Unit 1: Introduction to data
1. Data Collection +
Observational studies & experiments

Sta 101 - Fall 2016
Duke University, Department of Statistical Science

Dr. Mukherjee Slides posted at http://www2.stat.duke.edu/courses/Fall16/sta101.001/

1. Use a sample to make inferences about the population
   - Ultimate goal: make inferences about populations
   - Caveat: populations are difficult or impossible to access
   - Solution: use a sample from that population, and use statistics from that sample to make inferences about the unknown population parameters
   - The better (more representative) sample we have, the more reliable our estimates and more accurate our inferences will be

Suppose we want to know how many offspring female lemurs have, on average. It’s not feasible to obtain offspring data from on all female lemurs, so we use data from the Duke Lemur Center. We use the sample mean from these data as an estimate for the unknown population mean. Can you see any limitations to using data from the Duke Lemur Center to make inferences about all lemurs?

Sampling is natural

- When you taste a spoonful of soup and decide the spoonful you tasted isn’t salty enough, that’s exploratory analysis
- If you generalize and conclude that your entire 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)

2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier

**Simple random:**
Drawing names from a hat

**Cluster:** heterogeneous clusters
Sample all chosen clusters

**Stratified:** homogenous strata
Stratify to control for SES

**Multistage:**
Random sample in chosen clusters
Clicker question

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. Which approach would likely be the least effective?

(a) Simple random sampling
(b) Stratified sampling, where each stratum is a neighborhood
(c) Cluster sampling, where each cluster is a neighborhood

3. Sampling schemes can suffer from a variety of biases

- **Non-response**: 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**: Occurs when the sample consists of people who volunteer to respond because they 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

4. What type of study is this? What is the scope of inference (causality / generalizability)?

**Facebook Tinkers With Users’ Emotions in News Feed Experiment, Stirring Outcry**

*By VINDU GOEL  JUNE 29, 2014*

In an academic paper published in conjunction with two university researchers, the company reported that, for one week in January 2012, it had altered the number of positive and negative posts in the news feeds of 689,003 randomly selected users to see what effect the changes had on the tone of the posts the recipients then wrote.

The researchers found that moods were contagious. The people who saw more positive posts responded by writing more positive posts. Similarly, seeing more negative content prompted the viewers to be more negative in their own posts.

4. Experiments use random assignment to treatment groups, observational studies do not.

A study that surveyed a random sample of otherwise healthy adults found that people are more likely to get muscle cramps when they’re stressed. The study also noted that people drink more coffee and sleep less when they’re stressed. What type of study is this?

What is the conclusion of the study?

Can this study be used to conclude a causal relationship between increased stress and muscle cramps?

5. Four principles of experimental design: randomize, control, block, replicate.

We would like to design an experiment to investigate if increased stress causes muscle cramps:
- Treatment: increased stress
- Control: no or baseline stress

It is suspected that the effect of stress might be different on younger and older people: block for age.

Why is this important? Can you think of other variables to block for?

6. Random sampling helps generalizability, random assignment helps causality.

<table>
<thead>
<tr>
<th>Ideal experiment</th>
<th>Random assignment</th>
<th>No random assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random sampling</td>
<td>Causal conclusion, generalized to the whole population.</td>
<td>No causal conclusion, correlation statement generalized to the whole population.</td>
</tr>
<tr>
<td>No random sampling</td>
<td>Causal conclusion, only for the sample.</td>
<td>No causal conclusion, correlation statement only for the sample.</td>
</tr>
</tbody>
</table>

Generalizability

- Random sampling: Causal conclusion, generalized to the whole population.
- No random sampling: No causal conclusion, correlation statement generalized to the whole population.

Causation

- Random assignment: Causal conclusion, generalized to the whole population.
- No random assignment: No causal conclusion, correlation statement only for the sample.

Correlation

1. Use a sample to make inferences about the population
2. Ideally use a simple random sample, stratify to control for a variable, and cluster to make sampling easier
3. Sampling schemes can suffer from a variety of biases
4. Experiments use random assignment to treatment groups, observational studies do not
5. Four principles of experimental design: randomize, control, block, replicate
6. Random sampling helps generalizability, random assignment helps causality