Chapter 1 Notes and elaborations for STAT 141-Introduction to Statistics

Assignment

Read all of Chapter 1 except for the following: you may skip the parts about nominal, ordinal, interval, and ratio levels of measurement. What the book calls the **Hawthorne effect** is what most of the rest of the world (including myself) call the **placebo effect**. You may skip section 1-5. Also read chapter 14-1. for another take on "Stratified Sampling" and "Cluster Sampling." In all places in this book you may skip the discussion about using a table of random numbers. We have better ways these days.

You should be able to do the following exercises: (Answers are in an appendix in the book.)

1-1 problems 1-15 only the odds on page 5.
1-2 problems 1-15 only the odds on page 10.
1-3 problems 7-15 only the odds on page 17.
1-4 problems 1, 3, 7, 9, 11,15, 17, 23, 25 starting on page 24.

The book does a pretty good job most of the time. However, I think the book misses quite a few key points in Chapter 1 and emphasizes some items that aren't relevant to the scope of this course. Below I will attempt to fix what I think is broken and/or needs some emphasis. After reading the above material, please read what I've written below.

There will be a quiz on the material in Chapter 1.

What is statistics?

Statistics is a pretty broad field of the mathematical sciences, and for most definitions of the field, one can find examples that break the definition, including mine. Statistics is the body of knowledge pertaining to understanding data. This includes both the art and science of: experimental design, organization and summarization of data, data analysis, and making predictions and drawing conclusions from data. It is the interface of science with both observation and modeling. It is almost completely predicated on probability theory.

A combination of probability theory and statistics will undoubtedly make you a better hold'em player. Although I say that tongue-in-cheek, much of basic probability was discovered via games-of-chance by Gerolamo Cardano in the 1500s.

A fun link (but unnecessary): http://en.wikipedia.org/wiki/Gerolamo_Cardano

There are two primary reasons that you should know something about statistics.

1) Your field of study and/or job requires it.

• Statistics is the interface to scientific experiments and observations, and every pure scientific field, i.e., astronomy, physics, chemistry, medicine, biology, zoology, engineering, etc., utilizes the power of statistics. Indeed, they would not be what they are without it. Virtually everything scientific in nature requires experiments and observations of real data, which is statistics.

• The social sciences, such as political science, sociology, economics, and psychology to name just a few, rely on surveys, designed experiments, regression theory, etc., to do just about anything.

• Statistics is also prevalent in the professional fields, such as business, medicine, education, etc. In fact, statistics is huge here. Reliability analysis, market predictions, development and testing of product, allocation of educational funds, and so much more, are all based on statistics.

• And don't forget industry; Six Sigma programs are everywhere (Six Sigma is a data-driven approach and methodology, i.e., statistical approach for eliminating defective products in industry). Japan's post World War II economic boom is at least partially due to industrial statistics. Mil-specs, ISO, etc.

• Maybe the fine arts as a field doesn't really need it.

2) You are a citizen in a capitalistic republic. Everyone is trying to sell you their products, and every politician is trying to gain your favor. Understanding the basics of statistics is a prerequisite for not being swindled.

• 'Political statistics' is more than just reading poll results, but understanding if they are constructed correctly and what actual information can be taken from them. If your BFF Doug Awell takes a poll on you favorite social-networking site and finds that 84% of the people support hydraulic fracturing in Pennsylvania, what information do you actually have? (Answer: you know that the type of people that answer polls on headbuk.com are in favor or fracking, and that's about it.)

• Oftentimes, whomever is selling you something, whether it be an idea or a product, will feed you numbers/information biased in their favor, for whatever purpose. As a consumer, you should question where the numbers/information came from so you can make an informed decision about the accuracy of the numbers/information. A basic understanding of statistics will help you do this.

• For example, I'm sure you've all heard the middle class is disappearing. Is it true? Are the rich getting richer while the poor are getting poorer? What effect will tax changes have? People debate this all the time, sometimes bitterly, and they all cite facts and figures from 'studies'. What is the truth? Or maybe a better question is, what information is closest to the truth? How can we tell?

One last thought...

Statistics is one of the most profoundly misunderstood and abused fields. I can not emphasize this enough.

• "Figures lie and liars figure" and "lies, damned lies, and statistics" are a result of decades of poor statistics understanding both on the part of experimenters (both intentional and unintentional) and the public.

• Statistics, when done right, can't lie. Indeed, we can use statistics to get as close to the truth as possible. Mind you, I am not saying the pure truth can ever be obtained.

So, let's get started...

Sampling Theory
We begin with basic definitions relative to sampling theory. We then briefly discuss some basic sampling methods and we end with the two types of studies.

1.1 **Definitions**

Definition: A **population** consists of all the subjects being studied. The population might not be human or animal. It could very well be jelly-donuts.

Example 1.1.0

i) If you're looking at gathering data concerning the nutritional value of the lunches served in public schools in the last year, your population consists of every lunch served in the public schools in the last year. That's about 4.9 billion lunches based on statistics for the year 2019.

ii) If you're studying side effects of a new medicine, your population consists of every individual that might every take the new medicine.

iii) If you're trying to determine if a particular medication will stop the onset of type II diabetes, then your population probably consists of all people that are likely to get type II diabetes.

iv) If you're trying to ascertain what percentage of bass (a type of fish) in Cowanesque Lake have a particular disease destroying their reproductive abilities, your population consists of all the bass in Cowanesque Lake that are mature enough to reproduce.

v) If you want to find the average number of chocolate chips in a Chips O'Boy cookie on any given day, your population is every Chips O'Boy cookie in the entire world on that day. If you want the average number of chocolate chips in a Chips O'Boy cookie over a given time period, say for example, during October 2013, then your population is every Chips O'Boy cookie in the world during that month.

Note that, most often, it is entirely impractical and/or too costly to study the entire population. So, we take samples. There are many ways to take samples, which will be discussed in more detail later. Keep in mind that our ultimate goal is to take samples (minimizing bias, which we will define later) and use the information we gather from these samples to make inferences about the entire population in general.

Definition: A **sample** is a subset of the population.

Example 1.1.1

Considering example (i) above about school lunches: Clearly, it's inefficient to record data about 4.9 billion lunches. So, we take a sample. For example, we could record data from two school districts in each county every third day for one month of the year. This sample will give a good estimate of the overall nutritional value of the school lunches served in our nation's public schools.

Considering (ii) above about a medication: It is not hard to see that we will most likely never get our hands on the entire population in this scenario. Indeed, we must be extra cautious in the sampling procedure we decide to use in such a situation. For example, it would probably be easiest to maintain a hotline (where the number is on the medication packaging) for people to call in and report side effects. But this assumes that (a) the consumer actually reads the packaging and (b) the consumer will take the time to call and report possible side effects, if they are even distinguishable from other factors, e.g, was it the new medicine, or the all you can eat shrimp at Red Lobster? At this point, it should be clear that an important question we should ask every time we take a sample: is the subset of likely respondents truly representative of the population?

Considering (iii) above: The only way we could get the whole population here is if we knew how to predict diabetes with 100% accuracy, and we don't, so getting the entire population is literally impossible in this scenario. Things get trickier in this situation, too. Intuitively, we take a sample of people who are at a certain level of risk for type II diabetes and give them the medicine. But we have to be careful in choosing this sample. Remember that a sample is a subset of the population. So, we want every person in our sample to have a very high probability of getting type II diabetes.

Considering (iv) above: We could actually get the whole population here, but we'd probably have to drain the lake or something to be sure we had them all. So, we take a sample, usually via capture and release method. This is a method of sampling commonly used in studying wildlife where we put some kind of mark on the wildlife and release it back into the wild. This way, if we take another sample, we don't record data twice for the same animal.

Considering (v) above: If only we could get the entire population of cookies! But we can't, so we'd have to take a sample. In this scenario, it's probably easiest to go to all the factories that make the cookies and grab samples before they are shipped to the stores, though this may not be the most cost effective sampling method.

Before going any further, I'd like to briefly comment on sample sizes. In many statistical books, you find the erroneous conclusion that $n \geq 30$ (n is the sample size) is some magical number that allows us to perform (essentially) more simple analysis than if $n < 30$. With regards to sample size and accuracy, it is true that a big sample is better than a small sample, but it isn't nearly as helpful as one might think.

Generally, to get twice the accuracy one must collect 4 times more data. Loosely stated, to double the accuracy of a sample of size 1000, you need a sample of 4000 subjects. And unless the sample is a large portion of the population, the population size doesn't matter at all. That is, a poll of 1000 people in Pennsylvania, with about 12.5 million people yields an accuracy that is virtually identical to a poll of 1000 people from Wyoming (population about 500,000) even though Pennsylvania has 25 times as many people as Wyoming. Not convinced? Suppose you make a small pot of chili. Provided the chili is well mixed, you can decide if it tastes good from a teaspoonful or two. Now suppose that you make enough chili for an entire university in a large vat. Provided it is well mixed, how much do you have to taste? Certainly not a gallon! (Although I've used this excuse before). A representative sample is much more important than the sample size (unless your sample is the whole population, of course). Moreover, there is an entire mathematical theory of optimizing sample sizes.

Definition: A **parameter** is a numerical fact about a population.

In most situations, the true value of a parameter is not and never will be known. In theory, we can discuss the average amount of vitamin B12 in school lunches, but we will never know the true average unless we study the entire population, which we have already stated would be impractical. The true average amount of vitamin B12 in school lunches is a parameter. Likewise, we can talk about the inferred population ratio of people experiencing side effects (inferred from the sample we take), but we will never know the true ratio of people experiencing side effects. It is a parameter.

Mathematical notation for parameters generally consist of lower-case Greek letters. For example, it is standard practice to use the letter μ (spelled as mu, spoken as "muh-you") to denote population average and σ^2 (that's "sigma-squared") to denote population variance, two parameters that we will discuss later-on. In these notes, μ denotes the parameter population average and σ^2 denotes the parameter population variance.

Definition: A **statistic** is numerical fact about a sample. (For the mathematically inclined, it is a function of all the collected data.)

Although we can not attain the true average amount of vitamin B12 in school lunches, we can estimate this parameter using the average amount of vitamin B12 in the school lunches that were part of our sample. The average amount of vitamin B12 that were part of the sample is a statistic. Said another way, the statistic that is the average amount of vitamin B12 in the school lunches that were part of our sample helps us to estimate the true average amount of vitamin B12 in school lunches nationwide. Whether or not it will be a good estimate depends on both the statistic and the experiment itself.

In general, we use a statistic to estimate a parameter. Note that I will use the words statistic and estimator interchangeably, usually without mention.

In many cases, the computation of a statistic is done in a similar manner as one would compute a parameter. For example, consider the average. We find averages by adding all the data and dividing by the size of the data set. The average of 4, 6, and 11 is 7. (.) Both population averages and sample averages are computed in this same manner.

Mathematical notation for statistics differs from that for parameters. The sample average is usually denoted $\overline{X}$. The sample variance is usually denoted by s^2 (or sometimes by S^2). We will follow this notation in these notes.

Example 1.1.2

i) The sample average, $\overline{X}$ is an estimator of μ, the population average.

ii) The statistic s^2, which is the sample variance estimates σ^2, the population variance.

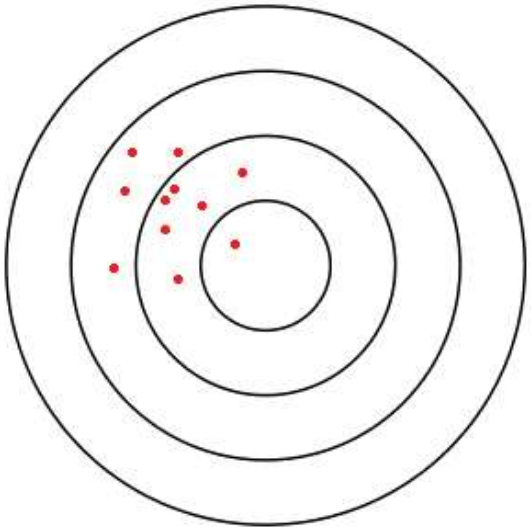
iii) When studying ratios, the sample ratio or proportion is usually denoted as $\hat{p}$ (we say 'p hat'). It is an estimator of the population proportion p.

Definitions 1.1.4 All three examples given above are **unbiased** estimators for the parameters they are estimating. Just like everything in statistics, there is a precise mathematical definition for this, and not too surprisingly, it requires the tools of calculus.

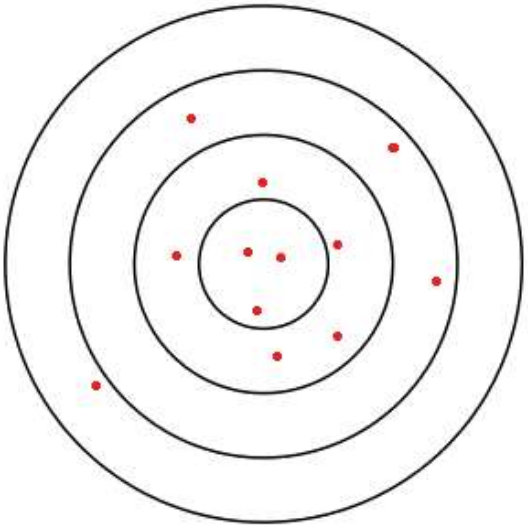
An estimator is **unbiased** if the expectation of the estimator is equal to the parameter it is estimating. This isn't the best example here, but this might help a bit. If a dart-thrower is aiming for a bullseye, and the long term average of the throws is right in the middle, then the dart-thrower is unbiased. If most of the time the throws are off in one direction then the darts are biased.

`

Biased towards the left and a bit upwards.



Although the variability is higher, this looks mostly unbiased.



Example 1.1.3

Suppose a researcher is interested in the weight of the heaviest human that has ever lived. This weight is a parameter. It is a numerical fact about the entire population of humans throughout history. There is an value for this parameter, but nobody knows it, and without a time machine, nobody ever will. One way to estimate it is to consider the largest ever measured weight of a human. This is a statistic, based on the sample of all humans ever weighed. It is an example of a biased estimator, as well. An estimator is biased if it tends to be either too big or too small relative to the parameter it is estimating. The odds are that the true weight of the heaviest human ever is larger than the statistic, but it cannot be smaller (think about this). So, this particular statistic (which is an estimator) is expected to be smaller than or equal to the parameter it is estimating, i.e., the expected value of the statistic is less than the parameter being estimated. Thus, this is a biased estimator.

Definition 1.1.5 A **biased sample** is a sample obtained via a sampling method that produces values which systematically differ from the values of the population being sampled, i.e., it is not representative of the population being studied.

The following examples illustrates some different types of bias in sampling. Note that sampling humans is generally very difficult.

Example 1.1.4

The Literary Digest was a popular magazine that ran from 1890-1938. They are well-known for a massive poll they took for the 1936 presidential election. The election was between Alfred Landon (Republican) and the incumbent democrat, Franklin D. Roosevelt. The Digest had correctly called the previous 5 elections.

They sampled simply by mailing surveys –10 million of them. By contrast, most modern polls use samples of between a few hundred and a few thousand people. They gathered names from lists of automobile owners, subscribers, voter registration records, and social clubs like the Elks, Moose, etc. They got a remarkable 23% response rate; 2.3 million ballots were returned. You may also consider that the total population of the US was around 130 million at the time, and not all of them can vote.

At the same time, George Gallup took a sample of 10,000 people, still a pretty large survey. Here's a table of the actual popular vote for FDR, the Literary Digest (TLD) prediction of the popular vote for FDR, Gallup's prediction, and amusingly, Gallup's prediction of what TLD would predict.

	FDR Popular vote percentage
Actual:	60.8%
The Literary Digest prediction:	43%
Gallup's prediction:	56%
Gallup's prediction of TLD's prediction:	44%

This is a spectacular failure of epic proportions! To contrast this with modern polls: in 2020 the best ten polls or so used between 800 and 9000 people in their sample. All of them predicted Joe Biden would receive between 54% and 56% of the popular vote. The actual was 51.3%. (The error in the polls was in general due to bias about who would actually cast their vote and other technical matters.)

So, why was Literary Digest so wrong? They used a biased sampling method. The technique they used is called convenience sampling, which is a catch-all term for a sample with no reasonable probabilistic basis. It has just about every feature you don't want in a sample. A couple of examples:

• It most certainly contained nonresponse bias, where people who don't respond may be different than those that do respond. It's very likely that the people who returned the survey were systematically different than those who didn't.

• It certainly had selection bias, i.e., it was not a representative sample of the population. That is, the people who they sampled were different than those they didn't sample. In this case, poorer and more rural people were not sampled nearly as often as those that were affluent. As is the case now, poorer and rural people vote differently.

Gallup looked the part of a hero here, although their technique was quite flawed, as well. The Gallup poll of 1936 used a method called quota sampling (described below), which is also a biased sampling method. It just happens to be less biased than what Literary Digest used.

1.2 Sampling techniques/methods

• **Census**: A census is a sample that consists of the whole population.

Sometimes, this is simply impossible. For example, the national census, although called that, can never be a true census because, for example, of the many homeless that are inevitably not-counted.

Sometimes, taking the whole population doesn't make sense. Consider that, to test cement blocks for strength, you destroy the block. A census would leave no cement blocks!

Usually, though, a census is just too expensive and/or time consuming. Suppose that your population consists of 1 million people, and your study costs 2 dollars per person (a relatively cheap experiment). So you need at least 2 million dollars to conduct your census.

• **Random**: A random sample has the property that every element in the population is equally likely to be chosen.

Random is good. We like random. Random sampling guards against any pattern in the data that is either known or unknown. More often than not, the patterns are unknown. Say you're a teacher, and you want to survey your students' reactions to some method you've implemented into your lecture to try and gauge its effectiveness. But, you're in a lecture hall with 300 students. It would be very inefficient to read 300 surveys. So, you randomly select a student and then have students count off by ten, taking each 10th student to be in your sample. This is a random sample. But this is not a simple random sample.

• **Simple Random** (**SRS**): A simple random sample has the property that every subset of size n of the population is equally likely to be chosen. After this chapter, all our samples will be assumed to be SRS unless otherwise indicated.

This is even better than random. Using the prior example, once you've randomly selected the initial student, not every subset of size 30 has an equally likely chance of being selected. So, instead, say that students draw from a hat, and whoever draws a red ticket takes the survey (there are 30). This is a simple random sample. It is a random sample, also. Every subject in the population has an equally likely chance of being in the sample, and every possible sample of size 30 has an equally likely chance of being selected. So, simple random samples are random samples, but random samples are not necessarily simple random samples.

• **Stratified**: For this type of sample, the population is divided into groups (strata) and then simple random samples are taken from within the groups.

In the hands of a statistician, this sampling technique is more powerful (in that we get better estimates) than SRS when information is known about the groups. Consider a study of the average education-level of residents of Pennsylvania. If we are stratifying by districts, a stratified sample might take a simple random sample of 100 people from each school district in each of the ten state regions. If we are stratifying by region, a stratified sample might consist of taking a simple random sample of a few school districts in each of the ten regions and then taking a simple random sample of the people in each of the districts.

• **Cluster**: This sampling method is similar to stratified sampling, but in this approach, the population is divided into groups and entire groups are chosen via random selection and every member of the chosen groups is in the sample.

Consider the example given above for stratified sampling. If instead of taking a sample from each district, we instead randomly choose say 20 districts and sample everybody in these 20 districts, we will have a cluster sample. If we randomly chose 4 regions and took a sample in every district in those 4 regions, we will have a cluster sample.

Notice here how easy it is to combine stratified and cluster sampling. For example, we could randomly choose 4 regions (cluster) and then take a simple random sample of school districts in each of the 4 regions (stratified). In reality, we do what's called complex sampling, and it often incorporates both stratified and cluster sampling.

• **Convenience**: Surveys without statistical methodologies, website polls, newspaper and TV call-ins, and many more, all fall into this category. This is a terrible way to sample. This is a catch-all category of "bad" techniques.

If it was an easy sample to get, chances are it is this type of sample. Convenience sampling is a catch-all term for any sample that has no statistical methodology. These are junk. They typically only include those members of the population who are available and/or willing to participate in the study. In other words, not every eligible member of the population has an equally likely chance of being in the sample. They are the epitome of biased, and show up all too frequently.

• **Quota**: This sampling method is like stratified sampling, but convenience sampling, instead of the SRS method, is used within each group. This is a terrible way to sample.

The Gallup poll of 1936 from Example 2 used this method. This is not quite as junky as convenience sampling alone, but it's still junky, which is why Gallup's numbers for this particular poll were considerably different than the true numbers, as well.

• Systematic: This type of sample consists of every k th element of the population.

This type of sampling isn't interesting to me; we have good random number generators these days.

• Complex, Multistage and others: These are beyond the scope of this class. Indeed, there are many more sampling methods (some quite exotic), but they are all well beyond the scope of this class.

Example 1.2.0

Suppose that we have 100 pieces of chocolate candy. They are separated into 7 piles, made by 7 different people, that I'll label with letters.

Pile A has 10 pieces.
Pile B has 10 pieces.
Pile C has 10 pieces.
Pile D has 15 pieces.
Pile E has 15 pieces.
Pile F has 20 pieces.
Pile G has 20 pieces.

If we were to put a unique number on each piece of chocolate, completely disregarding the pile that it came from, and sample 15 randomly (where each of the 100 pieces equally-likely to be sampled) then we have taken a simple random sample.

If we were to choose 2 pieces randomly from each pile, and with equal-likelihood within each pile, then we have taken a stratified sample.

If we were to randomly choose two piles and sample all the chocolate from the two piles then we would have taken a cluster sample.

Before we sample, though, we must design a study since the design of the study will often dictate the sampling method.

1.3 Two types of studies

Before we discuss the two types of studies, we first need a couple of definitions.

Definition 1.3.0 In research, the term **variable** refers to the measurable characteristics, qualities, traits, or attributes of a particular individual, object, or situation being studied.

Definition 1.3.1 Confounding variables are any extra variables not controlled for that can affect the outcome of a study.

There are **observational studies** and there are **experimental studies**. Experimental studies involve the deliberate manipulation of variables while observational studies are passive. Observational studies are any studies that are not experimental in nature. These are almost always cheaper and faster, but they are much more prone to confounding variables. The following example illustrates a confounding variable within an observational study.

Example 1.3.0

Every year or so an observational studies concerning the health benefits or drawbacks of drinking coffee is published. For many years it was believed that drinking coffee lead to a lower life span. A more carefully planned study noticed that, within the population of coffee drinkers, there is a much higher smoking rate. That is, people who drink coffee are far more likely to be smokers. (They figured out why as well. But it's off topic.) By employing advanced techniques, this can be properly controlled, and consequently, most studies show that moderate consumption of coffee is not a large risk, and sometimes there are benefits, e.g. it seems to lower the suicide rate. My personal observation is that drinking coffee leads to studying mathematics as a profession. :-)

These studies are observational and not experimental. People drink the amount of coffee that they decide, not how much the doctors tell them to drink. It would be experimental if people were made to drink a certain amount of coffee. The observational study above was confounded by smoking. That is, "the amount of cigarettes smoked" is a hidden third variable that interfered with the study.

Sometimes experimental studies are impossible because there are no variables that can be manipulated, or the variables that can be manipulated have serious ethical issues attached. Consider, for example, studying the cancer rate of smokers. We simply can not have people begin smoking or smoke more so that we can test relative cancer rates. Following is an example of a study that is forced to be observational in nature.

Example 1.3.2

Suppose we are studying whether or not performance on the Brilliance-Standard (BS) exam and shoe size are related. This is an observational study since it can not be experimental; we can not manipulate peoples' shoe size to see if it affects their BS score and vice versa. So, a sample of people take a BS test and their shoe size is measured. It is found that people with larger feet score much higher on the test. The researcher conducting the study concludes that people with larger feet have higher BS scores. Is their conclusion valid? Of course not. There is a confounding variable here. (No, there is no Brilliance-Standard exam, and any resemblance to any other exam is purely coincidental.)

The catch here is that children were included in this fictional example. The hidden bit of information, the confounding variable, is age. Generally speaking, people score better on exams as their age increases.

Although this example is kind of silly, it isn't far off from some real examples. As was previously noted, for years it was assumed that coffee was fairly dangerous to your health if consumed in relatively large quantities. Then someone realized that the heaviest smokers drank the most coffee. Caffeine is metabolized at almost twice the rate in smokers which allows smokers to drink twice as much coffee as nonsmokers without having adverse side-effects. So, until that realization, it was thought that the heaviest coffee-drinkers had the most health problems because of coffee. But, many of the health issues were due to smoking, which was a confounding variable in the studies. Most studies have potential to be affected by confounding variables, and so good researchers design their study accordingly.

Suppose now that we constructed the following (unethical) experimental study on the effects of drinking coffee.

Example 1.3.1

Say we have 300 subjects available for our experiment. We randomly assign 100 of these subjects to drink 6 cups of coffee each day, 100 to drink 2 cups of coffee each day, and 100 to abstain from coffee (oh no!). If the assignation of our subjects is truly random and the total sample is representative of the human population, then the confounding variable of smoking is controlled for since the smokers are randomly distributed in each of the three groups. The randomness protects the experiment from both known-confounding variables and unknown confounding variables.

When experimental studies are properly designed, confounding variables are controlled for, and the experiment is called a **controlled experiment**.

Moreover, when done properly, experimental studies are much, much more accurate. They have a very formal process and planning procedure. I will not cover this here, except for one very important part. Before an experiment is performed, each and every decision is already made. You would: design the study, state the research question, the sampling method, decide how to handle missing data, state each and every statistical method you will use, and so on. In fact, before a single datum is collected, there must be a roadmap for anyone to finish the entire project. A failure to do such is an invitation for data snooping and data fishing. Unfortunately this sort of malpractice is widespread and often unnoticed.

Experiments are better and generally more expensive, but when they involve humans they can sometimes have their own issues. There is a very real effect called the Placebo effect.

Definition 1.3.2 A **placebo** is a fake treatment, e.g., an inactive substance like sugar, distilled water, or saline solution that is, ideally, completely indistinguishable from the actual active treatment.

Definition 1.3.3 The **Placebo effect** is a phenomenon in which a placebo can sometimes improve a patient's condition simply because the person has the expectation that it will be helpful.

This is a real effect; we've measured it time and time again. If people know they are in a study, something psychological, but very real happens: their belief(s) can (and do) actually alter the outcome of the experiment. For example, people given pills with nothing but inert chemicals claim pain-relief.

A **randomized controlled double-blind experiment** controls for the Placebo effect. Controlled means some subjects receive no treatment, i.e., a group of subjects gets the placebo while the rest get the real medication. These are called the control and treatment groups, respectively. Randomized means that the control and treatment groups are chosen at random (as opposed to selected on some characteristic, even subconsciously), and double-blind means that neither the evaluator nor the subject knows who is in the control group or the treatment group.

For example, to decide if vitamin X really is good at preventing heart attacks, the following (unethical) example could be employed.

Example 1.3.3

A statistician picks two simple random samples from the population. (Ideally people from around the world. If it were people only from the United States it could very well be biased.) An "advanced vitamin X substitute" that is completely indistinguishable from the regular vitamin X is given by medical doctors to one of the samples, while vitamin X is administered by doctors to the other sample. The statistician makes sure the medical doctors don't know whether they are giving the substitute or the real vitamin X. The doctors then evaluate the people and report their findings to the statistician. The statistician then performs the exact statistical procedures that were decided upon before the experiment even began.

Why all the fuss? Humans have problems with bias, and we've taken all human-induced bias away in this example. The patients and doctors don't know if they are in the control group or the experimental group. Historically, either group knowing has been a huge problem. This is the "double-blind" part. It is a "controlled experiment" because we have people either getting a treatment or not getting a treatment, and they are randomized as well. (Letting people volunteer has historically sorted them into two different types of groups.) The last part about the statistician performing procedures decided upon before the experiment is equally important. In statistics, there is usually a large number of procedures that could be used in a given situation. If you allow the statistician to pick how they will analyze the data after the data is obtained, they could try them all and pick the one that works best for their particular purpose. This is another type of bias called **data snooping**. Picking your statistical methods after collecting the data completely invalidates your experiment.

So, although experimental studies are better than observational studies, keep in mind that a good experiment is hard to run and requires significant planning. (There is a rich field of mathematics called design of experiment.)