Chapter 8

Producing Data: Sampling

Population and Sample

- Researchers often want to answer questions about some large group of individuals (this group is called the **population**)
- Often the researchers cannot measure (or survey) all individuals in the population, so they measure a subset of individuals that is chosen to represent the entire population (this subset is called a **sample**)
- The researchers then use *statistical techniques* to make conclusions about the population based on the sample

Bad Sampling Designs

- Voluntary response sampling
  - allowing individuals to choose to be in the sample
- Convenience sampling
  - selecting individuals that are easiest to reach
- Both of these techniques are biased
  - systematically favor certain outcomes

Voluntary Response

- Advice columnist Ann Landers asked her readers, "If you had it to do over again, would you have children?"
- A few weeks later, her column was headlined: "70% OF PARENTS SAY KIDS NOT WORTH IT."
- The people who responded felt strongly enough to take the trouble to write Ann Landers. Their letters showed that many of them were angry at their children.
- These people don't fairly represent all parents.
- A statistically designed opinion poll on the same issue a few months later found that 91% of parents would have children again.

Convenience Sampling

- Sampling mice from a large cage to study how a drug affects physical activity
  - lab assistant reaches into the cage to select the mice one at a time until 10 are chosen
- Which mice will likely be chosen?
  - could this sample yield biased results?

Simple Random Sampling

- Each individual in the population has the same chance of being chosen for the sample
- Each group of individuals (in the population) of the required size (*n*) has the same chance of being the sample actually selected
- Random selection:
  - "drawing names out of a hat"
  - table of random digits
  - computer software

Table of Random Digits

- Table B on pg. 692 of text
  - each entry is equally likely to be any of the 10 digits 0 through 9
  - entries are *independent* of each other (knowledge of one entry gives no information about any other entries)
  - each pair of entries is equally likely to be any of the 100 pairs 00, 01,…, 99
  - each triple of entries is equally likely to be any of the 1000 values 000, 001, …, 999

Choosing a Simple Random Sample (SRS)

STEP 1: Label each individual in the population

STEP 2: Use Table B to select labels at random

Probability Sample

- a sample chosen by chance
- a SRS gives each member of the population an equal chance to be selected

Stratified Random Sample

- first divide the population into groups of similar individuals, called **strata**
- second, choose a separate SRS in each stratum
- third, combine these SRSs to form the full sample

Stratified Random Sample Example

Suppose a university has the following student demographics:

Undergraduate	Graduate	First Professional	Special
55%	20%	5%	20%

A stratified random sample of 100 students could be chosen as follows: select a SRS of 55 undergraduates, a SRS of 20 graduates, a SRS of 5 first professional students, and a SRS of 20 special students; combine these 100 students.

Multistage Sample

- several stages of sampling are carried out
- useful for large-scale sample surveys
- samples at each stage may be SRSs, but are often stratified
- stages may involve other random sampling techniques as well (cluster, systematic, random digit dialing, …)

Cautions about Sample Surveys

- Undercoverage
  - some individuals or groups in the population are left out of the process of choosing the sample
- Nonresponse
  - individuals chosen for the sample cannot be contacted or refuse to cooperate/respond
- Response bias
  - behavior of respondent or interviewer may lead to inaccurate answers or measurements
- Wording of questions
  - confusing or leading (biased) questions; words with different meanings

Nonresponse

- Advice columnist Ann Landers asked her readers, "If you had it to do over again, would you have children?"
- A few weeks later, her column was headlined: "70% OF PARENTS SAY KIDS NOT WORTH IT."
- The people who responded felt strongly enough to take the trouble to write Ann Landers. Their letters showed that many of them were angry at their children.
- These people don't fairly represent all parents.
- A statistically designed opinion poll on the same issue a few months later found that 91% of parents would have children again.

Response Bias

- A door-to-door survey is being conducted to determine drug use (past or present) of members of the community. Respondents may give *socially acceptable answers* (maybe not the truth!)
- For this survey on drug use, would it matter if a police officer is conducting the interview? (*bias from interviewer*)

Response Bias

Asking the Uninformed

Washington Post National Weekly Edition (April 10-16, 1995, p. 36)

- A 1978 poll done in Cincinnati asked people whether they "favored or opposed repealing the 1975 Public Affairs Act."
  - There was no such act!
  - About one third of those asked expressed an opinion about it.

Wording of Questions

A newsletter distributed by a politician to his constituents gave the results of a "nationwide survey on Americans' attitudes about a variety of educational issues." One of the questions asked was, "Should your legislature adopt a policy to assist children in failing schools to opt out of that school and attend an alternative school--public, private, or parochial--of the parents' choosing?" From the wording of this question, can you speculate on what answer was desired? Explain.

Wording: Deliberate Bias

- "If you found a wallet with $20 in it, would you return the money?"
- "If you found a wallet with $20 in it, would you do the right thing and return the money?"

Wording: Unintentional Bias

- "I have taught several students over the past few years."
  - How many students do you think I have taught?
  - How many years am I referring to?
- "Over the past few days, how many servings of fruit have you eaten?"
  - How many days are you considering?
  - What constitutes a serving?

Wording: Unnecessary Complexity

- "Do you sometimes find that you have arguments with your family members and co-workers?"
  - Arguments with family members
  - Arguments with co-workers

Wording: Ordering of Questions

- "How often do you normally go out on a date? about ___ times a month."
- "How happy are you with life in general?"
  - Strong association between these questions.
  - If the ordering is reversed, then there would be no strong association between these questions

Inferences about the Population

- Values calculated from samples are used to make conclusions (*inferences*) about unknown values in the population
- Variability
  - different samples from the same population may yield different results for a particular value of interest
  - estimates from random samples will be closer to the true values in the population if the samples are larger
  - how close the estimates will likely be to the true values can be calculated -- this is called the margin of error