TOPIC 6

Overview of Methods of Data Collection

- Census
- Sample Survey
- Experiment
- Observational Study

In the real world, time and cost considerations usually make it impossible to analyze an entire population. Does the government question you and your parents before announcing the monthly unemployment rates? Does a television producer check every household's viewing preferences before deciding whether a pilot program will be continued? In studying statistics we learn how to estimate *population* characteristics by considering a *sample.* For example, later in this book we will see how to estimate population means and proportions by looking at sample means and proportions.

To derive conclusions about the larger population, we need to be confident that the sample we have chosen represents that population fairly. Analyzing the data with computers is often easier than gathering the data, but the frequently quoted "Garbage in, garbage out" applies here. Nothing can help if the data are badly collected. Unfortunately, many of the statistics with which we are bombarded by newspapers, radio, and television are based on poorly designed data collection procedures.

CENSUS

A *census* is a complete enumeration of an entire population. In common use, it is often thought of as an official attempt to contact every member of the population, usually with details regarding age, marital status, race, gender, occupation, income, years of school completed, and so on. Every 10 years the U.S. Bureau of the Census divides the nation into nine regions and attempts to gather information about everyone in the country. A massive amount of data is obtained, but even with the resources of the U.S. government, the census is not complete. For example, many homeless people are always missed, or counted at two temporary residences, and there are always households that do not respond even after repeated requests for information. It was estimated that the 2000 census missed about 3.3 million people (1.2% of the population).

In most studies, both in the private and public sectors, a complete census is unreasonable because of time and cost involved. Furthermore, attempts to gather complete data have been known to lead to carelessness. Finally, and most important, a well-designed, well-conducted sample survey is far superior to a poorly designed study involving a complete census. For example, a poorly worded question might give meaningless data even if everyone in the population answers.

SAMPLE SURVEY

The census tries to count everyone; it is not a sample. A sample survey aims to obtain information about a whole population by studying a part of it, that is, a sample. The goal is to gather information without disturbing or changing the population. Numerous procedures are used to collect data through sampling, and much of the statistical information distributed to us comes from sample surveys. Often, controlled experiments are later undertaken to demonstrate relationships suggested by sample surveys.

However, the one thing that most quickly invalidates a sample and makes useful information impossible to obtain is *bias.* A sample is biased if in some critical way it does not represent the population. The main technique to avoid bias is to incorporate *randomness* into the selection process. Randomization protects us from effects and influences, both known and unknown. Finally, the larger the sample, the better the results, but what is critical is the sample size, not the percentage or fraction of the population. That is, a random sample of size 500 from a population of size 100,000 is just as representative as a random sample of size 500 from a population of size 1,000,000.

EXPERIMENT

In a controlled study, called an *experiment,* the researcher should randomly divide subjects into appropriate groups. Some action is taken on one or more of the groups, and the response is observed. For example, patients may be randomly given unmarked capsules of either aspirin or acetaminophen and the effects of the medication measured. Experiments often have a *treatment group* and a *control group*; in the ideal situation, neither the subjects nor the researcher knows which group is which. The Salk vaccine experiment of the 1950s, in which half the children received the vaccine and half were given a placebo, with not even their doctors knowing who received what, is a classic example of this *double-blind* approach. Controlled experiments can indicate cause-and-effect relationships.

The critical principles behind good experimental design include *control* (outside of who receives what treatments, conditions should be as similar as possible for all involved groups), *blocking* (the subjects can be divided into representative groups to bring certain differences directly into the picture), *randomization* (unknown and uncontrollable differences are handled by randomizing who receives what treatments), *replication* (treatments need to be repeated on a sufficient number of subjects), and *generalizability* (ability to repeat an experiment in a variety of settings).

OBSERVATIONAL STUDY

Sample surveys are one example of what are called *observational studies.* In observational studies there is no choice in regard to who goes into the treatment and control

groups. For example, a researcher cannot ethically tell 100 people to smoke three packs of cigarettes a day and 100 others to smoke only one pack per day; he can only observe people who habitually smoke these amounts. In observational studies the researcher strives to determine which variables affect the noted response. While results may suggest relationships, it is difficult to conclude cause and effect.

Observational studies are primary, vital sources of data; however, they are a poor method of measuring the effect of change. To evaluate responses to change, one must impose change, that is, perform an experiment. Furthermore, observational studies on the impact of some variable on another variable often fail because explanatory variables are *confounded* with other variables.

Summary

- A complete census is usually unreasonable because of time and cost constraints.
- Estimate population characteristics (called parameters) by considering statistics from a sample.
- Analysis of badly gathered sample data is usually a meaningless exercise.
- A sample is biased if in some critical way it does not represent the population.
- The main technique to avoid bias is to incorporate randomness into the selection process.
- Experiments involve applying a treatment to one or more groups and observing the responses.
- Observational studies involve observing responses to choices people make.

TOPIC 7

Planning and Conducting Surveys

- Simple Random Sampling
- Characteristics of a Well-Designed, Well-Conducted Survey
- Sampling Error
- Sources of Bias
- Other Sampling Methods

Most data collection involves observational studies, not controlled experiments. Furthermore, while most data collection has some purpose, many studies come to mind after the data have been assembled and examined. For data collection to be useful, the resulting sample must be representative of the population under consideration.

SIMPLE RANDOM SAMPLING

How can a good, that is, a representative, sample be chosen? The most accurate technique would be to write the name of each member of the population on a card, mix the cards thoroughly in a large box, and pull out a specified number of cards. This method would give everyone in the population an equal chance of being selected as part of the sample. Unfortunately, this method is usually too time-consuming and too costly, and bias might still creep in if the mixing is not thorough. A *simple random sample*, that is, **one in which every possible sample of the desired size has an equal chance of being selected**, can more easily be obtained by assigning a number to everyone in the population and using a random number table or having a computer generate random numbers to indicate choices.

EXAMPLE 7.1

Suppose 80 students are taking an AP Statistics course and the teacher wants to randomly pick out a sample of 10 students to try out a practice exam. She first assigns the students numbers 01, 02, 03, . . . , 80. Reading off two digits at a time from a random number table, she ignores any over 80 and ignores repeats, stopping when she has a set of ten. If the table began 75425 56573 90420 48642 27537 61036 15074 84675, she would choose the students numbered 75, 42, 55, 65, 73, 04, 27, 53, 76, and 10. Note that 90 and 86 are ignored because they are over 80, and the second and third occurrences of 42 are ignored because they are repeats.

CHARACTERISTICS OF A WELL-DESIGNED, WELL-CONDUCTED SURVEY

A well-designed survey always incorporates chance, such as using random numbers from a table or a computer. However, the use of probability techniques is not enough to ensure a representative sample. Often we don't have a complete listing of the population, and so we have to be careful about exactly how we are applying "chance." Even when subjects are picked by chance, they may choose not to respond to the survey or they may not be available to respond, thus calling into question how representative the final sample really is. The wording of the questions must be neutral—subjects give different answers depending on the phrasing.

EXAMPLE 7.2

Suppose we are interested in determining the percentage of adults in a small town who eat a nutritious breakfast. How about randomly selecting 100 numbers out of the telephone book, calling each one, and asking whether the respondent is intelligent enough to eat a nutritious breakfast every morning?

Answer: Random selection is good, but a number of questions should be addressed. For example, are there many people in the town without telephones or with unlisted numbers? How will the time of day the calls are made affect whether the selected people are reachable? If people are unreachable, will replacements be randomly chosen in the same way or will this lead to a certain class of people being underrepresented? Finally, even if these issues are satisfactorily addressed, the wording of the question is clearly not neutral—unless the phrase *intelligent enough* is dropped, answers will be almost meaningless.

SAMPLING ERROR: THE VARIATION INHERENT IN A SURVEY

No matter how well-designed and well-conducted a survey is, it still gives a sample *statistic* as an estimate for a population *parameter*. Different samples give different sample statistics, all of which are estimates for the same population parameter, and so error, called *sampling error*, is naturally present. This error can be described using probability; that is, we can say how likely we are to have a certain size error. Generally, the chance of this error occurring is smaller when the sample size is larger. However, the way the data are obtained is crucial—a large sample size cannot make up for a poor survey design or faulty collection techniques.

EXAMPLE 7.3

Each of four major news organizations surveys likely voters and separately reports that the percentage favoring the incumbent candidate is 53.4%, 54.1%, 52.0%, and 54.2%, respectively. What is the correct percentage? Did three or more of the news organizations make a mistake?

Answer: There is no way of knowing the correct population percentage from the information given. The four surveys led to four statistics, each an estimate of the population parameter. No one made a mistake unless there was a bad survey, for example, one without the use of chance, or not representative of the population, or with poor wording of the question. Sampling differences are natural.

SOURCES OF BIAS IN SURVEYS

Poorly designed sampling techniques result in *bias,* that is, in a tendency to favor the selection of certain members of a population. If a study is biased, size doesn't help—a large sample size will simply result in a large worthless study. Think about bias before running a study, because once all the data comes in, there is no way to recover if the sample was biased. Sometimes pilot testing with a small sample will show bias that can be corrected before a larger sample is obtained. Although each of the following sources of bias is defined separately, there is overlap, and many if not most examples of bias involve more than one of the following.

TIP

Sampling error is to be expected, while bias is to be avoided.

Household bias: When a sample includes only one member of any given household, members of large households are underrepresented. To respond to this, pollsters sometimes give greater weight to members of larger households.

Nonresponse bias: A good example is that of most mailed questionnaires, as they tend to have very low response percentages, and it is often unclear which part of the population is responding. Sometimes people chosen for a survey simply refuse to respond or are unreachable or too difficult to contact. Answering machines and caller ID prevent easy contacts. To maximize response rates, one can use multiple follow-up contacts and cash or other incentives. Also, short, easily understood surveys generally have higher response rates.

Quota sampling bias: This results when interviewers are given free choice in picking people, for example, to obtain a particular percentage men, a particular percentage Catholic, or a particular percentage African-American. This flawed technique resulted in misleading polls leading to the *Chicago Tribune* making an early incorrect call of Thomas E. Dewey as the winner over Harry S. Truman in the 1948 presidential election.

Response bias: The very question itself can lead to misleading results. People often don't want to be perceived as having unpopular or unsavory views and so may respond untruthfully when face to face with an interviewer or when filling out a questionnaire that is not anonymous. Patients may lie about following doctors' orders, dieters may be dishonest about how strictly they've followed a weight loss program, students may shade the truth about how many hours they've studied for exams, and viewers may not want to admit they they watch certain television programs.

TIP

Think about potential bias before collecting data.

Selection bias: An often-cited example is the *Literary Digest* opinion poll that predicted a landslide victory for Alfred Landon over Franklin D. Roosevelt in the 1936 presidential election. The *Digest* surveyed people with cars and telephones, but in 1936 only the wealthy minority, who mainly voted Republican, had cars and telephones. In spite of obtaining more than two million responses, the *Digest* picked a landslide for the wrong man!

Size bias: Throwing darts at a map to decide in which states to sample would bias in favor of geographically large states. Interviewing people checking out of the hospital would bias in favor of patients with short stays, since due to costs, more people today have shorter stays. Having each student pick one coin out of a bag of 1000 coins to help estimate the total monetary value of the coins in the bag would bias in favor of large coins, for example, quarters over dimes.

Undercoverage bias: This happens when there is inadequate representation. For example, telephone surveys simply ignore all those possible subjects who don't have telephones. In the 2008 presidential election surveys, phone surveys went only to land line phones, leaving out many young adults who have only cell phones. Another example is *convenience samples,* like interviews at shopping malls, which are based on choosing individuals who are easy to reach. These interviews tend to produce data highly unrepresentative of the entire population. Door-to-door household surveys typically miss college students and prison inmates, as well as the homeless.

Voluntary response bias: Samples based on individuals who offer to participate typically give too much emphasis to people with strong opinions. For example, radio call-in programs about controversial topics such as gun control, abortion, and school segregation do not produce meaningful data on what proportion of the population favor or oppose related issues. Online surveys posted to websites are a modern source of voluntary response bias.

Wording bias: Nonneutral or poorly worded questions may lead to answers that are very unrepresentative of the population. To avoid such bias, do not use *leading* questions, and write questions that are clear and relatively short. Also be careful of sequences of questions that lead respondents toward certain answers.

Note: Again, it should be understood that there is considerable overlap among the above classifications. For example, a nonneutral question may be said to have both *response* bias and *wording* bias. *Selection* bias and *undercoverage* bias often go hand in hand. *Voluntary response* bias and *nonresponse* bias are clearly related.

OTHER SAMPLING METHODS

Time- and cost-saving modifications are often used to implement sampling procedures other than simple random samples.

Systematic sampling involves listing the population in some order (for example, alphabetically), choosing a random point to start, and then picking every tenth (or hundredth, or thousandth, or *k*th) person from the list. This gives a reasonable sample as long as the original order of the list is not in any way related to the variables under consideration.

TIP

Know the difference between strata and clusters.

In *stratified sampling* the population is divided into *homogeneous* groups called *strata,* and random samples of persons from all strata are chosen. For example, we can stratify by age or gender or income level or race and pick a sample of people from each stratum. Note that all individuals in a given stratum have a characteristic in common. We could further do *proportional sampling,* where the sizes of the random samples from each stratum depend on the proportion of the total population represented by the stratum.

In *cluster sampling* the population is divided into *heterogeneous* groups called *clusters,* and we then take a random sample of clusters from among all the clusters. For example, to survey high school seniors we could randomly pick several senior class homerooms in which to conduct our study. Note that each cluster should resemble the entire population.

Multistage sampling refers to a procedure involving two or more steps, each of which could involve any of the various sampling techniques. The Gallup organization often

follows a procedure in which nationwide locations are randomly selected, then neighborhoods are randomly selected in each of these locations, and finally households are randomly selected in each of these neighborhoods.

EXAMPLE 7.4

Suppose a sample of 100 high school students from a school of size 5000 is to be chosen to determine their views on the death penalty. One method would be to have each student write his or her name on a slip of paper, put the papers in a box, and have the principal reach in and pull out 100 of the papers. However, questions could arise regarding how well the papers are mixed up in the box. For example, how might the outcome be affected if all students in one homeroom toss in their names at the same time so that their papers are clumped together? Another method would be to assign each student a number from 1 to 5000 and then use a random number table, picking out four digits at a time and tossing out repeats and numbers over 5000 (simple random sampling). What are alternative procedures?

Answer: From a list of the students, the surveyor could simply note every fiftieth name (systematic sampling). Since students in each class have certain characteristics in common, the surveyor could use a random selection method to pick 25 students from each of the separate lists of freshmen, sophomores, juniors, and seniors (stratified sampling). The researcher could separate the homerooms by classes; then randomly pick five freshmen homerooms, five sophomore homerooms, five junior homerooms, and five senior homerooms (cluster sampling); and then randomly pick five students from each of the homerooms (multistage sampling). The surveyor could separately pick random samples of males and females (stratified sampling), the size of each of the two samples chosen according to the proportion of male and female students attending the school (proportional sampling).

It should be noted that none of the alternative procedures in the above example result in a *simple random* sample because every possible sample of size 100 does not have an equal chance of being selected.

Summary

- A simple random sample (SRS) is one in which every possible sample of the desired size has an equal chance of being selected.
- Sampling error is not an error, but rather refers to the natural variability between samples.
- Bias is the tendency to favor the selection of certain members of a population.
- Nonresponse bias occurs when a large fraction of those sampled do not respond (most mailed questionnaires are good examples).
- Response bias happens when the question itself leads to misleading results (for example, people don't want to be perceived as having unpopular, unsavory, or illegal views).
- Undercoverage bias occurs when part of the population is ignored (for example, telephone surveys miss all those without phones).

- Voluntary response bias occurs when individuals choose whether to respond (for example, radio call-in surveys).
- Systematic sampling involves listing the population, choosing a random point to start, and then picking every nth person for some n.
- Stratified sampling involves dividing the population into homogeneous groups called strata and then picking random samples from each of the strata.
- Cluster sampling involves dividing the population into heterogeneous groups called clusters, and then picking everyone in a random sample of the clusters.
- Multistage sampling refers to procedures involving two or more steps, each of which could involve any of the sampling techniques.