

## Statistics 111 - Lecture 3

### Collecting Data

#### Surveys and Sampling

January 24, 2007

Stat 111 - Lecture 3 - Sampling

1

### Administrative Notes

- First recitation this Friday
- Homework 1 available today on website
  - Due in recitation on Friday, February 8th

January 24, 2007

Stat 111 - Lecture 3 - Sampling

2

### Outline for Lecture

- Another Example of Confounding
- Introduction to Sampling
  - Voluntary Response Samples
  - Simple Random Samples
- Sources of Sampling Bias
- More complicated sampling schemes
- Preview of Inference
- Bias versus Variability

January 24, 2007

Stat 111 - Lecture 3 - Sampling

3

### Confounding Example: Simpson's Paradox

- The relationship between two variables can change dramatically after considering a third confounding variable
- Example: observational study of UC Berkeley graduate admissions in 1971

Males Applicants		Female Applicants	
Number	Admitted	Number	Admitted
8442	44 %	4321	35 %

- Seems to indicate a sex bias towards males

January 24, 2007

Stat 111 - Lecture 3 - Sampling

4

### Confounding Example: Simpson's Paradox

- Trend seems to reverse when we subdivide applicants into their different departments

Department	Males Applicants		Female Applicants	
	Number	% Admitted	Number	% Admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

- A greater proportion of females applying to departments with lower acceptance rates
- Association between gender and admission rates does not imply causation !

January 24, 2007

Stat 111 - Lecture 3 - Sampling

5

### Survey Definitions

- **Population:** entire group of objects or people about which information is sought
- **Census:** survey of an entire population
- **Sample:** survey that examines only a portion of the population
- **Parameter:** a numerical characteristic of the population
- **Statistic:** a numerical characteristic of the sample

January 24, 2007

Stat 111 - Lecture 3 - Sampling

6

## Why Sample?

- **Expense: cheaper than a census**
  - Nielson ratings: based on 5000 out of an estimated 105.5 million US households with TVs
- **Time: quicker than a census**
  - Exit polls: gives news agencies valuable (?) information on election day in order to project election before all votes (census) are counted
- **Sampled units must sometimes be destroyed (or changed) to measure characteristics**
  - Reliability studies: testing lifetime of light bulbs, strength of windshields, etc.

January 24, 2007

Stat 111 - Lecture 3 - Sampling

7

## Sampling Bias

- Systematic errors that result in a sample that is not representative of the overall population of interest
- Just like in experiments, we must be cautious of potential sources of bias in our sampling results

January 24, 2007

Stat 111 - Lecture 3 - Sampling

8

## Voluntary Response Samples

- People choose to be included in sample themselves by responding to a general appeal
  - Eg. Amazon consumer ratings
- Results are often biased because people with strong opinions (usually negative) are more likely to respond and be included in the sample

January 24, 2007

Stat 111 - Lecture 3 - Sampling

9

## Hite Report: Women and Love (1987)

- Hite mailed 100,000 questionnaires to groups of women professionals, counseling centers, church societies, senior citizens centers. Only 4.5% were returned
- 84% of women are "not satisfied emotionally with their relationships" (p. 804)
- 70% of all women "married five or more years are having sex outside of their marriages (p. 856)
- 95% of women "report forms of emotional and psychological harassment from men with whom they are in love relationships" (p. 810)
- 84% of women report forms of condescension from the men in their love relationships (p. 809)

January 24, 2007

Stat 111 - Lecture 3 - Sampling

10

## Simple Random Sampling (SRS)

- Just as an experiment can be improved by randomization, so can sampling
- Each individual in the population has an equal chance of being included in the sample
- Does not allow self-response or evaluators to influence makeup of the survey (kinda like double-blinding in experiments)

January 24, 2007

Stat 111 - Lecture 3 - Sampling

11

## Example: Presidential Elections

- In 1912 *Literary Digest* began using surveys to predict US presidential elections
  - "The poll represents 30 years constant evolution and perfection..."
- In the 1936 Roosevelt vs Landon election, they polled 10 million voters:
  - 1,293,669 said they would vote for Landon
  - 972,897 said they would vote for Roosevelt
- Reality: Landslide victory (61% to 37%) for FDR
- What went wrong?

January 24, 2007

Stat 111 - Lecture 3 - Sampling

12

## Biases in Random Samples

- Randomization doesn't correct for certain problems with sampling
- **Bias 1: Undercoverage:** some groups in the population are left out of the process of choosing the sample
- **Bias 2: Nonresponse:** sampled individuals can not be contacted or do not cooperate
- Eg. 1936 presidential polls
  - Low response rate: less than 25% of questionnaires returned
  - Undercoverage of poorer demographics: sample of voters relied heavily on lists of automobile and telephone owners, which were generally more affluent voters
- Well, at least we learned from those mistakes, right?

January 24, 2007

Stat 111 - Lecture 3 - Sampling

13

## Recent Presidential Elections

- Using exit polls, several networks report early that Gore wins Florida on 2000 election
- Using exit polls, several pundits predict Kerry will win Ohio in 2004 election
- In general, we have gotten better, but still can make mistakes (especially when difference itself is so small)

January 24, 2007

Stat 111 - Lecture 3 - Sampling

14

## More Potential Problems with Surveys

- **Response Bias:** respondents may not answer truthfully to survey questions
  - Illegal or unpopular behaviour such as drug usage
  - Controversial topics such as teen sexual activity
  - Race or gender of interviewer can influence answers about race or gender-related questions
  - Respondents often have trouble remembering past events eg. yearly nutrition and health surveys

January 24, 2007

Stat 111 - Lecture 3 - Sampling

15

## More Potential Problems with Surveys

- **Wording of questions** can be confusing or intentionally lead the respondent
  - Do you favor a ban on disposable diapers ?
  - It is estimated that disposable diapers account for less than 2% of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard wastes account for 21% of the trash in landfills. Given this, would it be fair to ban disposable diapers?
- Complicated multi-part forms that require lots of skipped questions lead to a drop of in response

January 24, 2007

Stat 111 - Lecture 3 - Sampling

16

## More Complicated Random Surveys

- Weakness of simple random sampling is that you cannot use extra information about population (similar to blocking in experiments)
  - small group of people in one part of the city are known to be wealthier
- **Stratified random sampling:** individuals are divided into groups called strata
  - Simple random sampling done within each stratum
- National surveys can be even more complicated by using **multistage sampling** (cheaper)

January 24, 2007

Stat 111 - Lecture 3 - Sampling

17

## Dinner and Drugs Study

- Study by CASA that linked frequent family dining to reduced risk of substance abuse

“There is no more important thing that a parent can do”
- Some problems with study that relate to what we know about **surveys** and **observational studies**
- Problem 1: **undercoverage** of minority groups
  - Survey not representative of teen population

January 24, 2007

Stat 111 - Lecture 3 - Sampling

18

### Dinner and Drugs Study II

- Problem 2: high level of **non-response** in survey
  - Many households declined to answer, didn't complete survey or denied permission to use
- Problem 3: **observational study** with lots of potential confounding variables
  - Drug use itself wasn't measured, but rather a risk score for drug use
  - Study isn't adjusted for age, which is also associated with drug use

**“Most statisticians would be really cautious interpreting a study such as this one”**

January 24, 2007

Stat 111 - Lecture 3 - Sampling

19

### Next Class - Lecture 4

- Exploring Data: Graphical summaries of a single variable
- Moore and McCabe: Section 1.1

January 24, 2007

Stat 111 - Lecture 3 - Sampling

20