

# 1 Getting Started

## 1.1 What is Statistics?

THE BIG IDEA: Since the Renaissance, we have relied more and more on science to answer questions about our world. Science answers questions by gathering (or generating) data—numbers. Alas, the data that we collect are incomplete—if I want to know the average salary of an American worker, it is quite impossible to actually measure each and every one of them!

If you can't measure all of them, then what do you do? Well, measure some of them, of course! The question, then, becomes: *what can our observations on this small group tell us about the large group?* This is the central idea of statistics:

**We have some large group, about which we have a question—that could be answered, if we could only measure every person (thing, event, whatever) in this group. However, the group is too large—so we measure some smaller portion of the group, and then use mathematics to take our observations from the small group and make reasonable observations (i.e., educated guesses) about the large group.**

INDIVIDUALS AND VARIABLES: We use the word *measure* quite loosely here. When *you* think "measure," you probably think about rulers and scales. However, simply noting the color of your shoes is a measurement. Asking someone if they think that AOL is for losers is taking a measurement. Anything that you can record is a measurement. More on this in a bit...

**The physical things that we measure are individuals.** Shoes; people; cars; salaries; bags of sugar; anything.

**The qualities that we measure are variables.** Color; weight; mass; length; opinion; whatever.

So, if our big question is "What is the average salary of an American worker?" then the individuals would be American workers, and the variable would be salary.

If our big question was "What percentage of people wear flip-flops?" (some of us don't!) then our individuals are people (who could wear flip-flops; it would probably be useless to ask someone who didn't have feet or couldn't afford shoes or something else like that); the variable would be whether or not a person wears flip-flops (yes or no).

QUANTITATIVE AND QUALITATIVE: Variables come in two important categories. The book goes a bit farther, breaking it down into four categories—these will only become important if you continue your studies in statistics.

**Quantitative variables measure quantities**—amounts; things that *must* be measured with a number. Weight, mass, length, time, amperage, volts, candelas, degrees...the list goes on.

**Qualitative variables measure qualities**—observations that do not necessarily need numbers. Color, gender, hometown, marital status, playing position, favorite music type...again, the list goes on. Often, when we measure this type of variable, we change the measurement a bit, so that the response is either yes or no—Is that red? Are you female? Were you born in Irmo?

DISCRETE AND CONTINUOUS: Of the two types, quantitative variables are far more important. Quantitative variables can be further divided into two sub-categories.

**Discrete variables have lots of gaps in the possible measurements.** If you remember some math from long ago, you once learned about Integers. Integers are discrete—there are gaps between every pair of integers.

*Discrete variables almost always are counters*—they count how many of something you've got. You can have two things, or three things, but not two and a half.

**Continuous variables have few (or no) gaps in the possible measurements.** Continuing the trip down memory lane, the Real numbers are continuous.

If the only thing keeping you from taking a more accurate measurement is the measuring device you're using, then you've probably got a continuous variable. *If it's quantitative, and it's not discrete, then it's continuous.*

**The group of all individuals that you *could* measure is the Population.** This is the group that, if you could measure all of them, would answer your question.

**The number that would answer your question is the Parameter.** The only way to actually know the value of a parameter is to measure every individual in the population. Of course, as we've pointed out, that's impossible.

**The group of individuals that you *actually* measure is the Sample.** Naturally, this group will be quite a bit smaller than the population.

**The number that you measure from the sample is the Statistic.** The average salary of these 1000 American workers; the percentage of people who wear flip-flops; the number of red Skittles in the bag—these are all statistics.

**Descriptive Statistics deals with describing samples.** We measure a sample and report what we saw. This is where we will focus our efforts.

**Inferential Statistics deals with taking the value of the statistic and making a reasonable observation (educated guess) about the value of the parameter.** We will not be doing (much of) this!

## ***1.2 Random Samples***

RANDOM SAMPLES: So—we can't measure the entire population. Instead, we'll measure a sample. How will this sample be collected? How will we decide which individuals to measure? It turns out that this is a terribly important question—it was only recently (about 100 years ago—that's quite recent in the grand scheme of things!) that it was determined that randomization is the key to the answer!

There are many ways to take random samples. Here are a few choice selections.

SIMPLE RANDOM SAMPLE: The King of all sample techniques. **A Simple Random Sample (SRS) is one where every group has an equal chance of being selected.** The best idea to have about the SRS is drawing names out of a hat (without replacement). Imagine that you have a hat (box, barrel; whatever) big enough to hold one slip of paper for each individual in the population. Now, stir them up good, then reach in and grab a handful. That's what an SRS is.

SYSTEMATIC RANDOM SAMPLE: **In this technique, you select every  $n^{\text{th}}$**  (fifth; or tenth; or forty-ninth; or one hundredth; the choices are endless) **subject.** More on this technique in a few moments.

**STRATIFIED RANDOM SAMPLE:** In this technique, you first divide the population into **homogenous groups** (groups where the individuals in the group have one thing in common), **then select an SRS out of each group**. For example, if you want to sample high school students, you might first divide them into freshmen, sophomores, juniors and seniors; then select an SRS out of each group.

**CLUSTER SAMPLE:** In this technique, you select a group of individuals that are already **together for some reason**. If you are sampling Skittles, then a cluster would be one bag. If you are selecting students, then a class of students form a cluster.

**CONVENIENCE SAMPLE:** In this technique, you select from a group of individuals that are **convenient to reach**. For example, if your population is paper clips, and you select from those that available at Wal-Mart, then you've taken a convenience sample. If your population is all high school students, and you select from those in the Greater Columbia area, then that's a convenience sample also.

**RANDOM NUMBER TABLE:** Putting names into a hat can be quite tedious. An equivalent method is to assign some sort of identifier (number) to each individual, and then to use a Table of Random Digits to randomly select some of the identifiers (and thus, some of the individuals).

Here's an excerpt of a Random Number Table.

```
18903 70748 95462 96071 37840 69113 39642 82605 05990 45506
41461 80792 53544 04455 85335 76521 61813 22135 70104 99081
50964 60996 41960 81466 72610 58137 37304 82159 09342 53251
78076 21813 83730 79511 08266 24344 27862 55050 30945 34410
56737 03541 50643 86219 99195 35797 72244 22952 83216 73054
```

Digit after digit...usually grouped in fives to make it easier to read.

EXAMPLES:

[1.] Let's say that we have 1800 high school students, and we want to select an SRS of 20 of them to measure. To use the table, we'd first have to assign identifiers (digits) to each student. How about we take an alphabetical listing of the students, and identify the first student as 1, the second as 2, the third as 3, and so on...

There is one problem. Each student must receive an identifier that has the same number of digits as everyone else—our method gives one digit to students 1-9; two digits to students 10-99; etc. How can we fix this?

Well, the last student gets four digits—1800—so everyone must have four digits. I suppose we could write the number 1 as 0001...that would do it!

Since every student has a four digit identifier, we must read four digits at a time from the table. Here are those digits again, with separators after every four digits.

```
1890|3 707|48 95|462 9|6071| 3784|0 691|13 39|642 8|2605|
```

The first group of four digits is 1890—but our last student has number 1800! These digits don't identify anyone, so we throw them out. The second, third, fourth, fifth and sixth groups have the same problem.

The seventh group is 0691, so student number 691 is selected to be a part of our sample.

The eighth group is 1339, so student number 1339 is selected.

Keep this process up until you've got all 20 students that you need.

[2.] Perhaps you noticed that in the last example, we had to throw out a lot of digits that we read from the table. This can happen when selecting an SRS. The systematic sample does not suffer from this as much. Remember that the systematic selects every  $n^{\text{th}}$  individual. To do this with a table of random digits, you need only to read ONE individual from the table; the rest are generated systematically.

In our example, we are selecting 20 of 1800 students. In order to fix the system (to know what to use for  $n$ ), we need to know how many groups of 20 there are in 1800. How do you do that?

Divide, of course! 1800 divided by 20 is 90; thus, this system will select every 90<sup>th</sup> individual.

Let's use a fresh set of random digits, already divided into groups of 4:

4146|1 807|92 53|544 0|4455| 8533|5 765|21 61|813 2|2135| 7010|  
4 990|81 50|964 6|0996| 4196|0 814|66 72|610 5|8137| 3730|

You've got to go quite a while before you get a hit—0996. But now you're done with the table. Starting with 996, we now select every 90<sup>th</sup> number after that.  $996+90 = 1086$ .  $1086+90 = 1176$ .  $1176+90=1266$ . Keep going!

1356, 1446, 1536, 1626, 1716...oops! If we add another 90, we get 1806. Our last student ended at 1800!

Here's how we fix that: if the last student is 1800, then 1801 could represent 1; 1802 could represent 2; 1806 could represent 6!

So we wrap around, back to the beginning of the list.

6, 96, ...eventually you should come back to your starting number, at which point, you're done.

**SIMULATION:** Another use for random digit tables is simulation—modeling experiments or activities. Let's illustrate the idea with an example.

**EXAMPLE:**

[3.] How often does a family of three children have three boys? Let's answer this with a simulation that assumes boys and girls are equally likely (they aren't, really...).

Let's read a single digit from the table, letting an even number represent a girl, and an odd number representing a boy. Every group of three digits represents a family with three children. Here goes:

509|64 6|099|6 41|960| 814|66 7|261|0 58|137| 373|04 8|215|

The first family has 2 boys; the second has none; the third has two; keep going. Of the thirteen "families," two have all boys. So, our simulation resulted in two of thirteen (about 15%) families with all boys.

### ***1.3 Introduction to Experimental Design***

**CENSUS VS. SAMPLE:** A **Census is a technique where every individual in a population is measured**. As was mentioned earlier, this is rarely possible. Instead, a sample is taken. We hope (well, we take steps to ensure) that our sample is representative, so that a description of the sample can be used to make inferences about the population.

**OBSERVATIONAL STUDY VS. EXPERIMENT:** There are basically two ways (reasons) to collect data. An **Observational Study only observes**—there is no attempt to change anything about the

sample. **An Experiment changes something**—a treatment is applied to the individuals in the sample.

For example, if we want to know whether or not SAT preparation courses actually help students increase their SAT scores, we might just go and find some students who had taken such a course, and some others that had not, and look at their scores. This would be an observational study—nothing was changed. On the other hand, we might take a group of students, randomly select some of them to take an SAT prep course, and then measure SAT scores. This would be an experiment, since something was changed (we changed whether the students had taken a prep course).

An experiment is the better choice. Unfortunately, it isn't always possible. For example, if we want to see if smoking causes cancer, we can't force some people to smoke just to see if they develop cancer!

When we conduct an experiment, we take a group of individuals (a sample), and apply a treatment (some kind of change) to *some* of the individuals, in order to see if there is a corresponding change in some variable (the measurement). **The individuals that are not changed form the Control Group. The individuals that are changed for the Treatment Group.**

In order to know if our change actually caused the change in the variable, the individuals that are in the Treatment Group need to be almost identical to those in the Control Group—ideally, we want the only difference between the two groups to be the change that we induce. Thus, the experimenter must apply **Control** to other things that might cause differences in the two groups. There are lots of potential differences that can crop up—you could take a course (or two) on just that. We will focus on just a few of the big ones.

Imagine that we are testing a new pain medication. We give everyone the new medication, and measure the degree of pain relief. Do we really know if our new medication did the trick? Is it possible that our individuals experienced pain relief simply because we gave them a pill? Is it possible that our individuals would have experienced relief even if we had given them a sugar pill (which should have no effect whatsoever).

Yes!

**THE PLACEBO EFFECT: A change in the variable that is caused by the individuals' knowledge that they are part of an experiment.** Psychology! I give you a pill that (I say) should make you feel better. Often, you'll get better, even if the pill I gave you shouldn't have done anything. That's the placebo effect. **A Placebo is a fake treatment**—something which should not affect the variable that is being measured. Although the word placebo originally referred to a sugar pill, it now refers to *anything* that should have no effect.

Now, imagine that we have given our subjects their pills, and now we're asking them how much pain relief they've felt. Is it possible that we might (subconsciously) interpret their responses differently, based on whether they received the real medication or the placebo?

Yes!

**DOUBLE-BLIND EXPERIMENT:** An experiment where neither the individual receiving the treatment, nor the person measuring the variable, know whether the individual is in the Control Group or the Treatment Group. Failure to make an experiment double-blind may introduce hidden bias—we may do, say, suggest or interpret things differently because of our knowledge of the experiment.

**LURKING AND CONFOUNDING VARIABLES:** As was mentioned earlier, we want the individuals in the Treatment and Control Groups to be as alike as possible, so that any differences in our measurements can be attributed to the change (treatment) that we introduced. **Any other variable that might have an effect on our experiment is called a Lurking Variable.** The book suggests that Lurking and Confounding are the same; they are not, in fact—but the difference is subtle, and beyond our needs. In any case, we must control lurking variables—we must make the Control and Treatment Groups as alike as possible.

**RANDOMIZED EXPERIMENT:** There will always be lurking variables that are unknown to us—no one can think of everything! In order to deal with unknown lurking variables, we resort to randomization—we randomly assign individuals to either the Control or Treatment Groups. **A Randomized Experiment is one where individuals are randomly assigned to these groups.**

EXAMPLE:

[4.] Let's suppose that we want to find out if coffee grounds help make plants grow better (yes, some people claim that this is true!). Furthermore, let's suppose that we have 16 identical plants and a nice plot of ground to plant them in. How can we design an experiment to determine if the coffee grounds make the plants grow better?

Perhaps we would first randomly assign each plant to a part of our plot. A table of random digits can be used for this.

Next, we should randomly assign some of the plants (half is best) to receive the treatment—in this case, we'd want to pick eight that will get the coffee grounds. Again, a table of random digits is a great way to go here.

OK, lay out those coffee grounds. Now, everything else that we do to these plants should be the same—they should all get the same amount of water, or bug spray, or plant food...whatever one gets, they all get.

After some suitable amount of time, we can go and measure the height of each plant. Ideally, the person who takes the measurements should not know which plants got the coffee grounds (so that the experiment is double-blind).

We've talked quite a bit about experiments—and deservedly so, since they can give us answers that studies cannot. However, that does not mean that studies can be ignored. Observational studies are still conducted, and must be conducted properly in order to achieve the desired (accurate) results. The biggest issue with observational studies is when we survey people. Thus, we must look at sources of bias in surveys.

**BIAS:** A systematic difference between the population and the sample. If I conduct an Internet poll about how much people like using the Internet, what kinds of responses do you think I'm going to get? If I ask people at a high school basketball game about their favorite sport,

what sort of answers do you think they'd give? In both cases, the problem is in how I selected the sample. Of course, that's not the only problem.

NONRESPONSE BIAS: **This occurs when the people that I have selected refuse to answer the questions.** This is a big problem for telephone surveys.

VOLUNTARY RESPONSE BIAS: **This occurs when the people select themselves.** For example, call-in polls—you've got to decide to participate in order to answer the question(s).

And there are more. Many more. We've just touched the tip of the iceberg—opened your mind to possibilities.