Exploring Sampling Concepts

Data

We will consider data from many different studies, but the main one will be the `evals` dataset from course evaluations at UT Austin.

```
Rows: 463
Columns: 5
$ prof_ID <int> 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5,~
$ score   <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.5, 4.~
$ age     <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 40, 40~
$ gender  <fct> female, female, female, female, male, male, male, male, male, ~
$ rank    <fct> tenure track, tenure track, tenure track, tenure track, tenure~
```

Deciding a Study Population

Often when reading a statistical study, you will not know who the researchers thought of as their target population, and are left to imagine who the study population may have been.

Comment on the study population for the following studies:

1. `hsb2`: "Two hundred observations were randomly sampled from the High School and Beyond survey, a survey conducted on high school seniors by the National Center of Education Statistics."

2. `evals`: "The data on the 463 courses at UT Austin."

Random Sampling

3. What is the central tenant of random sampling? i.e. How are observations selected from the study population?

4. Comment on the "random-ness" of the following studies:

- A Cal Poly administrator wants to know the average income of all graduates in the last 10 years. So they get the records of five randomly chosen graduates, contact them, and obtain their answers.

- You want to know the prevalence of illegal downloading of TV shows among students at a local college. You get the emails of 100 randomly chosen students and ask them, "How many times did you download a pirated TV show last week?".

Sampling Randomly

5. Suppose we have a database of every professor at UT Austin, and are interested in studying the relationship between evaluations and age.

- How would we go about randomly sampling observations from the database?

- Would we allow for individuals to be selected more than once?

- Would we expect that our sample look like the population at UT Austin?

Representative Sampling

6. What is the central tenant of representative sampling? i.e. How are observations selected from the study population?

7. Comment on the "representative-ness" of the following studies:

- The Royal Air Force wants to study how resistant all their airplanes are to bullets. They study the bullet holes on all the airplanes on the tarmac after an air battle against the Luftwaffe (German Air Force).

- You want to know the average number of people in each household in your city. You randomly pick out 500 phone numbers from the phone book and conduct a phone survey.

Sampling Representatively

gender	rank	perc
female	teaching	6%
female	tenure track	20%
female	tenured	10%
male	teaching	4%
male	tenure track	4%
male	tenured	56%

- Female teaching = 28%
- Female tenure track = 15%
- Female tenured = 12%

8. Suppose we want to ensure that we have a representative proportion of faculty who are women and faculty of different tenure statuses. How would we go about collecting a representative sample that accounts for these demographic characteristics?

Sampling Issues

In the `evals` data, the courses included in the dataset were only taught by 94 unique professors.

```
count(evals, prof_ID)
```

```
# A tibble: 94 x 2
   prof_ID     n
     <int> <int>
 1       1     4
 2       2     3
 3       3     2
 4       4     8
 5       5     6
 6       6     7
 7       7     5
 8       8     7
 9       9     7
10      10    10
# ... with 84 more rows
```

9. How would you use sampling to remedy this situation?

Population Parameter

To investigate the relationship between course evaluation score and the professor's age, we would carry out a simple linear regression.

10. What would be the population parameter we are interested in? Do we know its value?

11. What would be our point estimate? What is its purpose?

Repeated Samples

Repeated samples are necessary for us to create a sampling distribution like the one below.

12. What are we assuming when we plot the result of every sample on the **same** distribution?