Chapter 1

Key Ideas

Terms: Data (Quantitative vs. Qualitative), Discrete vs. Continuous Data, Statistics, Population, Census, Sample, Parameter, Statistic, Observational Study vs. Experiment

Bias: Voluntary Response, Small Samples, Misleading Graphs/Percentages, Loaded Questions, Nonresponse, Missing Data, Correlation vs. Causality, Self-Interest, Precision, Partial Pictures

Obs. Studies: Cross-Sectional Study, Retrospective (Case-Control) Study, Prospective (Cohort) Study

Experiments: Confounding Variables, Single/Double Blind, Blocking, Randomization, Replication

Sampling: Convenience, Simple Random, Stratified, Cluster, Systematic, Multistage, Sampling vs. Nonsampling Error

Section 1-1: Overview

Why Statistics?

Before beginning the study of statistics, it is important to understand why it is even needed in the first place. We need statistics because we want to know something about the world, but because of *random* processes, we can only make educated guesses. For example, let's say you want to start a new organization on campus, but you are not sure how much of the student body would even be interested in it. You ask some people who live on your residence hall floor, and they seem to be excited about it. However, when you ask some people in your classes, they don't like the idea. It would be great if you could find a way to ask everyone on campus if they would be interested in your organization. Unfortunately, this would take too much time. This is where statistics comes in. Even though you can't know for sure what percentage of the population likes your idea, you could find a way to get a small group of people together who represent most of the variation in the student body (in terms of race, gender, sexual orientation, economic status, political affiliation, etc.). Then this group of people should be a rough microcosm of the entire campus, and their responses to your questions should generally match the campus as well. This is how statistics is used. We want to know something about the world that is either too difficult or impossible to observe. Therefore, we use logical thinking and mathematical principles (often it's common sense) to make an educated guess about what the true value is. The good thing about statistics is that we can even quantify how accurate and precise that measurement is (e.g. margin of error in polling). Since statistics is not tied to any one application, it is used in any situation where something is uncertain (business, medicine, aeronautics, physics, politics, athletics, weather forecasting, and so forth).

Here is some basic terminology we'll be using in class:

- **Population** – The entire group you want to know something about (like the entire student body above)
- **Sample** – The group you use to infer something about the population (the representative group you chose in the example above)
- **Data** – Collected observations from a study, experiment, etc. (the yes/no responses from the students in the sample)
- **Census** – Collection of data from everyone/everything in the population (usually hard to do or impossible)

Section 1-2: Types of Data

More terminology:

- **Parameter** – A value measuring some trait of the population (the percent of all students on campus who like your idea)
- **Statistic** – A value measuring some trait of the sample (the percent of students in your smaller sample who like your idea)

A note about the word "data": A lot of people don't realize it, but *data* is plural for *datum*. So although people often say things like "the data shows that…", they should really be saying "the data show that…". Just thought you'd like to know.

Data comes in two flavors: Quantitative and Qualitative.

- **Quantitative** data is in number form… this is also sometimes called **measurement** data.
  Examples: Height, Weight, Age (Years), Die Roll on a 6-sided die, Distance (miles), Shoe Size, etc.
  It can also be subdivided into two sub-classes:
  - *Discrete* – Numbers that aren't densely packed together (you can separate all the possible response values and count them)
    Examples: Age (Years), Die Roll, Shoe Size
  - *Continuous* (*Numerical*) – Numbers that can be very close to each other (there are infinitely many possible response values)
    Examples: Height, Weight, Distance (miles)
  - As a general rule of thumb, you can think of continuous data as data that can have decimal places, whereas discrete data do not (except for maybe .5 on the end in the shoe size example, since they run in half sizes too).

- **Qualitative** data (also called *Categorical* or *Attribute* data) can be separated into different categories and don't use numbers.
  Examples: Grade (A, B, C, D, F), Gender (M, F, Other), Economic Class (Lower, Middle, Upper), etc.

You can probably already see that it can be difficult to stick some variables into certain data types without knowing the context. If you take something like "Age," it could actually be in any of these groups. If Age is measured in years, but decimals are allowed (e.g. someone could be 14.236 years old), then it is continuous. If it is rounded off, then it is discrete. Furthermore, if you are doing a study on infants, say, and you only consider Age 0 (newborn) and 1 (one year old), then it could be thought of as qualitative data, since you have 2 categories. The point here is that these are just labels for data, and nothing is set in stone. You can use whatever terms you want to describe the data as long as you can make an argument for why the data should fall into that group.

Section 1-3: Critical Thinking

One general method for finding something out about a population is to take a sample and use it to make an inference. This is done by first determining the appropriate *statistic* for that sample, and then using it as an estimate of the *parameter* in question. However, unless that sample is large enough and representative of the population, the estimate can be very inaccurate. Furthermore, when the results of a study are displayed, there are often ways of exaggerating or dampening the relationships between groups that lead to misleading conclusions. The textbook has a lot of great examples concerning these issues, which are paraphrased below. Bias, by the way, is just a way of saying things are flawed, or skewed in one direction.

Sampling Biases

Voluntary Response Samples (*Self-Selected Samples*) – This happens when, instead of the researcher drawing a sample from the population, the people are choosing to be included in the sample. This often happens with internet polls on popular sports or political websites. What ends up happening is that only people who care enough about the issue (and visit that site) end up responding. This gives results that are incredibly biased.

Small Samples – If a sample is too small, the estimates of the parameter are less accurate. As an extreme example, suppose on election day, someone wants to see which president received more votes by giving an exit poll and seeing who people voted for. However, they only give the poll to two people. That isn't very likely to be accurate, is it?

Loaded Questions – Sometimes, the way a question is worded can change the way people answer it. Surveys about controversial topics are often places where this occurs. Asking "Do you support a woman's right to choose an abortion" is a lot different than asking "Do you support aborting a fetus which would otherwise grow into a healthy human being?" A researcher must be careful to present questions in a way that does not lead someone to a particular response. Similarly, writing questions in a particular order may influence how people respond as well.

Nonresponse – This occurs when someone is selected as part of a sample, but gives incomplete information. Most nationwide polls have very high nonresponse rates, and this can often be a problem if the nonresponse is a result of the survey design. It could be the case that most of the nonresponders had something in common, which isn't reflected in the results of the study.

Missing Data – This is the result of nonresponse, and it can result from survey design (low income people may not report their yearly salary) or random chance (in a phone survey, someone may not answer their phone). There is an entire field of statistics concerning how to deal with missing data.

Misleading and/or Erroneous Results

Misleading Graphs – The scale on graphs can often be increased or decreased to make differences appear larger or smaller than they actually are. This is very common, because most people look at the pictures when they read about studies, and not the numbers reported.

Percentages – Percentages are often misused in a variety of contexts. In my personal experience, I once saw a sign at a clothing store claiming that clothes that were currently on sale for 25% off would now by discounted by an additional 50%, for a total of 75% off. This isn't true. It's really just a little over 60% off, since the 50% off doesn't include the 25% already discounted before.

Correlation vs. Causality – Often, people make the mistake of assuming that since two variables are correlated (increasing one increases the other as well), that one must cause the other. This is not usually the case, however. Consider a fictional study where someone records the amount of ice cream eaten each month, as well as the number of drowning deaths. These two variables would be highly correlated. As ice cream sales increased, so would drowning. It is clear that eating ice cream does not cause drowning, however. This happens because *both* of the variables are controlled by a third variable, which is time of year. In the summer, the days are hotter – more people eat ice cream and go swimming. In the winter, no one eats ice cream or goes swimming.

Precise Numbers – When very specific numbers are stated in results, the tendency is for the reader to assume there is a reason the number is not rounded off. Often they assume this reason is that the number is more accurate. However, this is not the case. It is still just an estimate, even though it does not appear to be that way.

Partial Pictures – If a television company told you "All of our TVs sold in the last 50 years are still in operation," you would think the company made incredibly durable TVs. However, what would you think if that company had only been making and selling TVs for 6 months? While the claim is technically true, it isn't as fantastic as it is made out to be.

Section 1-4: Design of Experiments

There are two broad categories of studies to collect data:

Observational Study – Data are collected by observing characteristics of subjects, but they are not modified in any way.

Experiment – Data are collected by modifying subjects (by applying a treatment) and observing its effects.

Observational Studies

There are three general types of observational studies:

Cross-Sectional – Data are collected at one point in time.

Example: You want to see if kids who get allowances of $5/week or more like Pokemon, so you sample 30 different kids and measure both their allowance and whether they like Pokemon.

Retrospective (*Case-Control*) – Data are collected by looking back into the past.

Example: You want to see if smoking causes lung cancer. You take a sample of 30 people who already *have* lung cancer, then look back and see if they were long-term smokers.

Prospective (*Cohort* or *Longitudinal*) – Data are collected by following a group of people with common traits into the future.

Example: You want to see if smoking causes lung cancer. You take a sample of 30 people who currently smoke, then follow them for 10-20 years and see if they develop lung cancer.

Note: Prospective studies are often costly, because they take place over the course of many years. However, they have the advantage of letting the researcher form a group of people with traits of interest in common. Retrospective studies have the advantage of letting researchers examine causes of rare events, like particular diseases. If one wished to see if a particular characteristic led to a heightened risk of a rare disease, it is easy to draw a sample of people who already contracted the disease, and see if they share that characteristic. It is not so easy to form a group of people with that characteristic and then wait several years to see if any of them develop the disease (it would be likely that none of them would, simply due to the rarity of the disease).

Experiments

Experiments are carefully controlled studies in which the researcher tries to minimize **confounding**, which is an inability to separate the effects of multiple sources of variation. A source of variation is called a *factor*. This is generally done by three methods.

1. **Control** – This does not mean that every experiment needs to have a control group. It means that the researcher needs to identify potential confounding factors and design the study in such a way that those factors cannot have an effect on the response. There are a couple of ways to do this depending on which factors may be a problem.
   - *Blinding* – This is often used to take care of the "placebo effect," where people who take medicine may start to feel better simply because they swallowed a pill, even if the medicine in the pill has no effect. In medical studies, researchers don't want to have a situation where a pill actually does nothing to cure a disease, but people still feel better because of mental reasons. As a result, they will often use a *Single-Blind* experiment, where the subjects are not told whether they are taking a placebo (fake pill) or the real thing. Often, *Double-Blind* experiments are used as well, where even the doctors giving the pills to the subjects do not know whether the pill is real or not.
   - *Blocking* – This technique is used to deal with factors that cannot be controlled or assigned to subjects. Perhaps a scientist wants to see if a drug helps to prevent heart disease, but realizes that it may have different effects on men and women. Since subjects cannot be assigned a gender, blocking can be used. Essentially, each block of subjects (in this case the blocks are males and females) is treated separately, and the experiment is run on the blocks. Then, the researcher can compare the effects of the drug within each block (which eliminates the gender effect) and/or compare the effects between the blocks (to see if there is a difference between the effect on men and women).
2. **Randomization** – In order to take care of confounding factors that have not yet been identified, scientists always randomly assign treatments to the subjects in an experiment. The goal is the "even out" the differences between the subjects through a random process, and severely dampen any effects those unknown factors may have.
3. **Replication** – The more data the researcher has, the more accurate the results will be. Furthermore, if another person is able to repeat the results of a study, the conclusions are more powerful. Scientists often try to replicate (run more trials, use more subjects, etc.) if possible within the constraints of time and money.

Sampling Methods

Convenience Sample – Obtain the easiest sample you can get (this is a bad idea)

Random Sampling – Any method where every member of the population has an equal chance of being selected. (all of the methods below fall into this category)

Simple Random Sample – Assign a number to every member of the population, and then use a random mechanism to select n of them, where n is the sample size.

Stratified Sample – Split the population into groups (strata) of members sharing common characteristics. Ensure that each member belongs to one and only one stratum. Then take a simple random sample from each stratum. The goal here is for the strata to be homogeneous (the members are very similar).

Cluster Sample – Split the population into groups (clusters) that are heterogeneous (the members are very different). Again, each member should belong to one and only one cluster. Then randomly select a few clusters and sample *all* members of the clusters.

Systematic Sample – Put all of the members of the population into number order, then select every k^{th} member, where k is an integer.

Multistage Sample – A combination of random sampling schemes at different stages in the design. (see page 29 for an example)

A Final Note: There are two kinds of errors arising from sampling schemes. The first is *sampling error*, which is the naturally occurring variation between data collected from a sample compared to that of the population. The second is *nonsampling error*, which results from a biased study (see Sampling Biases) or poor sampling design (e.g. Convenience Sampling)