Agenda

1. Data Journalism: Income Mobility by Colleges
2. Data and Sampling
3. HW #1 due Wednesday (Name on Moodle)
4. First lab on Thursday!

Activity: Data Collection

1. Find two poeple who you have never met, someone other that the person you worked with last time!
2. Take turns answering the following questions:
   (a) What is your name?
   (b) What is your email address?
   (c) What year are you?
   (d) In which house do you live?
   (e) What is your hometown?
   (f) How many siblings do you have?
   (g) What is the furthest away from Northampton (in miles) you were over the summer?

   While one person is answering, the other two groupmates will write down their answers.
3. Next, open this Google Spreadsheet, and start entering data.
4. What are the *cases* in this data set?

5. For each of the variables in the Spreadsheet, describe the type of variable that it is (e.g. categorical/numerical, discrete/continuous, ordinal, etc.)
   (a) Name

   (b) Sheet Color

   (c) Class Year

   (d) House

   (e) Hometown

   (f) # of Siblings

   (g) Distance over break

Sampling It is important to keep in mind the distinction between the *population* and the *sample*. We collect a sample of data, analyze it, and try to use that information to make inferences about the population.

Three sampling schemes: simple random sampling, stratified sampling, and cluster sampling (see Figure 1.14 on page 15)

1. Suppose that in a company there are the following 180 staff members: 90 women who work full-time, 18 women who work part-time, 9 men who work full-time, and 63 men who work part-time. We are asked to take a sample of 40 staff, stratified according to the above categories. Devise a sampling scheme to do this.

2. A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. Briefly assess the strengths and weaknesses of each approach. Which approach would likely be the *least* effective? Why?

   (a) Simple random sampling

   (b) Cluster sampling

   (c) Stratified sampling

   (d) Anecdotal sampling

3. A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true? Why?

   (a) Some of the mailings may have never reached the parents.

   (b) The school district has strong support from parents to move forward with the policy approval.

   (c) It is possible that the majority of the parents of high school students disagree with the policy change.

   (d) The survey results are unlikely to be biased because all parents were mailed a survey.