Chapter 1 Introduction &
1.1: Analyzing Categorical
Data

"Numbers don’t lie. That’s where we come in."

Introduction
Data Analysis: Making Sense of Data

After this section, you should be able to...

- DEFINE “Individuals” and “Variables”
- DISTINGUISH between “Categorical” and “Quantitative” variables
- DEFINE “Distribution”
- DESCRIBE the idea behind “Inference”

What is the Study of Statistics?!

- Statistics is the science of data.
- In this course we study four different aspects of statistics:
  - Data Analysis (Chapters 1 to 3)
    - The process of organizing, displaying, summarizing, and asking questions about data.
  - Data Collection (Chapter 4)
    - The process of conducting and interpreting surveys and experiments.
  - Anticipating Patterns/Probability (Chapter 5 to 7)
    - The process of using probability and chance to explain natural phenomena.
  - Inference (Chapter 8 to 12)
    - The process of making predications and evaluations about a population from a sample.

Variable - any characteristic of an individual or object

Categorical Variable
- Usually an adjective
- Rarely a number

Examples:
- Gender
- Race
- Grade in School (Sophomore, Jr., Sr.)
- Zip Code

Quantitative Variable
- Always a number
- Must be able to find the mean of the numbers

Examples:
- Weight
- Height
- GPA
- # of AP Classes taken
- Square footage

Distribution

- Distribution: describes what values a variable takes and how often it takes those values
- Essentially “distribution” replaces the words “data” or “graph”.
- The median of the distribution is 28.
- The distribution is skewed left.

Dotplot of MPG Distribution
Section 1.1
Analyzing Categorical Data

After this section, you should be able to...

- CONSTRUCT and INTERPRET bar graphs and pie charts
- RECOGNIZE “good” and “bad” graphs
- CONSTRUCT and INTERPRET two-way tables
- DESCRIBE relationships between two categorical variables
- ORGANIZE statistical problems

Distribution & Categorical Variables

The distribution of a categorical variable lists the count or percent of individuals who fall into each category.

<table>
<thead>
<tr>
<th>Favorite Course</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>8</td>
<td>16%</td>
</tr>
<tr>
<td>Foreign Language</td>
<td>4</td>
<td>8%</td>
</tr>
<tr>
<td>History</td>
<td>11</td>
<td>22%</td>
</tr>
<tr>
<td>Math</td>
<td>15</td>
<td>30%</td>
</tr>
<tr>
<td>Science</td>
<td>12</td>
<td>24%</td>
</tr>
</tbody>
</table>

Displaying Categorical Data

Frequency tables can be difficult to read. Sometimes it is easier to analyze a distribution by displaying it with a bar graph or pie chart.

Bar graphs compare several quantities by comparing the heights of bars that represent those quantities. Our eyes react to the area of the bars as well as height. Be sure to make your bars equally wide. Avoid the temptation to replace the bars with pictures for greater appeal...this can be misleading!

Graphs: Good and Bad

This ad for DIRECTV has multiple problems. How many can you point out?

Two-Way Tables

Two-Way Tables: describe two categorical variables, organizing counts according to a row variable and a column variable.

When a dataset involves two categorical variables, we begin by examining the counts or percents in various categories for one of the variables.

<table>
<thead>
<tr>
<th></th>
<th>Member of No Clubs</th>
<th>Member of One Club</th>
<th>Member of 2 or More Clubs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rides the School Bus</td>
<td>55</td>
<td>33</td>
<td>20</td>
<td>108</td>
</tr>
<tr>
<td>Does not Ride Bus</td>
<td>16</td>
<td>44</td>
<td>82</td>
<td>142</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
<td>77</td>
<td>102</td>
<td>250</td>
</tr>
</tbody>
</table>
• What proportion of males have “a good chance” at being rich?
• What proportion of females have a “50-50 chance” at being rich?
• What proportion of young adults that have an “almost certain” chance of being rich are male?

<table>
<thead>
<tr>
<th>Young adults by gender and chance of getting rich</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td></td>
</tr>
<tr>
<td>Almost no chance</td>
<td>96</td>
<td>98</td>
<td>194</td>
</tr>
<tr>
<td>Some chance, but probably not</td>
<td>426</td>
<td>286</td>
<td>712</td>
</tr>
<tr>
<td>A 50-50 chance</td>
<td>696</td>
<td>720</td>
<td>1416</td>
</tr>
<tr>
<td>A good chance</td>
<td>488</td>
<td>587</td>
<td>1063</td>
</tr>
<tr>
<td>Almost certain</td>
<td>586</td>
<td>875</td>
<td>1461</td>
</tr>
<tr>
<td>Total</td>
<td>2367</td>
<td>2469</td>
<td>4826</td>
</tr>
</tbody>
</table>

Comparing Categorical Distributions

<table>
<thead>
<tr>
<th>Sophomore</th>
<th>Junior</th>
<th>Senior</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Two</td>
<td>1</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>Three</td>
<td>4</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Four</td>
<td>7</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Five</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>14</td>
<td>33</td>
</tr>
</tbody>
</table>

Comparing Categorical Distributions

Writing to Compare Categorical Distributions

• Cite specific numerical values/proportions.
• Use comparison words.
  – Greater, smaller, less, while only, more, wider, narrower, fewer, etc.
• Use transition words
  – However, whereas, similarly, additionally, etc.
• Discuss at least two points of comparison.

Comparing Categorical Distributions

Is there an association between after-school club participation and whether or not the student rides the school bus? Support your answer with a discussion of the provided graphs.
Comparing Categorical Distributions

Sample Answer:
Yes, there is a clear association between after-school club participation and transportation. Only 11% of students who don’t ride the bus do not participate in after school clubs, whereas 51% of students who do ride the bus do not participate. Similarly, 58% of students who do not ride the bus are involved in 2 or more clubs, while only 19% of students riding the bus are involved in 2 or more clubs. However, the proportion of students who participate in one club is the same for students who ride and students who don’t ride the bus.