Chapter 12 - Sample Surveys

Exploratory Data Analysis - Exploring data using pictures and numbers.

Statistical Inference - Answering specific questions about a large group based on a subset of the large group.

To use a small group to answer questions about a larger group there are three "ideas" we need.

1st. Obtain the small group.
   Population (p. 303) - Entire group of individuals of interest.

   Sample (p. 304) - Smaller group of individuals or subset of the population.

   Sample Survey (p. 304) - A study that asks questions of a sample in order to learn something about the population. (Opinion poll)

Bias (p. 304) - Any failure of a sample to represent the population fairly by over- or underemphasizing some characteristics of the population.
It is vital to avoid bias, because any data collected from biased samples will be flawed and useless.   Bias usually cannot be fixed after sample is drawn.
   The most common historical example of a biased sample is the Literary Digest survey for the 1936 Presidential election. Literary Digest sent out 10 million ballots on election between Alf Landon and Franklin Delano Roosevelt. 2.4 million were returned and showed that Landon would win by a margin of 57% to 43%. Actually, Roosevelt won by a margin of 62% to 37%.

   What went wrong? This was the Depression; their list came from phone numbers, drivers' registrations and memberships in organizations like country clubs.

   Now phone numbers can be used, but they need to be randomly computer generated to avoid missing unlisted numbers, cell phones, and recently relocated people.

2nd. Randomization needs to be used.

Every individual has a fair and random chance of being selected for the sample. Best defense against bias and other factors that might influence results.   Helps to insure, on average, that the sample is like the population.

3rd. The Sample Size needs to be chosen.

The Sample Size is the number of individuals in a sample and is usually represented by n. It is not the percent or fraction of a population that is important, but the size of the sample that is important.

A larger sample will usually increase the accuracy of the results, but it may not be worth the cost.

Why not contact everyone? **Census** (p. 307) - A study that contacts every member of the population.

Problems: Cost, Complex, and Undercount or not being able to contact everyone.

Parameter (p. 307) - A numerical value used in the model of a population. Usually a Greek letter.

Statistic (p. 307) - A numerical value from a sample used to estimate a parameter.

Measure	Statistic	Parameter
Mean	$\overline{x}$	μ (mu)
Standard Deviation	s	σ (sigma)
Correlation	r	ρ (rho)
Proportion	$\hat{p}$	p
Regression Coefficient	b	β (beta)

Sample Variability (p. 310) - The natural tendency of randomly drawn samples to differ. This is not an error. This also means that the statistic will differ from sample to sample.

Good Types of Sampling

1. **Simple Random Sample (SRS)** (p. 309) – Method where every sample of n individuals from a population has an equal chance of being selected.

   Sampling Frame (p. 309) - The individuals in a population that a sample is chosen from.

2. **Stratified Sampling** (p. 310) – Method where population is divided into groups of similar individuals, called strata (like male and female). Then chose a separate Simple Random Sample in each stratum and combine to get the sample.

3. **Cluster Sampling** (p. 311) - Method where population is divided into representative parts or clusters. Then randomly select one or a few clusters and do census within the selected clusters. Clusters are representative of the population, so each cluster is similar.

4. **Multistage Sampling** (p. 312) – Method where several sampling schemes are combined to create a sample.
   Example: Divide population into geographical areas. Randomly select cities from each region. Then randomly select clusters of residents from each city.

5. **Systematic Sampling** (p. 313) - A sample drawn by selecting individuals systematically from a sampling frame. Example: Select every 7th person on an alphabetical list of students. The first selection must be randomly selected in order for the sample to be random.

Samples to Avoid (They will almost always be biased.)

1. **Voluntary Response Sample** (p. 318) – Method that consists of people who choose to respond to a survey. Examples: Call in Poll, Write in Poll, Internet Poll. Bias since people with strong opinions are more likely to respond. Also, people with negative opinions tend to respond more than those with positive opinions.

2. **Convenience Sampling** (p. 319) - Method that is chosen because the individuals are convenient to contact. Bias since sample will not be representative of the population. Surveys at a mall or on the internet are convenience samples. Only mall shoppers (who tend to have more money) will be represented, and if an interviewer is used they will tend to only contact people who seem "safe" to them.

Bias can be introduced into sample because of the following mistakes:

1. **Undercoverage** (p. 319) - When some groups of the population are left out. Examples: 900# polls, Telephone polls, Magazine Polls

2. **Nonresponse Bias** (p. 320) - When large number of individuals chosen for the sample cannot be contacted or choose not to cooperate.   If a survey is too long, people are more likely NOT to participate.

3. **Response Bias** (p. 320) - When something in the survey can influence the response of individuals.   Examples: Confusing or leading questions, Respondent may be unwilling to answer correctly questions about private, illegal, or inappropriate behavior. The sex, race, attire, or attitude of the interviewer can influence the response. The interviewer can lead individuals to responses.

   Surveys by special interest groups or for political parties will often include their side of an issue before asking the question.