**Announcements**

- Practice quiz on OpenIntro this weekend - to give you a chance to see what they'll be like and to test out your OpenIntro account.
  - To register go to [http://www.openintro.org/login.php](http://www.openintro.org/login.php)
  - Course ID: STAT101S12, Access code: 6VQ3T
  - Quiz will be open from Thursday 5 pm to Monday 8 am (note slight change from what was stated on the syllabus initially)
  - 5 multiple choice or short numerical answer questions each quiz
  - 1 hr to finish, must finish when you start, can’t save work
  - To be completed on your own, you may not collaborate with another student or anyone else, honor code applies
  - This week’s quiz not graded
- HW1 posted - due Feb 2.

---

**Review question**

How would you describe the shape of the distribution of number of piercings college students have?

(a) right skewed
(b) left skewed
(c) symmetric
(d) uniform

---

**Contingency table**

A table that summarizes data for two categorical variables is called a **contingency table**.

<table>
<thead>
<tr>
<th>major</th>
<th>cheat no</th>
<th>cheat yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>arts and humanities</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>natural sciences</td>
<td>15</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>social sciences</td>
<td>20</td>
<td>21</td>
<td>41</td>
</tr>
<tr>
<td>other</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>41</strong></td>
<td><strong>36</strong></td>
<td><strong>77</strong></td>
</tr>
</tbody>
</table>

**Note:** The survey question on cheating was worded as “Have you ever cheated on an assignment or exam?”
### Clicker question

Does there appear to be a relationship between major and whether or not a student cheated?

<table>
<thead>
<tr>
<th>major</th>
<th>cheat</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td>Total</td>
</tr>
<tr>
<td>arts and humanities</td>
<td>5</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>natural sciences</td>
<td>15</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td>social sciences</td>
<td>20</td>
<td>21</td>
<td>41</td>
</tr>
<tr>
<td>other</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>36</td>
<td>77</td>
</tr>
</tbody>
</table>

(a) yes  
(b) no

---

### Segmenting bar plots

A poll conducted by the Pew Research Foundation asked 2,142 people from the general public and 1,055 college presidents if they thought online courses offer an equal educational value compared with courses taken in a classroom. 621 people from the general public and 538 college presidents answered yes. Based on these data, does it appear that the general public and college presidents have differing opinions on this issue?

(a) Yes, the general public is more likely to view online education as comparable to classroom education.  
(b) Yes, college presidents are more likely to view online education as comparable to classroom education.  
(c) No, the two groups do not appear to have differing opinions on this issue.  
(d) We cannot tell from the information given.

http://pewresearch.org/pubs/2092/online-courses-students-colleges-universities-technology-laptops-tablets

---

### Bar plots

A bar plot is a common way to display a single categorical variable. A bar plot where proportions instead of frequencies are shown is called a relative frequency bar plot.
Considering categorical data

Segmented bar plots

Mosaic plots

Does there appear to be a relationship between gender and having used Adderall for an exam or to study?

Clicker question

Does there appear to be a relationship between gender and opinion on gay marriage?

Note: The survey question on gay marriage was worded as “Should gay marriage be legal?”

Side-by-side box plots

Does there appear to be a relationship between how much students dread this semester and their relationship status?

Clicker question

Based on the relative frequency histograms below, which of the following is the best estimate for the median number of Duke basketball games attended for students who were a Duke fan before becoming students here?

Note: 1 - not at all, 5 - a lot
Based on the relative frequency histograms below, compare the distributions of the number of Duke basketball games attended for students who were and were not Duke fans becoming students here?

Consider the following three research questions:

1. What is the average mercury content in swordfish in the Atlantic Ocean?
2. Over the last 5 years, what is the average time to degree for UCLA undergraduate students?
3. Does the drug sulphinpyrazone reduce the number of deaths in heart attack patients?

Consider the following three research questions:

1. A man on the news got mercury poisoning from eating swordfish, so the average mercury concentration in swordfish must be dangerously high.
2. I met two students who took more than 10 years to graduate from UCLA, so it must take longer to graduate at UCLA than at many other colleges.
3. My friend's dad had a heart attack and died after they gave him sulphinpyrazone. The drug must not work.

These are all anecdotal evidence, based on a limited sample size that might not be representative of the population.

Anecdotal evidence is typically composed of unusual cases that we recall based on their striking characteristics.

Anti-smoking research started in the 1930s and 1940s when cigarette smoking became increasingly popular.

While some smokers seemed to be sensitive to cigarette smoke, others were completely unaffected.

Anti-smoking research was faced with resistance based on anecdotal evidence such as “My uncle smokes three packs a day and he’s in perfectly good health”.

Back then it was concluded that “smoking is a complex human behavior, by its nature difficult to study, confounded by human variability.”

In time researchers were able to examine larger samples of cases (smokers) and trends showing that smoking has negative health impacts became much clearer.
Wouldn’t it be better to just include everyone and “sample” the entire population?
- Such a special sample is called a **census**.

There are problems with taking a census:
- It can be difficult to complete a census: there always seem to be some individuals who are hard to locate or hard to measure. And there may be a certain characteristic about those individuals who are hard to locate.
- Populations rarely stand still. Even if you could take a census, the population changes while you work, so it’s never possible to get a perfect measure.
- Taking a census may be more complex than sampling.

Sampling is natural
- Sampling is a natural...
- Think about sampling something you are cooking - you taste (examine) a small part of what you’re cooking to get an idea about the dish as a whole.

Exploratory analysis to inference
- When you taste a spoonful of soup and decide it doesn’t taste salty enough, that’s **exploratory analysis**.
- If you generalize and conclude that your soup needs salt, that’s an **inference**.
- For your inference to be valid the spoonful you tasted (the sample) needs to be **representative** of the entire pot (the population).
  - If your spoonful comes only from the surface and the salt is collected at the bottom of the pot, what you tasted is probably not representative of the whole pot.
  - If you first stir the soup thoroughly before you taste, your spoonful will more likely be representative of the whole pot.
Random vs. biased sampling

- If we want to examine characteristics of all Duke undergraduates, we should take a random sample from the population.
- We could do this by getting a list of all Duke undergraduates from the registrar’s office and randomly selecting a number of students from that list.
- If we employ a method where all students are equally likely to be selected, this would be a simple random sampling that is representative of the population.

Can you suggest such a method?

- If instead we asked first-year students to sample a number of Duke students and they sample only from students in their classes and dorm, the sample would be biased towards first-year students.

A few sources of bias

- Non-response bias: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.
- Voluntary response bias: Occurs when those who respond have strong opinions on the issue since such a sample will also not be representative of the population.
- Convenience sample: Individuals who are easily accessible are more likely to be included in the sample.

Landon vs. FDR

A historical example of a biased sample yielding misleading results:

In 1936, Landon sought the Republican presidential nomination opposing the re-election of FDR.

The Literary Digest Poll

- The Literary Digest polled about 10 million Americans, and got responses from about 2.4 million.
- The poll showed that Landon would likely be the overwhelming winner and FDR would get only 43% of the votes.
- Election result: FDR won, with 62% of the votes.
- The magazine was completely discredited because of the poll, and was soon discontinued.
The Literary Digest Poll - what went wrong?

- The magazine had surveyed
  - its own readers,
  - registered automobile owners, and
  - registered telephone users.
- These groups had incomes well above the national average of the day (remember, this is Great Depression era) which resulted in lists of voters far more likely to support Republicans than a truly typical voter of the time, i.e. the sample was not representative of the American population at the time.

The Literary Digest election poll was based on a sample size of 2.4 million, which is huge, but since the sample was biased, the sample did not yield an accurate prediction.

Back to the soup analogy: If the soup is not well stirred, it doesn’t matter how large a spoon you have, it will still not taste right. If the soup is well stirred, it doesn’t matter whether you have a large or small spoon, it will taste fine either way.

Overview of data collection principles
Explanatory and response variables

Explanatory and response variables

- To identify the explanatory variable and the response variable in a pair of variables, identify which of the two is suspected of affecting the other.

  explanatory variable might affect response variable

- Labeling variables as explanatory and response does not guarantee the relationship between the two is actually causal, even if there is an association identified between the two variables.

Clicker question

A study is designed to test the effect of type of light on exam performance of students. 180 students are randomly assigned to three classrooms: one that is dimly lit, another with yellow lighting, and a third with white fluorescent lighting and given the same exam. Which is correct?

(a) explanatory: dimly lit, yellow, white fluorescent
response: exam performance

(b) explanatory: exam performance
response: dimly lit, yellow, white fluorescent

(c) explanatory: type of light (categorical with 3 levels)
response: exam performance

(d) explanatory: exam performance
response: type of light (categorical with 3 levels)
Observational studies and experiments

- **Observational study**: Researchers collect data in a way that does not directly interfere with how the data arise, i.e. they merely “observe”, and can only establish an association between the explanatory and response variables.

- **Experiment**: Researchers randomly assign subjects to various treatments in order to be able to establish causal connections between the explanatory and response variables (also called a randomized experiment).

- If you're going to walk away with one thing from this class, let it be “correlation does not imply causation”.

---