Questions about the Assignment

Collecting Data:
Sampling

Population vs. Sample
A population includes all individuals or objects of interest.
A sample is all the cases that we have collected data on; usually a subset of the population.
Statistical inference is the process of using data from a sample to gain information about the population.

The Big Picture

Population vs. Sample
What are some reasons why researchers would gather data from a sample of the population rather than gather data on the entire population?

Potential Problems with Sampling
Gore Takes FL in 2000
8pm
Bush Takes FL in 2000
2am
**Sampling Bias**

Sampling bias occurs when the method of selecting a sample causes the sample to differ from the population in some relevant way.

In what ways could the sample have been biased in the projected FL election results?

If sampling bias exists, we cannot make (accurate) inferences about the population based on the sample.

**Sampling**

Suppose you want to know what proportion of Duke students took a course during the summer session.

What is the population?

In what ways would the sample be biased if the sample was...
- every student in this class?
- every student who is an athlete?
- every student in the sociology department?

What would be a good (i.e., unbiased) sample to use in order to make inferences about the population?

**Take a Random Sample**

We can avoid sampling bias by taking a random sample?

Imagine putting the name of every Duke student into a hat, and drawing out names at random to be in the sample.

**Random Sampling**

Before the 2008 election, the Gallup Poll took a random sample of 2,847 Americans. 52% of those sampled supported Obama.

In the actual election, 53% of Americans voted for Obama.

2,847 is less than 1/1000th of 1% of the U.S. population.

Random sampling is a very efficient and powerful tool!!!
Selecting a Random Sample

Option 1: Actually draw names out of a hat
Option 2: Number all of the units in the population and generate random numbers to decide which unit gets sampled

Online Random Number Generator: www.random.org/integers

Random vs. Non-Random Sampling

Random samples provide averages that are centered around the correct number.

Random samples provide the best estimates for making generalizations to the population.

Non-random samples may suffer from sampling bias, and averages may not be centered around the correct number.

Jelly Bean Analogy

Think about a bowl of jelly beans…
Population = entire bowl of beans
Sample = whatever is in the handful you grab

If you grab non-randomly from the bowl (e.g., if you prefer black jelly beans), you may not get a very accurate representation of the jelly beans in the bowl.

If you grab randomly (e.g., with your eyes closed), the beans in your hand will be an accurate representation of the beans in the bowl.

Simple Random Sample

These methods generate a simple random sample
In a simple random sample, each unit of the population has the same chance of being selected.

Sampling Challenges

While a random sample is ideal, it often isn’t feasible.

A list of the entire population may not be available.
It may be too difficult to contact all members of the population.

Sometimes your population of interest has to be altered to something more feasible to sample from.
In these cases, the inferences you make based on the results are limited to the population that was actually sampled from.

Quasi-Random Sampling

Suppose you want to estimate the average number of hours that Duke students spend studying each week.

What are the potential problems associated with the proposed sampling techniques?

Go to Perkins Library and ask all the students there how much they study.
Sampling units based on something obviously related to the variables you are analyzing.

Email all Duke students asking how much they study, and use all the data you get.
Letting your sample be comprised of whoever chooses to participate (volunteer bias).
Data Collection and Bias

Even with a random sample, data can still be biased, especially when collected on people.

Bias exists when the method of collecting data causes the sample data to inaccurately reflect the population.

Other forms of bias to watch out for in data collection include:

1. **Non-Response**
   - If the people who respond to the survey tend to answer differently than the people who do not respond, the results will be biased.
   - What is your income level?
   - Please rate your experience at our restaurant?

2. **Question Wording**
   - Do you support gay marriage?
   - Do you think the government should restrict same-sex partners from getting married?
   - A black interviewer asking a white respondent survey questions related to racism.
   - Asking questions about sensitive information that the respondent may not want to publicly disclose.

3. **Context**
   - In the last year, how many times have you attended a religious service?
     - Once a week
     - Once a month
     - 1-2 times
     - Never
   - How old were you when your first child was born?
     - In what year was your first child born?
     - How old is your first child?
Social Desirability

In the last year, have you cheated on any academic assignment or exams?

How many hours a week do you study?

Other Forms of Bias

Be aware of the other forms of bias that may exist and appropriately account for them.

Assignment

Part I
Graded Problems
1.40 and 1.44

Additional Practice Problems (not to be turned in):  
1.39 and 1.41

Part II
Goto http://sda.berkeley.edu/cgi-bin/hda?harcsda+gss10 and find 10 variables of interest. Provide the variable name and the question associated with the variable.

Summary

Data is collected on a sample of the population, and we use this data to make inferences to the larger population.

Sampling bias can occur when the sample is not representative of the population.

Sampling bias can be avoided by random sampling.

Data bias exists when the sample data do not accurately reflect the true population data.

When making conclusions based on data, examine the study for potential sampling and data bias.

Think critically about how the data were collected, and recognize that not all forms of data collection lead to valid inferences.