

**STA13-B**  
**Elementary Statistics**  
**Fall 2007**

**Lecture 2**

**Instructor: Katie Pollard**

# If you missed the first class...

- Class policies & assignments are on the course website (syllabus & lecture 1 slides):

<http://www.stat.ucdavis.edu/~kpollard/sta13/>

- Email: [s13b@wald.ucdavis.edu](mailto:s13b@wald.ucdavis.edu)
- Come to office hours if you have questions.

# Announcements

- 3 Exams: F 10/26, M 11/19, F 12/14  
Plan ahead (no early or late make-up exams)!
- Waitlist
- Textbook
- How to write up homework (due Thurs)

# Chapter 2

- Sources of sampling bias
- Sampling methods
- Study designs
  - Observational
  - Experimental
- Confounding

# Study Design

- A **study design** is an overall plan for collecting data.
- Should reflect the objectives of the study.
- Affects the type of conclusions that can be drawn from the results:
  - Generalizing to a population,
  - Inferring cause and effect.

# Sources of Sampling Bias

- A sample should be a good representation of the population.
- Systematic differences can occur at each step of the data collection process:
  1. Selection of experimental units for the sample
  2. Obtaining data from each experimental unit
  3. Measurement of the variable of interest

# 1. Selection Bias

Systematic exclusion/inclusion of individuals (experimental units) from the sample.

## Examples:

- People who give birth at home excluded from a maternity ward interview
- Mountain lions that live deep in wilderness areas missed in a study of nests observed by a ranger
- Beef-lovers more likely to fill out a survey if a steak house gift certificate is offered as a reward

## 2. Nonresponse Bias

Systematic exclusion/inclusion of individuals (experimental units) from data collection.

### Examples:

- Difficult to sequence regions of a genome
- People who work full time outside the home in a residential phone survey
- Organizations that file their taxes late in a study of tax forms from an accountant's office 4/15

## 3. Measurement Bias

Observed data differs systematically from the true value. Also known as **response bias**.

### Examples:

- People will tell their dentist that they floss more often than they actually do
- Faulty voting equipment
- "Would you vote to deprive our children of a decent education?" vs. "Are you in favor of a property tax increase?"

# Consequences of Bias

## Demonstration:

- Population=everyone in this class
- Sample=everyone in the front row today
- Variables of interest:
  - number of homework #1 problems completed
  - proportion of females
- Data collection: In class interview by Prof. Pollard on Wednesday

# Sampling Techniques

General tips for all methods

- Use a **sampling frame** or list of all experimental units in the population
- A **random number generator** can pick the sample from the sampling frame
- With or without **replacement**
- Avoid **convenience samples!**

# Sampling Methods

## 1. Simple random sampling (size $n$ )

- Every possible sample of size  $n$  has equal probability of being selected

## 2. Systematic sampling (1 in $k$ )

- Randomly pick one of the first  $k$  entries in the sampling frame
- Select every  $k$ 'th entry until the end
- OK if there is no repeating pattern in list

# Sampling Methods

## 3. Stratified random sampling

- strata=non-overlapping, homogeneous subgroups of the population
- Select a simple random sample from each strata

## 4. Cluster sampling

- clusters=non-overlapping, heterogeneous subgroups of the population
- Select a simple random sample of clusters and include all experimental units in these clusters

# Sampling Example

Four ways to sample 50 students from the population of all 350 students in STA13B:

1. **Simple random:** Number students 1-350 and pick 50 random numbers
2. **Systematic:** Pick 1 from first 7 on class list (alphabetic), then every 7th
3. **Stratified:** Random 25 males and 25 females
4. **Cluster:** All students in a random section

# Causal Inference

- Often one is interested in the relationship between two variables.

**Example:** Women's Health Initiative Data  
variable 1=heart attack rate in women  
variable 2=estrogen taken (Yes or No)

- Suppose an **association** is observed. Can a **cause-and-effect** conclusion be made?

# Observation vs. Experimentation

The ability to make a causal inference (or not) depends on how the data was collected.

- **Observational**=in the normal course of events
- **Experimental**=after manipulating events

**Example (cont.):**

observational: random sample of women

experimental: give estrogen to a random half the sample and a **placebo** to the other half

# Problem with Observational Data

- Suppose estrogen has no effect, but that a low fat diet reduces heart attacks by 50%.
- What if a higher percentage of estrogen-takers happen to eat a low fat diet?

	Estrogen=Y	Estrogen=N
Low fat=Y	300	200
Low fat=N	100	400
Total	400	600

# Advantage of Experimental Data

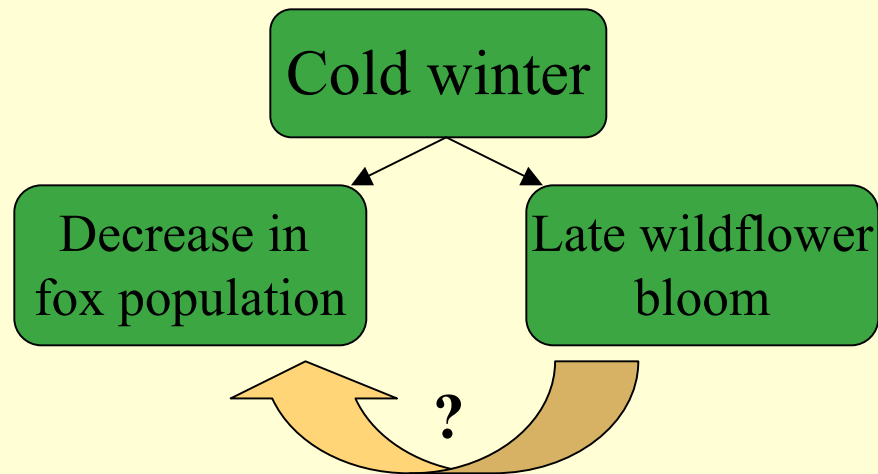
- In a well designed experimental study, the values of a confounding variable should be approximately the same in both groups.

	Estrogen=Y	Estrogen=N
Low fat=Y	120/400=30%	200/600=33.3%
Low fat=N	280/400=70%	400/600=67.7%
Total	400	600

In this case, correct **causal inference** is possible.

# Confounding

A **confounding variable** can create an observed association between two unrelated variables.



If it is not measured, a confounding variable can lead to incorrect conclusions.