Chapter 2

Displaying and Describing Categorical Data
Homework

p31 5, 6, 7, 8, 9, 10, 21, 22, 23, 25, 31, 32
Your Turn
The Three Rules of Data Analysis

1. **Draw a picture:** An illustration of the data can show any trends, patterns, or unusual characteristics of the collection that is not obvious in a simple list of the data.

2. **Draw a picture:** Patterns in the data can be seen in a visual representation of the data and you may see things that you would otherwise miss.

3. **Draw a picture:** A picture is worth a thousand words. Not sure I agree with that, but a picture allows you to more easily explain your words.
Frequency Tables

We can organize the data by counting the number of data values some category of interest.

We organize the counts into a frequency table which simply records the category names and the total frequency within each category.

<table>
<thead>
<tr>
<th>Class</th>
<th>Count Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>325</td>
</tr>
<tr>
<td>Second</td>
<td>285</td>
</tr>
<tr>
<td>Third</td>
<td>706</td>
</tr>
<tr>
<td>Crew</td>
<td>885</td>
</tr>
</tbody>
</table>
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**Frequency Tables**

A **relative frequency table** is virtually the same, but gives the percentages or proportion for each category in place of the absolute count in the category.

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>325</td>
</tr>
<tr>
<td>Second</td>
<td>285</td>
</tr>
<tr>
<td>Third</td>
<td>706</td>
</tr>
<tr>
<td>Crew</td>
<td>885</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>(f/total)100</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>14.77</td>
</tr>
<tr>
<td>Second</td>
<td>12.95</td>
</tr>
<tr>
<td>Third</td>
<td>32.08</td>
</tr>
<tr>
<td>Crew</td>
<td>40.21</td>
</tr>
</tbody>
</table>

\[
\frac{325}{2201} = 0.147660154... 
\]

We are going to experience relative frequencies often in this course. **Relative frequency** is a proportion, percent, and empirical probability.
Frequency Tables

- Both types of tables show how cases are distributed across the categories.

- Frequency tables illustrate the distribution of a categorical variable as the table names the possible categories and indicates how frequently each category occurs.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

Anything Wrong With This Picture?

Here is an artistic, clever way to show the number of people on the Titanic:

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>325</td>
</tr>
<tr>
<td>Second</td>
<td>285</td>
</tr>
<tr>
<td>Third</td>
<td>706</td>
</tr>
<tr>
<td>Crew</td>
<td>885</td>
</tr>
</tbody>
</table>

We will not be clever, or artistic in examining data!
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

The Area Principle

- The ship display makes it look like most of the people on the Titanic were crew members, with a few passengers along for the ride.

- When looking at each ship, we react to the area taken up by the ship, instead of the length of the ship. The length is the only valid characteristic.

- The ship display violates the area principle:

  - The area occupied by a part of the graph should match the magnitude of the value it represents.

  - Do not get clever, creative, cute, or fancy.....
Do NOT get fancy

A bar chart displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.

Thus, a much better display for the ship data is:

This bar chart stays true to the area principle.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**Bar Charts**

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

- Simply replace absolute counts with percentages in the data:

- A relative frequency bar chart also stays true to the area principle.

- The only difference between a frequency bar chart and a relative frequency bar chart is the vertical axis labels.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**Pie Charts**

- If you are interested in illustrating the relative size of parts of the whole, a **pie chart** might be your best choice.

- Pie charts show all the categories as sections of a circle.

- Pie charts divide the circle into sections whose size is a proportional fraction of the whole for each category.
Contingency Tables (Two Way Tables)

- A **contingency table** allows us to look at two categorical variables together.

- **Important**: We will be referring to contingency table several times later in this course. So remember this topic.

A contingency table shows how data frequencies are distributed for one variable, contingent upon each level of the other variable.

- In our example we can examine the class of ticket contingent upon the survival condition:

<table>
<thead>
<tr>
<th>Survival</th>
<th>Class</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Crew</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>203</td>
<td>118</td>
<td>178</td>
<td>212</td>
<td></td>
<td>711</td>
</tr>
<tr>
<td>Dead</td>
<td>122</td>
<td>167</td>
<td>528</td>
<td>673</td>
<td></td>
<td>1490</td>
</tr>
<tr>
<td>Total</td>
<td>325</td>
<td>285</td>
<td>706</td>
<td>885</td>
<td></td>
<td>2201</td>
</tr>
</tbody>
</table>
Contingency Tables

- The margins of the table, both on the right and on the bottom, give totals and the frequency distributions for each of the variables.

- The frequency distribution for each variable is called a marginal distribution of its respective variable in the contingency table.

The marginal distribution of Survival is:

<table>
<thead>
<tr>
<th>Survival</th>
<th>Alive</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>711</td>
<td>1490</td>
</tr>
</tbody>
</table>

The marginal distribution of Class is:

<table>
<thead>
<tr>
<th>Class</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Crew</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>325</td>
<td>285</td>
<td>706</td>
<td>885</td>
</tr>
</tbody>
</table>
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**Contingency Tables**

- Each **cell** of the table gives the count for a **combination** of two conditions.
- For example, the second cell in the crew column tells us that 673 crew members died when the Titanic sunk.

<table>
<thead>
<tr>
<th>Survival</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Crew</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td>203</td>
<td>118</td>
<td>178</td>
<td>212</td>
<td>711</td>
</tr>
<tr>
<td>Dead</td>
<td>122</td>
<td>167</td>
<td>528</td>
<td>673</td>
<td>1490</td>
</tr>
<tr>
<td>Total</td>
<td>325</td>
<td>285</td>
<td>706</td>
<td>885</td>
<td>2201</td>
</tr>
</tbody>
</table>

- crew and died
Conditional Distributions

A conditional distribution shows the distribution of one variable for just the individuals who satisfy a single condition on another variable.

The following is the conditional distribution of ticket Class, conditional on having survived:

<table>
<thead>
<tr>
<th>Survival</th>
<th>Class</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>Crew</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alive</td>
<td></td>
<td>203</td>
<td>118</td>
<td>178</td>
<td>212</td>
<td>711</td>
</tr>
<tr>
<td>Dead</td>
<td></td>
<td>122</td>
<td>167</td>
<td>528</td>
<td>673</td>
<td>1490</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>325</td>
<td>285</td>
<td>706</td>
<td>885</td>
<td>2201</td>
</tr>
</tbody>
</table>
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables, and bar graphs.

## Conditional Distributions

The following is the conditional distribution of ticket Class, conditional on having perished:

<table>
<thead>
<tr>
<th>Survival</th>
<th>Class</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
<td>Second</td>
</tr>
<tr>
<td>Alive</td>
<td>203</td>
<td>118</td>
</tr>
<tr>
<td>Dead</td>
<td>122</td>
<td>167</td>
</tr>
<tr>
<td>Total</td>
<td>325</td>
<td>285</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Class</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First</td>
</tr>
<tr>
<td>Dead</td>
<td>122</td>
</tr>
<tr>
<td></td>
<td>8.2%</td>
</tr>
</tbody>
</table>
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

## Conditional Distributions

- The conditional distributions tell us that there is a difference in the distribution of class for those who survived and those who perished.

- This is easily seen with pie charts of the two distributions:

- Note the obvious differences in section sizes between the two charts.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**Conditional Distributions**

- We see that the distribution of Class for the survivors is different from that of the non-survivors.

- This leads us to believe that Class and Survival are associated. That is, they are not independent.

- The variables would be considered independent when the distribution of categories of one variable in a contingency table is essentially the same for all categories of any other variables in the table.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

Segmented Bar Charts

A **segmented bar chart** displays the same information as a pie chart.

Here is the segmented bar chart for ticket Class by Survival status:

Each bar is treated as the “whole” and is divided proportionally into segments corresponding to the percentage in each group.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

Side-by-Side

Each bar corresponds to a “pie”. The information portrayed is the same in each picture but the emphasis is slightly different.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

Simpson’s Paradox

- Simpson’s paradox is a result of averaging done when averages can be misleading.
- Let us peek into the lives of two waiter’s in a local eatery, Gyade and JuChi.

Gyade and JuChi are competing for promotion to night manager. The restaurant manager decides to look at tip count to measure customer satisfaction.

<table>
<thead>
<tr>
<th></th>
<th>Gyade</th>
<th>JuChi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lunch</td>
<td>50 meals - $100</td>
<td>100 meals - $300</td>
</tr>
<tr>
<td>Dinner</td>
<td>100 meals - $600</td>
<td>50 meals - $400</td>
</tr>
<tr>
<td>Total</td>
<td>150 meals - $700</td>
<td>150 meals - $700</td>
</tr>
</tbody>
</table>
Simpson’s Paradox

Based on the marginal distribution of waiter it appears Gyade and JuChi are equally well regarded by customers.

Perhaps we should look a little closer.

Gyade averages $2/lunch and $6/dinner for tips.
JuChi averages $3/lunch and $8/dinner in tips.

Averaging tips across lunch and dinner is unreasonable and is an example of Simpson’s Paradox.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

What NOT to Do.

- Do not violate the area principle. In other words, do NOT get cute.

Some people might like the pie chart on the left better because of the three-dimensional effect. But it is much more difficult to compare fractions of the whole, which is the primary purpose of a pie chart.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

What NOT to Do.

Make certain your display is honest, and not intended to fool the reader. Your display should show what it purports to show.

This plot of the percentage of high-school students who engage in specified dangerous behaviors has a few problems.

List the problems you see.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**What NOT to Do.**

Does this make sense?

How common is red hair?

<table>
<thead>
<tr>
<th>Fair</th>
<th>Red</th>
<th>Medium</th>
<th>Dark</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>1455</td>
<td>286</td>
<td>2137</td>
<td>1391</td>
<td>116</td>
</tr>
</tbody>
</table>

Total = 5387
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

What NOT to Do.

- Depends on what you mean by “Fair”.

### Hair Color

<table>
<thead>
<tr>
<th></th>
<th>Fair</th>
<th>Red</th>
<th>Medium</th>
<th>Dark</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>27.01%</td>
<td>5.31%</td>
<td>39.67%</td>
<td>25.88%</td>
<td>21.9%</td>
</tr>
</tbody>
</table>

Total = 5387

$$\frac{1455}{5387} \times 100 = 27.01\%$$

Caithness County, Scotland
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**How to Lie With Statistics**

- Statistics do not lie, but there are people who will, intentionally or unintentionally, mislead you. This is especially true of using graphs.

- View graphs through knowledgeable eyes. Ask how the data was collected, from whom the data was collected, when was the data collected, where did you get your data, and most importantly, why are you being shown the graph.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

See anything wrong?

- 9.0% = 8.6%?
- The maximum differential is only 0.6% (less than 1%).
- Actually indicating a downward trend.
- Note the scale along the vertical axis.
**What Global Warming?**

- **1st, still inside 95% prediction.**

- **2nd, this graph shows air temp? What about ocean temp where most heat resides?**

- **3rd, and most importantly, we never trust extrapolation past data. We have no certainty of what comes next.**
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

What Global Warming?

- Sometimes, starting the y axis at zero hides important changes.
Let Me Illustrate

Here is an example we can all understand. Who has the fever?

Aah! Now we can see. Note the y-axis.

Anyone in here is toast, or better a popsicle.
What Recovery?

- The economy is a mess. Just look, the growth of the Gross Domestic Product (GDP) is flat.

- Well, you knew this was coming. I guess once Obama took office there was significant growth.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

College is a Bad Investment?

What about the earnings of those not attending college (Who)? Maybe the differential is worth the investment.

In this graphic the cost of a 4 year degree is compared directly to the average 1st year salary. To truly determine the value of that college education, we must compute the expected increase in earnings over not going to college after a lifetime of work.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

Graduation Rates Are Improving

- Here is another close to home. There are several problems with this graphic.

- **DO NOT** illustrate elements in a graph with pictures (the books here).

- 5 books equal 75%, thus one book is 15%. 82% should be 5.4667 books.

- This is how the bar chart should look.

- But the most serious problem is that the bar chart is not the appropriate graph for this data and what it is intended to show.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

**Time Series Graph**

- When the goal is to show changes over time (time series graph) it is preferable to use a frequency polygon (line chart, line graph).

- So yes, graduation rates did increase during Obama’s term, but that increase had begun around ‘96 during the Clinton and Bush years.
Objective: Students organize and describe distributions of categorical data by using frequency tables, pie charts, contingency tables and bar graphs.

Here Is My Personal Favorite!

The only explanation for this graph is an intentional attempt to mislead the reader.

Gun deaths in Florida

Number of murders committed using firearms

Source: Florida Department of Law Enforcement
C. Chan 16/02/2014