Chapter 12 Summary
Sample Surveys

What have we learned?

- A representative sample can offer us important insights about populations.
  - It's the size of the same, not its fraction of the larger population, that determines the precision of the statistics it yields.
- There are several ways to draw samples, all based on the power of randomness to make them representative of the population of interest:
  - Simple Random Sample, Stratified Sample, Cluster Sample, Systematic Sample, Multistage Sample
- Bias can destroy our ability to gain insights from our sample:
  - Nonresponse bias can arise when sampled individuals will not or cannot respond.
  - Response bias arises when respondents' answers might be affected by external influences, such as question wording or interviewer behavior.
- Bias can also arise from poor sampling methods:
  - Voluntary response samples are almost always biased and should be avoided and distrusted.
  - Convenience samples are likely to be flawed for similar reasons.
  - Even with a reasonable design, sample frames may not be representative.
    - Undercoverage occurs when individuals from a subgroup of the population are selected less often than they should be.
- Finally, we must look for biases in any survey we find and be sure to report our methods whenever we perform a survey so that others can evaluate the fairness and accuracy of our results.

Background

- We have learned ways to display, describe, and summarize data, but have been limited to examining the particular batch of data we have.
- We'd like (and often need) to stretch beyond the data at hand to the world at large.
- Let's investigate three major ideas that will allow us to make this stretch…

Idea 1: Examine a Part of the Whole

- The first idea is to draw a sample.
- We'd like to know about an entire population of individuals, but examining all of them is usually impractical, if not impossible.
- We settle for examining a smaller group of individuals—a sample—selected from the population.
- Sampling is a natural thing to do. Think about sampling something you are cooking—you taste (examine) a small part of what you're cooking to get an idea about the dish as a whole.
- Opinion polls are examples of sample surveys, designed to ask questions of a small group of people in the hope of learning something about the entire population.
  - Professional pollsters work quite hard to ensure that the sample they take is representative of the population.
  - If not, the sample can give misleading information about the population.

Bias

- Samples that don't represent every individual in the population fairly are said to be biased.
    - Bias is the bane of sampling—the one thing above all to avoid.
    - There is usually no way to fix a biased sample and no way to salvage useful information from it.
- The best way to avoid bias is to select individuals for the sample *at random*.
    - The value of deliberately introducing randomness is one of the great insights of Statistics.

Idea 2: Randomize

- Randomization can protect you against factors that you know are in the data.
    - It can also help protect against factors you are not even aware of.
- Randomizing protects us from the influences of *all* the features of our population, even ones that we may not have thought about.
    - Randomizing makes sure that *on the average* the sample looks like the rest of the population.
- Not only does randomizing protect us from bias, it actually makes it possible for us to draw inferences about the population when we see only a sample.
- Such inferences are among the most powerful things we can do with Statistics.
- But remember, it's all made possible because we deliberately choose things randomly.

Idea 3: It's the Sample Size

- How large a random sample do we need for the sample to be reasonably representative of the population?
- It's the size of the sample, not the size of the population, that makes the difference in sampling.
    - Exception: If the population is small enough and the sample is more than 10% of the whole population, the population size *can* matter.
- The *fraction* of the population that you've sampled doesn't matter. It's the *sample size* itself that's important.

Does a Census Make Sense?

- Why bother determining the right sample size?
- Wouldn't it be better to just include everyone and "sample" the entire population?
    - Such a special sample is called a census.
- There are problems with taking a census:
    - It can be difficult to complete a census—there always seem to be some individuals who are hard to locate or hard to measure.
    - Populations rarely stand still. Even if you could take a census, the population changes while you work, so it's never possible to get a perfect measure.
    - Taking a census may be more complex than sampling.

Populations and Parameters

- Models use mathematics to represent reality.
    - Parameters are the key numbers in those models.
- A parameter that is part of a model for a population is called a population parameter.
- We use data to estimate population parameters.
    - Any summary found from the data is a statistic.
    - The statistics that estimate population parameters are called sample statistics.

Notation

- We typically use Greek letters to denote parameters and Latin letters to denote statistics.

Name	Statistic	Parameter
Mean	$\bar{y}$	μ (mu, pronounced "meeoo," not "moo")
Standard deviation	s	σ (sigma)
Correlation	r	ρ (rho)
Regression coefficient	b	β (beta, pronounced "baytah"[5])
Proportion	$\hat{p}$	p (pronounced "pee"[6])

Simple Random Samples

- We draw samples because we can't work with the entire population.
    - We need to be sure that the statistics we compute from the sample reflect the corresponding parameters accurately.
    - A sample that does this is said to be representative.
- We will insist that every possible *sample* of the size we plan to draw has an equal chance to be selected.
    - Such samples also guarantee that each individual has an equal chance of being selected.
    - With this method each *combination* of people has an equal chance of being selected as well.
    - A sample drawn in this way is called a Simple Random Sample (SRS).
- An SRS is the standard against which we measure other sampling methods, and the sampling method on which the theory of working with sampled data is based.
- To select a sample at random, we first need to define where the sample will come from.
    - The sampling frame is a list of individuals from which the sample is drawn.
- Once we have our sampling frame, the easiest way to choose an SRS is with random numbers.
- Samples drawn at random generally differ from one another.
    - Each draw of random numbers selects *different* people for our sample.
    - These differences lead to different values for the variables we measure.
    - We call these sample-to-sample differences sampling variability.

Stratified Sampling

- Simple random sampling is not the only fair way to sample.
- More complicated designs may save time or money or help avoid sampling problems.
- All statistical sampling designs have in common the idea that chance, rather than human choice, is used to select the sample.
- Designs used to sample from large populations are often more complicated than simple random samples.
- Sometimes the population is first sliced into homogeneous groups, called strata, before the sample is selected.
- Then simple random sampling is used within each stratum before the results are combined.
- This common sampling design is called stratified random sampling.
- Stratified random sampling can reduce bias.
- Stratifying can also reduce the variability of our results.
    - When we restrict by strata, additional samples are more like one another, so statistics calculated for the sampled values will vary less from one sample to another.

Cluster and Multistage Sampling

- Sometimes stratifying isn't practical and simple random sampling is difficult.
- Splitting the population into similar parts or clusters can make sampling more practical.
    - Then we could select one or a few clusters at random and perform a census within each of them.
    - This sampling design is called cluster sampling.
    - If each cluster fairly represents the full population, cluster sampling will give us an unbiased sample.
- Cluster sampling is not the same as stratified sampling.
    - We stratify to ensure that our sample represents different groups in the population, and sample randomly within each stratum.
        - Strata are homogeneous, but differ from one another.
    - Clusters are more or less alike, each heterogeneous and resembling the overall population.
        - We select clusters to make sampling more practical or affordable.
- Sometimes we use a variety of sampling methods together.
- Sampling schemes that combine several methods are called multistage samples.
- Most surveys conducted by professional polling organizations use some combination of stratified and cluster sampling as well as simple random sampling.

Systematic Samples

- Sometimes we draw a sample by selecting individuals systematically.
    - For example, you might survey every 10th person on an alphabetical list of students.
- To make it random, you must still start the systematic selection from a randomly selected individual.
- When there is no reason to believe that the order of the list could be associated in any way with the responses sought, systematic sampling can give a representative sample.
- Systematic sampling can be much less expensive than true random sampling.
- When you use a systematic sample, you need to justify the assumption that the systematic method is not associated with any of the measured variables.

Who's Who?

- The *Who* of a survey can refer to different groups, and the resulting ambiguity can tell you a lot about the success of a study.
- To start, think about the population of interest. Often, you'll find that this is not really a well-defined group.
    - Even if the population is clear, it may not be a practical group to study.
- Who's Who? (cont.)
- Second, you must specify the sampling frame.
    - Usually, the sampling frame is not the group you *really* want to know about.
    - The sampling frame limits what your survey can find out.
- Then there's your target sample.
    - These are the individuals for whom you *intend* to measure responses.
    - You're not likely to get responses from all of them—nonresponse is a problem in many surveys.

Who's Who? (cont.)

- Finally, there is your sample—the actual respondents.
    - These are the individuals about whom you *do* get data and can draw conclusions.
    - Unfortunately, they might not be representative of the sample, the sampling frame, or the population.
- At each step, the group we can study may be constrained further.
- The *Who* keeps changing, and each constraint can introduce biases.
- A careful study should address the question of how well each group matches the population of interest.
- One of the main benefits of simple random sampling is that it never loses its sense of who's *Who*.
    - The *Who* in an SRS is the population of interest from which we've drawn a representative sample. (That's not always true for other kinds of samples.)

What Can Go Wrong?—or, How to Sample Badly

- Sample Badly with Volunteers:
    - In a voluntary response sample, a large group of individuals is invited to respond, and all who do respond are counted.
        - Voluntary response samples are almost always biased, and so conclusions drawn from them are almost always wrong.
    - Voluntary response samples are often biased toward those with strong opinions or those who are strongly motivated.
    - Since the sample is not representative, the resulting voluntary response bias invalidates the survey.
- Sample Badly, but Conveniently:
    - In convenience sampling, we simply include the individuals who are convenient.
        - Unfortunately, this group may not be representative of the population.
    - Convenience sampling is not only a problem for students or other beginning samplers.
    - In fact, it is a widespread problem in the business world—the easiest people for a company to sample are its own customers.
- Sample from a Bad Sampling Frame:
    - An SRS from an incomplete sampling frame introduces bias because the individuals included may differ from the ones not in the frame.
- Undercoverage:
    - Many of these bad survey designs suffer from undercoverage, in which some portion of the population is not sampled at all or has a smaller representation in the sample than it has in the population.
    - Undercoverage can arise for a number of reasons, but it's always a potential source of bias.

What Else Can Go Wrong?

- Watch out for nonrespondents.
    - A common and serious potential source of bias for most surveys is nonresponse bias.
    - No survey succeeds in getting responses from everyone.
        - The problem is that those who don't respond may differ from those who do.
        - And they may differ on just the variables we care about.
- Don't bore respondents with surveys that go on and on and on and on…
    - Surveys that are too long are more likely to be refused, reducing the response rate and biasing *all* the results.
- Work hard to avoid influencing responses.
    - Response bias refers to anything in the survey design that influences the responses.
    - For example, the *wording* of a question can influence the responses:

How to Think About Biases

- Look for biases in any survey you encounter—there's no way to recover from a biased sample of a survey that asks biased questions.
- Spend your time and resources reducing biases.
- If you possibly can, pretest your survey.
- Always report your sampling methods in detail.