STA 570 Spring 2009

Lecture 2 *Tuesday, Jan 18*

- Sampling and Measurement
  - Scales of Measurement
  - Methods of Collecting Data
  - Simple Random Sampling
  - Other Sampling Plans
  - Sampling and Nonsampling Error

Homework 1: Due this week in the lab.

Homework 2: Due next week in lab.

Review: Basic Definitions

- **Population**

  total set of all subjects of interest
- **Sample**

  subset on which the study collects data
- **Parameter**

  Parameters are calculated using the whole population
- **Statistic**

  Statistics are based on a sample

  (even if they describe a population)

Recall Example from Lecture 1

- The Current Population Survey of about 60,000 households in the United States indicated that 5.3% of married couple households in the United States have annual income below the poverty level.
- Is this number a statistic or a parameter?

Modified Example

- A census of all households in Lexington indicated that 6.2% of married couple households in Lexington have annual income below the poverty level.

- Is this number a statistic or a parameter?

Review: Scales of Measurement

Important to distinguish between these different types of data, because they require different statistical methods.

Nominal - Ordinal - Interval

Qualitative (Categorical) — Quantitative

Lowest level

Highest Level
- most information
- “best” statistical methods

Discrete and Continuous

- If a variable can take only a finite number of values, it is discrete
- Examples: number of children, number of teeth, etc
- Qualitative (categorical) variables are *always* discrete

Discrete and **Continuous**

- Continuous variables can (in theory) take an *infinite continuum* of possible real number values
- Example: time spent on STA 570 homework
  - can be 63 min. or 85 min.
    or 27.358 min. or 27.35769 min. or ...
  - can be **subdivided**
  - therefore **continuous**

Discrete or Continuous

- Quantitative variables can be discrete or continuous
- How about age, income, height?
- **It depends** on the scale
- Age is potentially continuous, but usually measured in years (discrete)

- NB: The distinction between discrete and continuous is not as important as the one between nominal/ordinal/interval

Methods of Collecting Data I
Observational Study

- An observational study observes individuals and measures variables of interest but does not attempt to influence the responses.
- The purpose of an observational study is to describe/compare groups or situations.
- Example: Select a sample of men and women and ask whether he/she has taken aspirin regularly over the past 2 years, and whether he/she had suffered a heart attack over the same period
- *Important Example: Sample surveys/Polls*

Methods of Collecting Data II
Experiment

- An experiment deliberately imposes some treatment on individuals in order to observe their responses.
- The purpose of an experiment is to study whether the treatment causes a change in the response.
- Example: Randomly select men and women, divide the sample into two groups. One group would take aspirin daily, the other would not. After 2 years, determine for each group the proportion of people who had suffered a heart attack.

Methods of Collecting Data III Observational Study/Experiment

- **Observational Studies** are passive data collection
- We observe, record, or measure, but don't interfere
- **Experiments** are active data production
- Experiments actively intervene by imposing some treatment in order to see what happens
- *Experiments are preferable if they are possible*
- Examples

Sampling Plans

- **Simple Random Sampling (SRS)**
- Stratified Random Sampling
- Cluster Sampling
- Systematic Sampling

Simple Random Sampling

- Each possible sample has the same probability of being selected.
- The sample size is usually denoted by *n*.
- Example: Population of 4 students: Adam, Bob, Christina, Dana
  - Select a simple random sample (SRS) of size n=2 to ask them about their smoking habits
  - 6 possible samples of size n=2:

    (1) A+B, (2) A+C, (3) A+D

    (4) B+C, (5) B+D, (6) C+D

How to choose a SRS?

- Each of the six possible samples has to have the same probability of being selected
- For example, roll a die (or use a computer-generated random number) and choose the respective sample
- Online Sampling Applet

How **not** to choose a SRS?

- Ask Adam and Dana because they are in your office anyway
  - "convenience sample"
- Ask who wants to take part in the survey and take the first two who volunteer
  - "volunteer sampling"

Problems with Volunteer Samples

- The sample will poorly represent the population
- Misleading conclusions
- BIAS
- Examples: Mall interview, Street corner interview, Kernel online poll

Famous Example

- 1936 presidential election
- Alfred Landon vs. Franklin Roosevelt
- Literary Digest sent over 10 million questionnaires in the mail to predict the election outcome
- More than 2 million questionnaires returned
- Literary Digest predicted a landslide victory by Alfred Landon

- George Gallup used a much smaller **random** sample
- Gallup predicted Digest would predict 44%
- Digest predicted 43%

- Gallup predicted 56% for the election
- Actual results were 62% for Roosevelt

- Why was the Literary Digest prediction so far off?

Other Examples

- TV, radio call-in polls
- "If you had to do it over again, would you have children?"
- Ann Landers had 10,000 responses and about 70% said "NO!"

Compare to SRS

- Newsday commissioned a nationwide SRS poll of n = 1373 parents
- Found that 91% would have children again!
- Which figure is correct?

Bias

- Follow-up by Landers
  - "People who are contented are rarely motivated to write and tell me how happy they are. Anger, hostility and resentment are often the fuel that moves people to action."

$\longrightarrow$ Bias in voluntary response "surveys".

Other Examples

- TV, radio call-in polls
- “should the UN headquarters continue to be located in the US?”
- ABC poll with 186,000 callers: 67% no
- Scientific random sample with 500 respondents: 28% no
- The smaller **random** sample is much more trustworthy because it has less bias

- Cool inferential statistical methods can be applied to state that “the true percentage of all Americans who want the UN headquarters out of the US is between 24% and 32%”
- These methods **can not** be applied to a volunteer sample.

Why are call-in polls usually biased?

- People are much more likely to call in if the feel strongly about an issue

  (e.g., Israel-Palestine-Lebanon, war in Iraq, health care, equal rights for homosexuals, pedestrian safety, development of downtown block, name of the UK mascot)

The UK mascot

- Wildcat named “Blue” is the official UK mascot

  (SEC info)
- The name was selected 2002 in an **online poll** where multiple voting was possible
- The choices were “Champ”, “Blue”, or “Tucky”
- Somebody felt strongly about it and voted often

Question Wording

- Kalton et al. (1978), England
- Two groups get questions with slightly different wording
- Group 1 is asked: “Are you in favor of giving special priority to buses in the rush hour *or not*?”
- Group 2 is asked: “Are you in favor of giving special priority to buses in the rush hour *or should cars have just as much priority as buses*?”

Question Wording

- Result: Proportion of people saying that priority should be given to buses.

	Without reference to cars	With reference to cars	Difference
All respondents	**0.69** (n=1076)	**0.55** (n=1081)	**0.14**
Women	**0.65** (n=585)	**0.49** (n=590)	**0.16**
Men	**0.74** (n=491)	**0.66** (n=488)	**0.08**
Non Car-owners	**0.73** (n=565)	**0.55** (n=554)	**0.18**
Car owners	**0.66** (n=509)	**0.54** (n=522)	**0.12**

Question Wording

- 2000 Republican primary; Bush camp asked voters in South Carolina
  - "Would you be more likely or less likely to vote for John McCain for president if you knew he had fathered an illegitimate black child?"
- Push Polls “push” voters in a certain direction.
- Inference with push poll becomes nonsense.

Question Order

- Two questions asked in different order during the cold war
- (1) “Do you think the U.S. should let Russian newspaper reporters come here and send back whatever they want?” 36% answered “Yes”
- (2) “Do you think Russia should let American newspaper reporters come in and send back whatever they want?”
- When question (2) was asked first, 73% answered “Yes” to question (1)

Don’t trust bad samples

- Whenever you see results from a poll, check
  - who sponsored the poll,
  - who conducted the poll,
  - how the questions were worded,
  - how the sample was selected (e.g., whether it was a random sample),
  - and how large the sample was
- ***If you can not find this information, the results may not be trustworthy***

Sampling Plans

- Simple Random Sampling (SRS)
- **Stratified Random Sampling**
- **Cluster Sampling**
- Systematic Sampling

Stratified Sampling

- Suppose the population can be divided into separate, non-overlapping groups ("***strata***") according to some criterion.
- Select a simple random sample independently from each group.

Why could stratification be useful?

- We may want to draw inference about population parameters for each subgroup
- Sometimes, ("proportional stratified sample") estimators from stratified random samples are more precise than those from simple random samples

Proportional Stratification

- The proportions of the different strata are the same in the sample as in the population
- Mathematically:

Population size N, subpopulation sizes N_i

Sample size n, subsample sizes n_i

$$\frac{n_i}{n} = \frac{N_i}{N}$$

Proportional Stratification

- Example:
  - Total population of the US: 304 Million
  - Population of Kentucky: 4 Million (1.3%)
  - Suppose you take a sample of size n=304 of people living in the US.
  - If stratification is proportional, then 4 people in the sample need to be from Kentucky
  - Suppose you take a sample of size n=1000. If you want it to be proportional, then 13 people (1.3%) need to be from Kentucky.

Cluster Sampling

- The population can be divided into a set of non-overlapping subgroups (the clusters)
- The clusters are then selected at random, and all individuals in the selected clusters are included in the sample
- Example: to conduct personal interviews of operating room nurses, it might make sense to randomly select a sample of hospitals and then interview all of the operating room nurses at that hospital.

Summary of Important Sampling Plans

- **Simple Random Sampling (SRS)**
  - Each possible sample has the same probability of being selected.
- **Stratified Random Sampling**
  - Non-overlapping subgroups (strata)
  - SRSs are drawn from each strata
- **Cluster Sampling**
  - Non-overlapping subgroups (clusters)
  - Clusters selected at random
  - All individuals in the selected clusters are included in the sample
- **Systematic Sampling**
  - Useful when the population consists as a list
  - A value *K* is specified. Then one of the first *K* individuals is selected at random, after which every *Kth* observation is included in the sample

Types of Bias

- **Selection Bias**
  - Selection of the sample systematically excludes some part of the population of interest
- **Measurement/Response Bias**
  - Method of observation tends to produce values that systematically differ from the true value
- **Nonresponse Bias**
  - Occurs when responses are not actually obtained from all individuals selected for inclusion in the sample

Biased or Unbiased Sample?

- Researchers state, “A systematic sample was drawn from the national membership lists of the American Society of Certified Accountants, National Association of Accountants, American Association of Women Accountants, and the Association of Government Accountants. An initial name on each list was selected at random, and every *K*th name was thereafter selected. *K* was computed by dividing membership list length by the desired sample size.”

Day, Bedeian (1991), Journal of Vocational Behavior, 38, 39-50

Bias?

- Pittsburgh is known to have a very good medical center. However, in “America’s Most Liveable cities”, Pittsburgh was marked down on health care.
- The variable used as a proxy for healthcare was “mortality rate in hospitals”.
- Why would a good medical center perform poorly on mortality rate?

Sampling and Nonsampling Error

- Assume you take a random sample of 100 UK students and ask them about their political affiliation (Democrat, Republican, Independent)
- Now take another random sample of 100 UK students
- Will you get the same percentages?

Sampling Error

- No, because of sampling variability.
- Also, the result will not be exactly the same as the population percentage, unless you take a “sample” consisting of the whole population of 30,000 students (this would be called a “census”)

or if you are very lucky

Sampling Error

- **Sampling Error** is the error that occurs when a statistic based on a sample estimates or predicts the value of a population parameter.
- In random samples, the sampling error can usually be quantified.
- In ***nonrandom*** samples, there is also sampling variability, but its extent is ***not predictable***.

Nonsampling Error

- Everything that could also happen in a census, that is when you ask the whole population
- Examples: Bias due to question wording, question order, nonresponse (people refuse to answer), wrong answers (especially to delicate questions)

Quiz

- Take a piece of paper and write your ***name*** and ***section number*** on top of it
- Please write your answers to the questions legibly…
- When you are done,
  - quietly leave your seat,
  - turn in the paper,
  - and quietly leave the room