<table>
<thead>
<tr>
<th>ID</th>
<th>Number of Cell Calls</th>
<th>Text While Driving</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>208</td>
<td>5</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>227</td>
<td>5</td>
<td>2</td>
<td>54</td>
</tr>
<tr>
<td>262</td>
<td>2</td>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>207</td>
<td>20</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>56</td>
<td>3</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>248</td>
<td>1</td>
<td>1</td>
<td>57</td>
</tr>
<tr>
<td>65</td>
<td>7</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>194</td>
<td>5</td>
<td>2</td>
<td>51</td>
</tr>
<tr>
<td>132</td>
<td>10</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>110</td>
<td>10</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>41</td>
<td>10</td>
<td>2</td>
<td>29</td>
</tr>
<tr>
<td>109</td>
<td>5</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>219</td>
<td>3</td>
<td>2</td>
<td>53</td>
</tr>
<tr>
<td>89</td>
<td>10</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>283</td>
<td>12</td>
<td>1</td>
<td>62</td>
</tr>
<tr>
<td>144</td>
<td>20</td>
<td>2</td>
<td>46</td>
</tr>
<tr>
<td>138</td>
<td>6</td>
<td>2</td>
<td>45</td>
</tr>
<tr>
<td>168</td>
<td>15</td>
<td>2</td>
<td>48</td>
</tr>
<tr>
<td>113</td>
<td>5</td>
<td>2</td>
<td>42</td>
</tr>
<tr>
<td>26</td>
<td>10</td>
<td>2</td>
<td>25</td>
</tr>
</tbody>
</table>
Would you be surprised if…

<table>
<thead>
<tr>
<th>A population looked like this</th>
<th>but the sample looked like this?</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Population" /></td>
<td><img src="image2.png" alt="Sample" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Population" /></td>
<td><img src="image4.png" alt="Sample" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Population" /></td>
<td><img src="image6.png" alt="Sample" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Population" /></td>
<td><img src="image8.png" alt="Sample" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Population" /></td>
<td><img src="image10.png" alt="Sample" /></td>
</tr>
</tbody>
</table>
Sampling Variability Early and Often!

Roxy Peck
Cal Poly, San Luis Obispo
rpeck@calpoly.edu
Sampling Variability

One of the most important concepts in the introductory statistics course.

Foundation for understanding

- Meaning of statistical significance
- Meaning of margin of error
The “Usual” Approach (without Technology)

Wait until Chapter X when you get to the dreaded topic of sampling distributions…

Then

1. Ask students to “imagine…”
2. Show a static picture of the distribution of $\bar{x}$
3. Pray that they understand the relevance and importance (they don’t).
4. Hope they can get through the rest of the course devoted to inference (some can, but usually by following rote procedures—the logic escapes them, but oh well…)
The “Usual” Approach (with Technology)

Wait until Chapter X when you get to the dreaded topic of sampling distributions…

Then

1. Show the students an applet (there are lots of good ones out there). Technology replaces the “imagine…” part of the no technology approach. **An improvement!**
2. Pray that they understand the relevance and importance (some do with less left to imagination!). **Some do, but...**
3. Hope they can get through the rest of the course devoted to inference (some can, but usually by following rote procedures—the logic escapes them, but oh well...) **Not much changes...**
Sampling variability and sampling distributions are central to understanding statistics.

These are the most abstract concepts we ask students to grapple with in the intro stat course.

SO

Why do we keep this hidden until we spring it on students late in the course (with or without technology) and expect them to get it?!?

Wouldn’t it be better to develop and nurture this idea from the beginning?

* In My Humble Opinion, with a nod to Chris Olsen!
There is a potential BIG payoff

BUT much easier to do if

- Plan for it in a thoughtful way.
- Have access to technology to facilitate this type of consistent development.
The Missed Opportunities

- Chapter 1: Sampling Methods
- Chapter 3: Graphical Methods
- Chapter 4: Numerical Summary Measures
- Chapter 5: Regression
- Chapter 8: Sampling Distributions
- Chapter 9: Confidence Intervals
- Chapter 10: Hypothesis Testing
Activities Based on Random Sampler

- Every student gets a different random sample from the same population.

- Students engage with the data and with other students to build understanding.
Random Samplers

- Many possibilities
  - Census@School database has a random sampler
    (http://www.amstat.org/censusatschool/)
  - Random sampling applets
    (one example: Sampling from a finite population
    http://www.rossmanchance.com/applets/OneSample.html)
  - Textbook resources
  - Statistical software packages with random sampling capabilities (such as Minitab and JMP)
Classroom Activities

- You can have students use technology to select their own samples, or (like we will do here today) you can preselect samples and distribute them to students.

- Examples that follow are based on sampling from specific populations but can be easily adapted to any population.

- The important point is what you ask the students to do and when.
Sampling

- Sample from a population of adults. Each student gets a different random sample of 20.

- Variables: number of cell phone calls made in a typical day, response to “have you ever sent a text message while driving?”
<table>
<thead>
<tr>
<th>ID</th>
<th>Number of cell calls</th>
<th>Text while driving</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>4</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>198</td>
<td>6</td>
<td>2</td>
<td>51</td>
</tr>
<tr>
<td>30</td>
<td>35</td>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>230</td>
<td>10</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>77</td>
<td>2</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>258</td>
<td>10</td>
<td>2</td>
<td>58</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>2</td>
<td>24</td>
</tr>
<tr>
<td>157</td>
<td>2</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>99</td>
<td>2</td>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>274</td>
<td>20</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>126</td>
<td>5</td>
<td>2</td>
<td>44</td>
</tr>
<tr>
<td>135</td>
<td>10</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>106</td>
<td>6</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>67</td>
<td>5</td>
<td>1</td>
<td>33</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>134</td>
<td>10</td>
<td>1</td>
<td>45</td>
</tr>
<tr>
<td>246</td>
<td>10</td>
<td>2</td>
<td>57</td>
</tr>
<tr>
<td>88</td>
<td>20</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>55</td>
<td>4</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>217</td>
<td>7</td>
<td>2</td>
<td>53</td>
</tr>
</tbody>
</table>
What can we ask students to think about...

- Make a dot plot using the data for number of cell phone calls. (Use the scale provided on the handout).
- What does the dot plot tell you about number of cell phone calls for the people in your sample?
- How is your dot plot similar to the following dot plot for the entire population of 300 people? How is it different?
Note: because there were 300 individuals in this population, some dots in the population graph represent up to 3 individuals.

Each symbol represents up to 3 observations.
And we could also ask...

- If you had taken a random sample of 50 people from this population instead of 20, do you think that your dotplot would have looked more like the population dotplot or less like the population dotplot? Why do you think this?
And we could also ask...

- Compare your dotplot to the dotplots of the other groups at your table.
  - Why aren’t all of the dotplots the same?
  - Even though not all of the sample dotplots are the same, in what ways are they similar?
And we could also ask...

- Now calculate the average number of cell phone calls for the people in your sample.
- The average number of cell phone calls made by the people in this population is 10.033. Did you get 10.033 for your sample average? Does this surprise you? Explain why or why not.
- Did everyone in the class get the same value for the sample average? Is this surprising?
And we could also ask…

- Calculate the proportion of people in your sample who said that they had sent or read a text message while driving by counting the number of yes responses (1 = yes) and dividing by 20.

- For the entire population, the proportion who responded yes is $120/300 = 0.40$. Is this what you got for your sample proportion? Does this surprise you?
And we could also ask…

- Did everyone in the class get the same sample proportion? Why did this happen?
- Of all the sample proportions, what was the smallest? What was the largest? Does this sample-to-sample variability surprise you?
- Suppose that everyone had selected a random sample of 50 people instead of 20 people. Do you think that the sample proportions for these new samples would have varied more or varied less from sample to sample than what we saw for samples of size 20?
Graphical Methods

- Each student gets a different random sample of size 50 from a population of adults

- Variables: Gender, age
What can we ask students to think about...

- Flip a coin two times and record the outcome (H or T) for each toss. These outcomes will determine which of the following class intervals you will use to create a frequency distribution and histogram of your age data.

<table>
<thead>
<tr>
<th>Coin Toss Outcomes</th>
<th>Use Class Intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH</td>
<td>15 to &lt; 25, 25 to &lt; 35, 35 to &lt; 45, 45 to &lt; 55, 55 to &lt; 65</td>
</tr>
<tr>
<td>HT</td>
<td>18 to &lt; 28, 28 to &lt; 38, 38 to &lt; 48, 48 to &lt; 58, 58 to &lt; 68</td>
</tr>
<tr>
<td>TH</td>
<td>15 to &lt; 20, 20 to &lt; 25, 25 to &lt; 30, ..., 55 to &lt; 60, 60 to &lt; 65</td>
</tr>
<tr>
<td>TT</td>
<td>18 to &lt; 23, 23 to &lt; 28, 28 to &lt; 33, ..., 58 to &lt; 63, 63 to &lt; 68</td>
</tr>
</tbody>
</table>
And we could ask students...

- Using the age data for the people in your sample, construct a frequency distribution.
- Draw a histogram of the sample age data.
- Describe the shape of the age histogram.
- Here is a histogram that displays the age distribution of the entire population. How is your histogram similar to the population histogram? How is it different?
And we could also ask...

- Compare your histogram to a histogram produced by another student who used different class intervals. How are the two histograms similar? How are the two histograms different?

- How do the histograms produced by the class support the following statement: Histograms based on random samples from a population tend to look like the population histogram.
And we could also ask…

- Now let’s compare the age distribution of the males and the age distribution of the females in your sample.
- Flip a coin to determine what type of comparative graphical display you will construct. If your flip results in a head, you will use a comparative dotplot. If your flip results in a tail, you will use a comparative stem-and-leaf plot. Which type of graphical display will you be constructing?
- Using the data in your sample, construct either a comparative dotplot or a comparative stem-and-leaf plot (depending on the outcome of the coin flip) that allows you to compare males and females. You will probably have different numbers of males and females in your sample, but this is not something to worry about.
- Based on your graphical display, write a few sentences comparing the two age distributions.
And we could also ask…

- Compare your graphical display to that of another student who used a different display. Did you both make similar statements when you compared the two age distribution even though you each used different random samples and constructed different types of displays? Does this surprise you?
Big Payoffs

- Illustrates
  - Effect of choice of bins on appearance of histogram.
  - Even with different bins and different random samples, there are similarities and they still resemble the population distribution in key ways.
  - Even with different types of graphical displays and different random samples, tend to draw the same conclusions.
  - All this and we are weeks away from formal treatment of sampling distributions!
Follow up Activity

Would you be surprised if...

<table>
<thead>
<tr>
<th>A population looked like this</th>
<th>but the sample looked like this?</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Population Graph" /></td>
<td><img src="image2" alt="Sample Graph" /></td>
</tr>
<tr>
<td><img src="image3" alt="Population Graph" /></td>
<td><img src="image4" alt="Sample Graph" /></td>
</tr>
<tr>
<td><img src="image5" alt="Population Graph" /></td>
<td><img src="image6" alt="Sample Graph" /></td>
</tr>
<tr>
<td><img src="image7" alt="Population Graph" /></td>
<td><img src="image8" alt="Sample Graph" /></td>
</tr>
<tr>
<td><img src="image9" alt="Population Graph" /></td>
<td><img src="image10" alt="Sample Graph" /></td>
</tr>
<tr>
<td><img src="image11" alt="Population Graph" /></td>
<td><img src="image12" alt="Sample Graph" /></td>
</tr>
</tbody>
</table>

AMATYC 2017
Regression

- Each student gets a different random sample from a population of adults.

- Variables: Number of text messages sent, age
We could ask students to think about...

- Find the equation of the least squares regression line that could be used to describe the relationship between
  \[ y = \text{number of text messages sent} \]
  and
  \[ x = \text{age} \]
  for your sample.
We could also ask students to think about...

Compare your regression equation with those of the other groups at your table.

- Why didn’t every one get the same regression equation?
- If two different groups at your table were to each use their regression equation to predict the number of text messages sent for a person who is 28 years old, would the predictions be the same? Do you think they would be similar?
- If every group had obtained a random sample of 50 people instead of 20, do you think the regression equations would have differed more or less from sample to sample than they did for samples of size 20?
Illustrates

- Sampling variability in the slope of the regression line
- Potential impact of sampling variability on predictions
Sampling Distributions

- Students are ready for these ideas!
- In fact, they have already done what you will now formalize several times and should be comfortable with the idea of sampling variability.
And it Continues…

- Confidence intervals
- Hypothesis tests
Confidence Intervals

Compare your confidence interval to the confidence interval obtained by another student.

- Are the two confidence intervals the same? Does this surprise you?
- Did both intervals contain the actual population proportion of 0.41?
- How many people in your class have a confidence interval that does not include the actual value of the population proportion? Is this surprising given the 90% confidence level associated with the confidence intervals?
Hypothesis Tests

Compare your p-value to the p-value obtained by another student.

- Are the two p-values the same? Does this surprise you?
- Did you both reach the same conclusion based on your random samples?
- For this population, the correct conclusion would be to fail to reject the null hypothesis. How many people in your class rejected the null hypothesis? Is this surprising given the significance level of 0.10 used for the test?
So, don’t leave it until Chapter X!

Engage with the ideas of sampling variability...

EARLY AND OFTEN!
Thanks for attending this session.

Questions?

Comments?

Contact information:

rpeck@calpoly.edu