STA 291
Lecture 3

- Data type:
  - Categorical/Qualitative and
  - Quantitative/Numerical

    within categorical (nominal and ordinal)

    within quantitative (discrete and continuous)

- How data are collected?

  experiments and surveys (polls)

Example of experiment: clinical trials testing the effectiveness of a new drug.

Example of survey: opinion polls.

In both methods, a key ingredient is randomness. "randomly select people to interview" (in survey)

"randomly divide patients into two groups" (in experiment)

Observational Study = Survey

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 age 18 and over and ask whether he/she smoke cigarette.

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: Volunteers, divided randomly into two groups. One group would take aspirin daily, the other would not. After 3 years, determine for each group the proportion of people who had suffered a heart attack. (This is an actual study)

Methods of Collecting Data 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 can tell what caused the change, if any.*

- Why random?
- To eliminate bias.

Collecting data for a poll

Simple Random Sampling

- Each possible sample has the same probability of being selected.
- The sample size is usually denoted by *n*.

Example: Simple Random Sampling

- 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

Q: How to choose a SRS?
A: "Label and table"

- Give each unit in the population a unique label (usually a number, like SSN, SID, or phone number etc) (product serial #)

- Go to random number table to see which label (unit) should be selected as sample. [this step often done by computer now]

Q: How to choose a SRS?
A: “Label and table”

- 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 random number Applet acts like a table

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 – and no way to pin it down (how much is the bias?)
- Examples: Mall interview, Street corner interview, internet click survey, TV show audience phone-in the opinion.

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 and predicted a clear victory by Franklin Roosevelt (modern technique were able to reduce the sample size *n* to 1500 or so)

- Roosevelt won with 62% of the vote

- Why was the Literary Digest prediction so far off?

Terminology

- Population
- Sample

- Parameter
- Estimator

Other Examples

- TV talk show, 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% etc.”
- These methods **cannot** be applied to a volunteer sample.

- http://www.pollster.com/pollster-faq/
- http://abcnews.go.com/PollingUnit/
- http://en.wikipedia.org/wiki/Clinical_trial

Collecting Data II --- Experiments

- **Example: testing of new treatments or drugs** via clinical trials.
- Testing a new product, etc.

- Clinical trial: Double blinded, placebo controlled, randomized.

- recruit volunteers that met specific requirements (have certain conditions). Statistician will decide how many subjects is enough. (usually from a few hundreds to a few thousands, depending on what you are looking for, what is the budget, how certain the result need be ….)

- Randomly decide if a volunteer is given the new drug or ***placebo*** (sugar pill). Usually 50%-50% chance.

- Neither the subject nor the attending doctor know which is given to the subject. (to minimize psychological effects, also called placebo effects)
- Only a high level committee know.

- The idea is to match as closely as possible the subjects of the two groups. The only difference is the drug.

- The phrase “if *everything else remain the same*, the use of the drug for XXX patients can reduce the 5 year mortality rate by X%”

- Resulting data are analyzed by statistical procedure. (will cover later)
- Conclusion might be “proven beyond reasonable doubt that the new drug is better”. Or …
- Inconclusive…either no effect or the results too noisy (effect too small) that you do not see it clearly, or
- Clearly No effect.

- More than 40% of clinical trials result in abandoning of the drug. (either because of no good effect, or bad side effects) Very costly. (Hundreds of millions $)
- Any drug company announcing the abandoning of a (phase III) clinical trial usually result in their stock price going down significantly.
- Vioxx, phen-fen, ……Purdue Pharma to Withdraw Palladone ….

- Success story:

- Martha Stewart went to jail because of selling a drug company stock with inside information and then lied about it.

- Info: ImClone's new drug (for cancer) was not ***statistically proven*** to be effective, Food and Drug Administration determined. So the stock price fall.

- Variations/refinements of SRS:

- Stratified sampling
- Cluster and multistage sampling
- Systematic sampling

Attendance Survey Question 2

- On a 4”x6” index card (or little piece of paper)
  - Name and section number
  - Today’s Question: The two ways of collecting data we covered today are

(1) Surveys, and

(2) _____ (one word, begin with E).