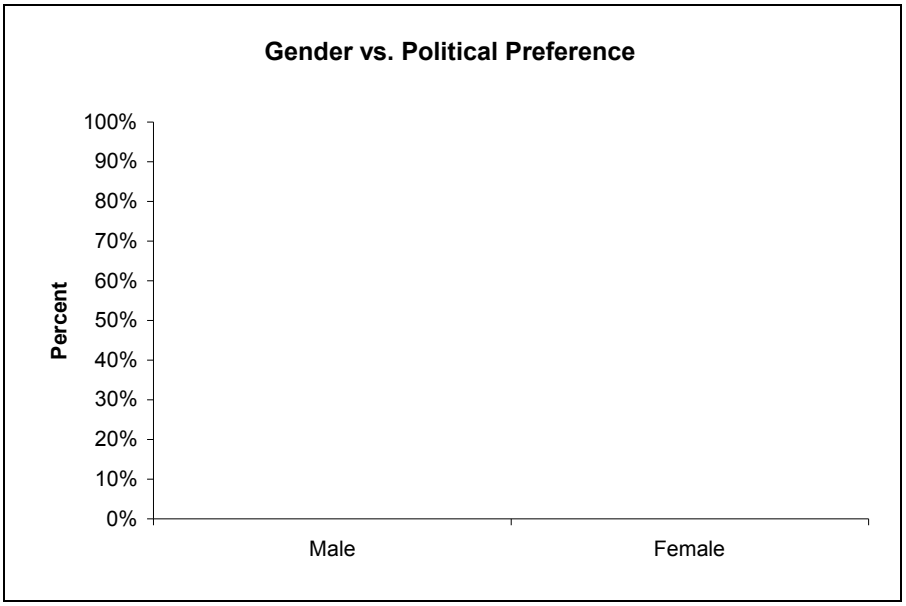
- What is the conditional relative frequency distribution of gender among conservatives?

If the conditional distributions are the same, we can conclude that the variables are not associated. Therefore, they are independent of one another.

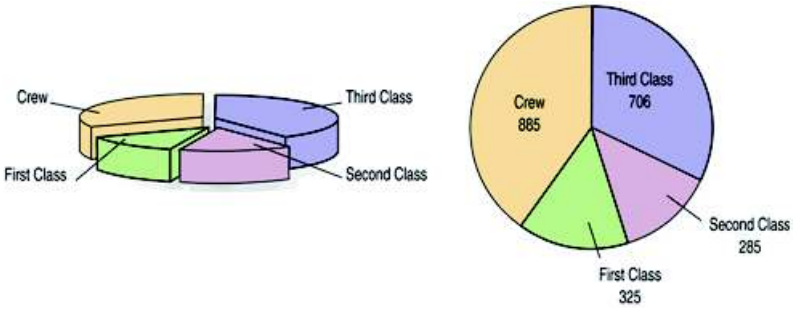
If the conditional distributions differ, we can conclude that the variables are somehow associated. Therefore, they are not independent of one another.

- Are gender and political view independent?

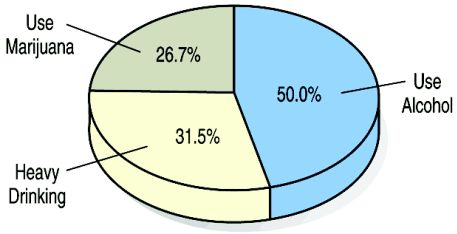
A segmented bar chart displays the same information as a pie chart, but in the form of bars instead of circles. Comparing segmented bar charts is a good way to tell if two variables are independent of one another or not.



- Explain how the graph on the left violates the "area principle."



- Explain what is wrong with the graph below.



Averaging one variable across different levels of a second variable can lead to **Simpson's Paradox**. Consider the following example:

It's the last inning of an important game. Your team is a run down with the bases loaded and two outs. The pitcher is due up, so you'll be sending in a pinch-hitter. There are 2 batters available on the bench. Whom should you send in to bat?

Player	Overall
A	33 for 103
B	45 for 151

- Compare A's batting average to B's batting average. Which player appears to be the better choice?
  Player A has a higher batting average (0.320 vs. 0.298), so he looks like the better choice.

Does it matter whether the pitcher throws right- or left-handed?

Player	Overall	vs LHP	vs RHP
A	33 for 103	28 for 81	5 for 22
B	45 for 151	12 for 32	33 for 119

- Compare A's batting average vs. a left-handed pitcher to B's. Compare A's batting average against a right-handed pitcher. Which player appears to be the better choice?
  Player B has a higher batting average against both right- and left-handed pitching, even though his overall average is lower. Player B hits better against both right- and left-handed pitchers. So no matter the pitcher, B is a better choice. So why is his batting "average" lower? Because B sees a lot more right-handed pitchers than A, and (at least for these guys) right-handed pitchers are harder to hit. For some reason, A is used mostly against left-handed pitchers, so A has a higher average.

Pooling the data together loses important information and leads to the wrong conclusion. We always should take into account any factor that might matter.