STA 291

Lecture 4 Jan 26, 2010

- **Methods of Collecting Data**

- **Survey**
- **Experiment**

Review: Methods of Collecting Data Observational Study vs. Experiment

- An observational study (survey) passively observes individuals and measures variables of interest but does not attempt to influence the responses

- An experiment deliberately imposes actively some treatment on individuals in order to observe their responses

- Population ←→ Parameter

- Sample ←→ Statistics (=estimator)

Sample size n

- Interview how many people?

"PARDNER, A BIG STATE NEEDS A BIG SAMPLE."

STATISTICS

- Required sample size changes very little regarding the population size. (say from 30 K to 300 Mil etc.)

- Very much depend on the required margin of error.

- Typical example: sample size n=1500, margin of error = 2.6 % (assume using SRS)

- Big N the size of population
- Small n the size of sample

- Unless N is very small (comparable to n) the reliability of the survey results depend minimally on N

Collecting Data II --- Experiments

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

- Clinical trials (3 Key features):

*R*andomized,

*P*lacebo controlled,

*D*ouble blinded.

- recruit volunteers that met specific requirements (have certain conditions). Statistician decide how many subjects are enough. (usually 100 to a few 1000, depending on what you are looking for, what is the budget, how certain the result need be ….)

- Randomly decide if a subject is given the new drug or placebo (sugar pill). Usually 50%-50% chance. [randomized]

- 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. [double blind]

- The two groups are usually called Treatment group and Control group or drug group and placebo group.

- The need of the control group, in a comparison.

- 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 this drug for XXX patients can reduce the 5 year mortality rate by X%” etc. [or “reduce the risk of heart attack by x%” etc.]

- 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 that you do not see it clearly, or
- Clearly No effect.

- What is/are the population(s) here?

- What is/are the population(s) here?

  --- there are two:

  those patients that treated with drug is usually called the treatment population

  those receive placebo usually called control population

- How many samples here? Two.

  treatment sample

  placebo sample or control sample.

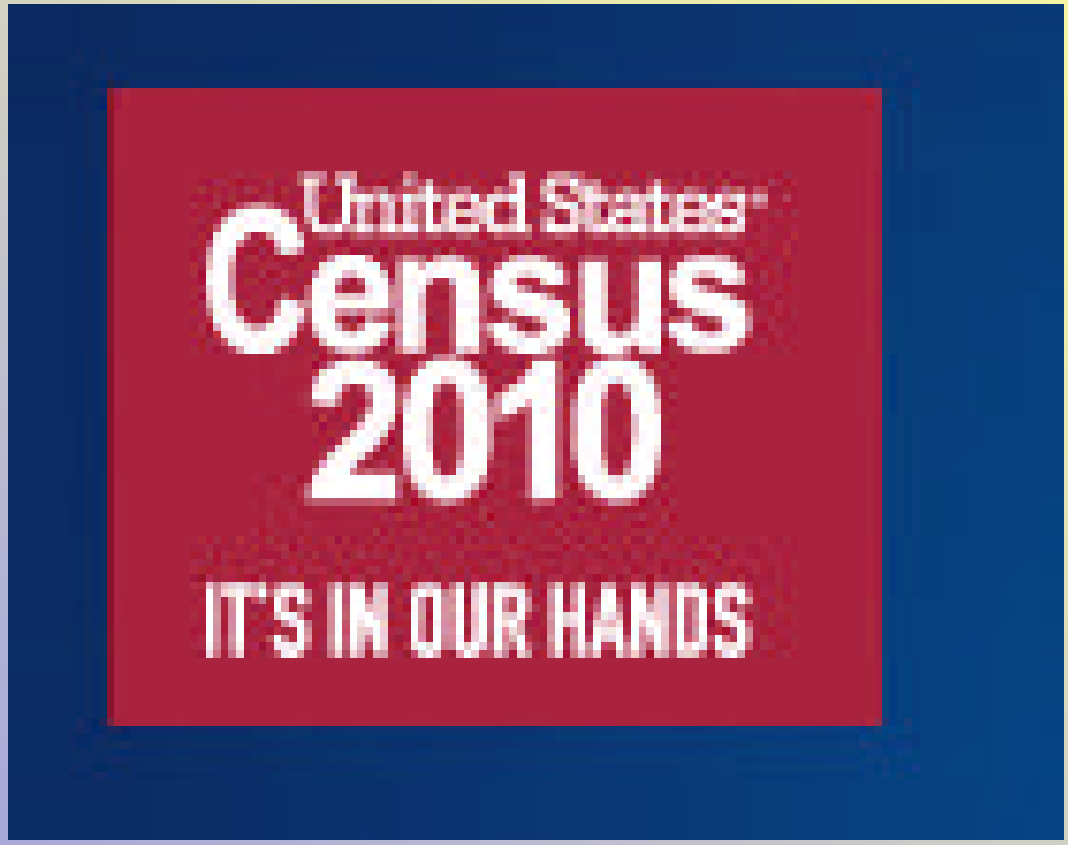
- More than 40% of clinical trials result in abandon of the drug. Very costly. (Hundreds of millions $)
- Any drug company announcing the abandoning of a (phase III) clinical trial usually result in their stock price going down significantly.
- Vioxx, phen-fen, …….Purdue Pharma to Withdraw Palladone ….

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

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

Why sample?

- Why not just measure all?



Question Wording

- Kalton et al. (1978), England
- Two groups get questions with slightly different wording

Question 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 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)

Stratified Sampling

- Suppose the population can be divided into separate, non-overlapping groups ("***strata***") according to some criterion.

  example: all voters in USA can be divided into Male voters, female voters.

- 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: 300 Million (2006)
  - Population of Kentucky: 4 Million (1.33%)
  - Suppose you take a sample of size n=300 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 (1.33%) need to be from Kentucky.

Summary: Important Sampling Plans

- **Stratified Random Sampling**
  - The population can be divided into a set of non-overlapping subgroups (the strata or sub-populations)
  - SRSs are drawn from each strata
- **Cluster and multistage Sampling**

- **Systematic Sampling**
  - A value *K* is specified. Then Randomly select a starting point, after which every *Kth* observation is included in the sample

Systematic sampling

- Digital music. MP3……sampling rate
- CD quality music Typically sample 44,100 times per second

- SRS has no bias.
- Stratified sampling, if done right, can also be no bias.
- But SRS is hard.

Where Does Bias Occur?

- **Selection Bias**
  - Selection of the sample systematically excludes some part of the population of interest

- **Nonresponse Bias**
  - Occurs when responses are not actually obtained from all individuals selected for inclusion in the sample

Biased or Unbiased Sample?

- Researchers state, “This study was conducted at a large, predominantly White southwestern university. On this campus, American Indians were the smallest racial and ethnic minority student group, consisting of only 2.3% of the student population. Recruited through education and liberal arts classes, students who volunteered to participate in this study completed the research packet and returned it during the next class period. A total of 83 American Indian undergraduates returned completed survey packets.”

Gloria, Kurpius (2001), Cultural Diversity and Ethnic Minority Psychology, 7, 88-102

Attendance Survey Question 4

- On a 4”x6” index card (or little piece of paper)
  - Please write down your name
  - Today’s Question:

    What is sampling scheme used in the Digital music? _____ sampling.

    [start with S.]

Next Definition: Sampling 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?

- 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, 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.