Chapter 3 - Displaying and Describing Categorical Data

August 25, 2010

Exploratory Data Analysis - The use of graphs or numerical summaries (values) to describe the variables in a data set and the relation between the variables.

Frequency Table (p. 21) - Table that organizes counts or frequencies into categories. A **Relative Frequency Table** (p. 22) uses the categories proportions or percents.

Color of Car Model	
Color	**Frequency**
White	135
Red	250
Blue	325
Green	190
	900

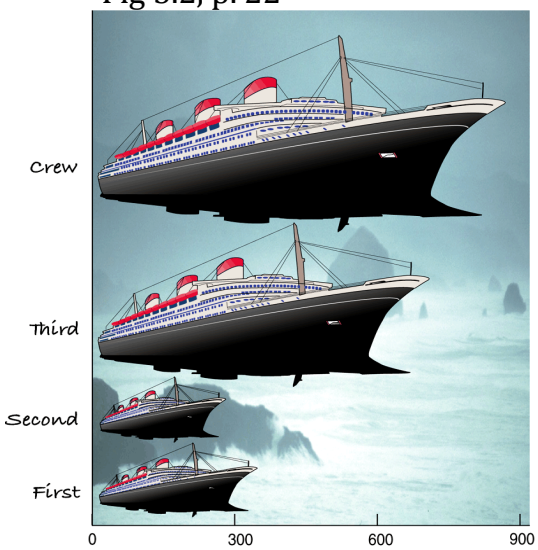
Color of Car Model	
Color	**Relative Frequency**
White	$\frac{135}{900} = 0.15$ or 15%
Red	$\frac{250}{900} = 0.278$ or 27.8%
Blue	$\frac{325}{900} = 0.361$ or 36.1%
Green	$\frac{190}{900} = 0.211$ or 21.1%

Both tables help describe the distribution of the categorical variable.

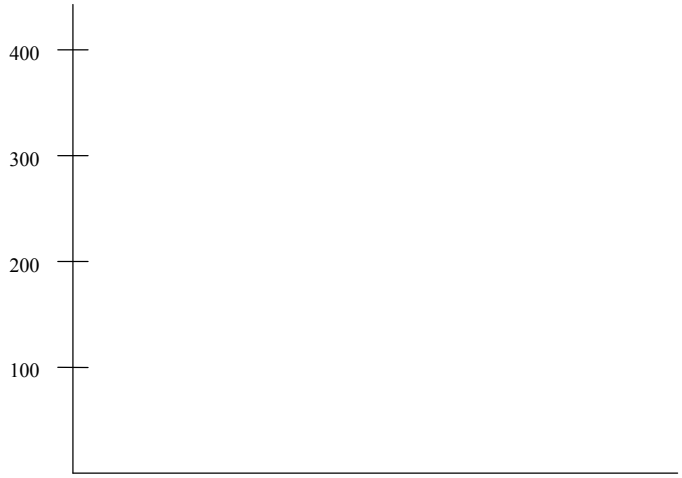
Distribution (p.22) - The pattern of variation of a variable. For categorical variables this will be how often categories occurs.

Area Principal (p. 22) - The area occupied by a part of the graph should correspond to the magnitude of the value it represents. This is important to remember because a graph can be misleading to the eye about the size of the area.

Fig 3.2, p. 22



Bar Chart (p. 23) - Graph that uses bars to represent the counts or proportion of each category. The bars have the same width, and the same amount of spaces between the bars. The space between the bars represent that the bars are freestanding and the bars can be arranged in any order. The bars can go up or to the side. Can be used to compare a few categories.



Pie Chart (p. 23) - Graph that shows the whole group as a circle, and the percent of each category is represented by a slice of the pie. All categories are present in the graph.

Contingency or Two-Way Table (p. 24) - Table that shows the distribution between two categorical variables.

> **Marginal distribution** (p. 25) is a frequency distribution of one of the variables along the margin of the contingency table. Frequencies or percentages can be used.
>
> **Conditional distribution** (p. 27) is a distribution of one variable that considers a smaller group of individuals that satisfy some condition on another variable.

The following is a table that compares the highest level of education of an individual with whether or not the individual smokes.

	Smoker	Nonsmoker	Total
High School	32	61	93
2-yr college	5	17	22
4+-yr college	13	72	85
Total	50	150	200

a) What is the percent of Nonsmokers who have completed 4+ years of college?

b) What is the percent of people that have completed 4+ years of college that are Nonsmokers?

	Smoker	Nonsmoker	Total
High School	32	61	93
2-yr college	5	17	22
4+-yr college	13	72	85
Total	50	150	200

c) What is the percent of Nonsmokers?

d) What is the percent of Smokers who have completed 4+ years of college?

e) What is the percent of Smokers given that they have completed 4+ years of college?

Side by Side Bar Chart (p.28) uses side by side bars of one categorical variable to show distribution versus other categorical variable.

Segmented Bar Chart (p. 31) - graph that treats each bar as a whole category and divides it proportionally into segments corresponding to the percentage of the other category.

Smoker	
High School	32 $\frac{32}{50} = 64\%$
2-yr college	5 $\frac{5}{50} = 10\%$
4+-yr college	13 $\frac{13}{50} = 26\%$
Total	50

Nonsmoker	
High School	61 $\frac{61}{150} = 40.7\%$
2-yr college	17 $\frac{17}{150} = 11.3\%$
4+-yr college	72 $\frac{72}{150} = 48\%$
Total	150

Simpson's Paradox (p. 35) - The comparison of different groups appears to contradict the results when the groups are combined into a single group.

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 two batters available. Batter A is 33 for 103 or an average of 0.320. Batter B is 45 for 151 for an average of 0.298.

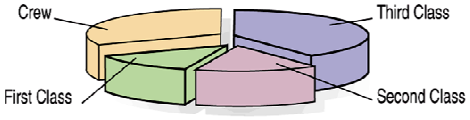
	vs. LHP	vs RHP
Batter A	28 for 81 or 0.346	5 for 22 or 0.227
Batter B	12 for 32 or 0.375	33 for 119 or 0.277

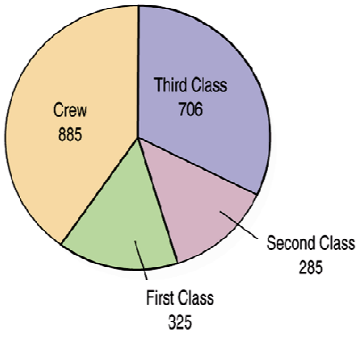
B is the better choice since B has a higher average against both LHP and RHP. However B faces many more RHP which have been harder to hit this season. A has faced mostly LHP.

Suppose you knew only that Batter A batted 0.227 against RHP and 0.346 against LHP what would A's overall average be. It would have to be a value between 0.227 and 0.346. B's average would fall between 0.277 and 0.375. Since these intervals overlap A could either be higher or lower than B.

What Can Go Wrong? (p.34)

- Don't violate the area principle.





- Make sure your display shows what it says it shows.