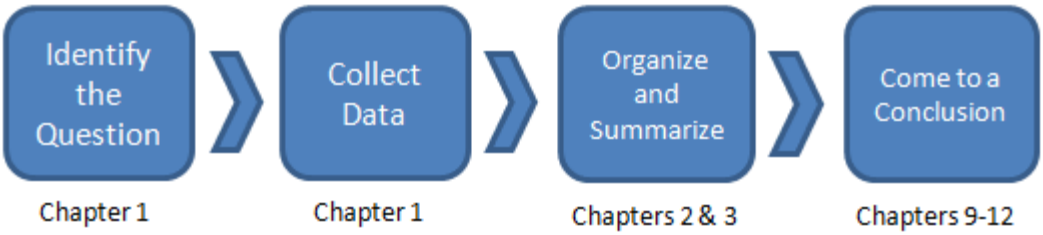
Chapter 1: Data Collection

Section 1.1
Section 1.2
Section 1.3
Section 1.4
Section 1.5
Section 1.6



Statistics is a process. The first step in that process is collecting data, which is what we'll focus on in this first chapter.

In Section 1.1, we introduce the process of statistics, and the different ways to classify data. In Section 1.2, we talk about different sources for data, and we introduce a couple of very important ones: observational studies and designed experiments. We'll introduce one way to select a random sample from a population in Section 1.3, and Section 1.4 will introduce several more.

In Section 1.5, various sources for error in sampling are discussed, while Section 1.6 covers how to design a statistical experiment.

By the end of this section, you should have a good understanding of the process of statistics, how to select a random sample from a population, how to avoid errors in that sample, and how to design statistical experiments. If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

Section 1.1: Introduction to the Practice of Statistics

1.1 Introduction to the Practice of Statistics
1.2 Observational Studies versus Designed Experiments
1.3 Simple Random Sampling
1.4 Other Effective Sampling Methods
1.5 Bias in Sampling
1.6 The Design of Experiments

Objectives

By the end of this lesson, you will be able to...

1. describe what "statistics" means in the context of this course.
2. explain the process of statistics
3. distinguish between qualitative and quantitative variables
4. distinguish between discrete and continuous variables
5. determine the level of measurement of a variable

The first thing we want to look at is exactly what "statistics" is. It should come as no surprise that your textbook has a definition for statistics. Here's what the author writes:

> **Statistics** is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions.

So what does that mean? Well, if you haven't already, you really need to read your textbook on this section. There's some great information on pages 3 & 4 about the study of statistics.

See, the reason we use data is that often the anecdotal information we have which might *appear* to be true actually is not.

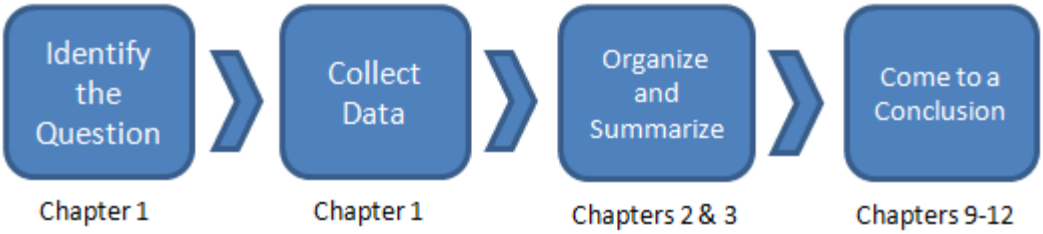
Case in point: Recently the math department at ECC decided to do some investigation into the success of students in the College Algebra course. Many instructors had poor experiences with students who placed into the course with an ACT score of 23. Those faculty members felt that the cut-off score for placement into College Algebra should be increased to at least 24.

An interesting thing happened when data was collected from a year's worth of students - the students with a 23 on their ACT did just as well as students who placed their via ECC's own placement exam or those who took Intermediate Algebra first. Oops! This was a reminder to even those of us in the math department that there's a reason why the study of statistics was developed - we often have a skewed sense of reality when we only trust our experiences.

> At this point, I'd strongly recommend beginning a list of terms with definitions. You might start to get overwhelmed with all the terminology, so a list of terms to refer to would be very helpful.

The Process of Statistics

So what exactly is the study of statistics? Well, it's really a process. Your textbook has a good summary of it, but I've included a bit of a visual here as well.



First, we must identify exactly what it is we're hoping to study. We must also determine what our population is.	Next, we select a representative sample using appropriate sampling techniques.	Once we have our data collected, we have to summarize it. We'll do this both numerically and visually with charts.	Finally, we need to analyze it and come to a conclusion.

The gaps in the middle - Chapters 4-8 - are a mix of sections. Chapter 4 really can stand on its own. It's all about analyzing the relationship between two variables. Chapters 5-8 involve probability and are intended as preparation for the meat of the course in Chapters 9-12.

Identifying the Question

A couple of key comments about identifying the question are needed here. The first thing we really need to consider is what our population is. The **population** is the group we're studying.

For example, if I'm interested in the studying habits of ECC students, then my population is *all ECC students*. Since asking every ECC student isn't possible, I would then take a **sample**, which is a subset of the population. The characteristics of the sample are key. If we select too few or the individuals selected don't represent the population, any conclusions we draw will be meaningless.

> A **statistic** is a numerical summary of a sample. By contrast, a numerical summary of a *population* is called a **parameter**.

For example, if we know from ECC data that the average age of all ECC students is 29 (ECC College Facts), that value is a *parameter*. On the other hand, if we take a sample of 100 students and find that 63% support a new initiative at the college, that is a *statistic* - since it is only a measure of the sample of 100 students, not the entire student population.

> When we simply describe or summarize data, we're using **descriptive statistics**. When we draw conclusions or extend our results to the population, we're using **inferential statistics**.

For example, the statistics of 63% from above would be a *descriptive statistic*, since it is simply a summary of our sample. If we, in turn, make a broad generalization and claim that 63% of *all* ECC students support the initiative, then that is *inferential statistics*.

Qualitative or Quantitative

In general, we classify data into two groups: *qualitative* or *quantitative*. Of course, your textbook has definitions for both:

> **Qualitative (or categorical) variables** allow for classification of individuals based on some attribute or characteristic.
>
> **Quantitative variables** provide numerical measures of individuals. Arithmetic operations such as

addition and subtraction can be performed on the values of a quantitative variable and will provide meaningful results.

Basically, if a variable describes a *quality* of an individual - i.e. hair color, political party, etc - then it is *qualitative*. If a variable is numerical **and those numbers have meaning**, then it is *quantitative*. (Not all numbers have meaning numerically - think of an individuals Social Security number.)

Example 1

So, which are they? Here are some examples of data that might be collected. Take a minute and make a note of whether each is qualitative or quantitative. When you're ready, check your answer below.

gender, IQ, ACT score, eye color, area code

[reveal answer]

Discrete or Continuous

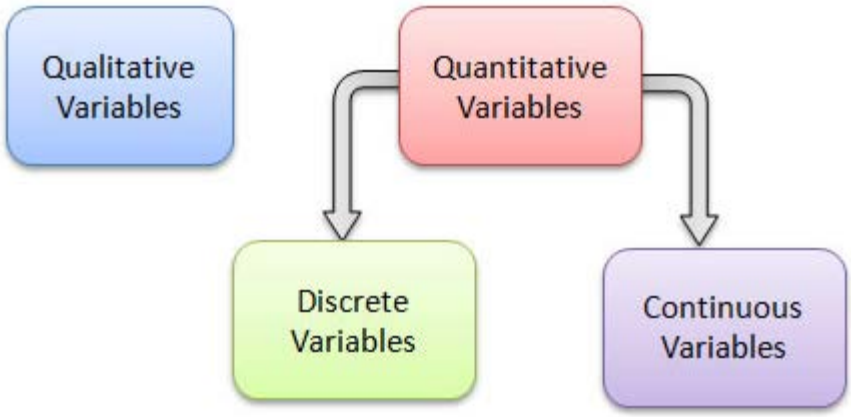
Quantitative variables can be further split into two groups.

A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of values. (*Countable* means that the values result from counting - 0, 1, 2, 3, ...)

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable.

Most variables are pretty clear, but some can be a bit tricky. An example of a tricky one is time. Say, for example, we're looking at how long we've been waiting for a bus. We *count* the minutes and seconds, but really those time units are only rounded. There are actually milliseconds, nanoseconds, etc - an infinite number of possibilities in the middle. So actually, any variable that is time is continuous.

Here's a graphical representation of the different ways to classify variables:



Example 2

Time for some examples. Take a minute and make a note of whether each quantitative variable is discrete or continuous. When you're ready, check your answer below.

IQ, ACT score, height, distance commuting, shoe size

[reveal answer]

Section 1.2: Observational Studies versus Designed Experiments

1.1 Introduction to the Practice of Statistics
1.2 Observational Studies versus Designed Experiments
1.3 Simple Random Sampling
1.4 Other Effective Sampling Methods
1.5 Bias in Sampling
1.6 The Design of Experiments

Objectives

By the end of this lesson, you will be able to...

1. distinguish between an observational study and a designed experiment
2. identify possible lurking variables
3. explain the various types of observational studies

To begin, we're going to discuss some of the ways to collect data. In general, there are a few standards:

- census
- existing sources
- survey sampling
- designed experiments

Most of us associate the word *census* with the U.S. Census, but it actually has a broader definition. Here's what your text defines it as:

> A **census** is a list of all individuals in a population along with certain characteristics of each individual.

The nice part about a census is that it gives us all the information we want. Of course, it's usually impossible to get - imagine trying to interview *every single ECC student*. That'd be over 10,000 interviews!

So if we can't get a census, what do we do? A great source of data is other studies that have already been completed. If you're trying to answer a particular question, look to see if someone else has already collected data about that population. As your textbook says, the moral of the story is this: **Don't collect data that have already been collected!**

Observational Studies versus Designed Experiments

Now to one of the main objectives for this section. Two other very common sources of data are **observational studies** and **designed experiments**. We're going to take some time here to describe them and distinguish between them - you'll be expected to be able to do the same in homework and on your first exam.

The easiest examples of observational studies are surveys. No attempt is made to influence anything - just ask questions and record the responses. By definition,

> An **observational study** measures the characteristics of a population by studying individuals in a sample, but does not attempt to manipulate or influence the variables of interest.

For a good example, try visiting the Pew Research Center. Just click on any article and you'll see an example of an observational study. They just sample a particular group and ask them questions.

In contrast, *designed experiments* explicitly do attempt to influence results. They try to determine what affect a particular treatment has on an outcome.

A **designed experiment** applies a treatment to individuals (referred to as **experimental units** or **subjects**) and attempts to isolate the effects of the treatment on a **response variable**.

For a nice example of a designed experiment, check out this article from National Public Radio about the effect of exercise on fitness.

So let's look at a couple examples.

Example 1

Visit this link from Science Daily, from July 8th, 2008. It talks about the relationship between Post-Traumatic Stress Disorder (PTSD) and heart disease. After reading the article carefully, try to decide whether it was an *observational study* or a *designed experiment*

What was it?

Example 2

Visit this link from the Gallup Organization, from June 17th, 2008. It looks at what Americans' top concerns were at that point. Read carefully and think of the how the data were collected. Do you think this was an observational study or a designed experiment? Why?

Think carefully about which you think it was, and just as important - why? When you're ready, click the link below.

What was it?

Example 3

This last example is regarding the "low-carb" Atkins diet, and how it compares with other diets. Read through this summary of a report in the New England Journal of Medicine and see if you can figure out whether it's an observational study or a designed experiment.

What was it?

Probably the biggest difference between observational studies and designed experiments is the issue of *association* versus *causation*. Since observational studies don't control any variables, the results can only be *associations*. Because variables are controlled in a designed experiment, we can have conclusions of *causation*.

Look back over the three examples linked above and see if all three reported their results correctly. You'll often find articles in newspapers or online claiming one variable *caused* a certain response in another, when really all they had was an *association* from doing an observational study.

The discussion of the differences between observational studies and designed experiments may bring up an interesting question - why are we worried so much about the difference?

We already mentioned the key at the end of the previous page, but it bears repeating here:

Observational studies only allow us to claim *association*,not *causation*.

The primary reason behind this is something called a *lurking variable* (sometimes also termed a *confounding factor*, among other similar terms).

A **lurking variable** is a variable that affects both of the variables of interest, but is either not known or is not acknowledged.

Consider the following example, from The Washington Post:

Example 4 **Coffee may have health benefits and may not pose health risks for many people**

By Carolyn Butler Tuesday, December 22, 2009

Of all the relationships in my life, by far the most on-again, off-again has been with coffee: From that initial, tentative dalliance in college to a serious commitment during my first real reporting job to breaking up altogether when I got pregnant, only to fail miserably at quitting my daily latte the second time I was expecting. More recently the relationship has turned into full-blown obsession and, ironically, I often fall asleep at night dreaming of the delicious, satisfying cup of joe that awaits, come morning.

[...] Rest assured: Not only has current research shown that moderate coffee consumption isn't likely to hurt you, it may actually have significant health benefits. **"Coffee is generally associated with a less health-conscious lifestyle -- people who don't sleep much, drink coffee, smoke, drink alcohol,"** explains Rob van Dam, an assistant professor in the departments of nutrition and epidemiology at the Harvard School of Public Health. **He points out that early studies failed to account for such issues and thus found a link between drinking coffee and such conditions as heart disease and cancer, a link that has contributed to java's lingering bad rep.** "But as more studies have been conducted-- larger and better studies that controlled for healthy lifestyle issues --the totality of efforts suggests that coffee is a good beverage choice."

[...]

Source: Washington Post

What is this article telling us? If you look at the parts in bold, you can see that Professor van Dam is describing a lurking variable: lifestyle. In past studies, this variable wasn't accounted for. Researchers in the past saw the relationship between coffee and heart disease, and came to the conclusion that the coffee was *causing* the heart disease.

But since those were only observational studies, the researchers could only claim an *association*. In that example, the lifestyle choices of individuals was affecting both their coffee use *and* other risks leading to heart disease. So "lifestyle" would be an example of a lurking variable in that example.

For more on lurking variables, check out this link from The Math Forum and this one from The Psychology Wiki. Both give further examples and illustrations.

With all the problems of lurking variables, there are many good reasons to do an observational study. For one, a designed experiment may be impractical or even unethical (imagine a designed experiment regarding the risks of smoking). Observational studies also tend to cost much less than designed experiments, and it's often possible to

obtain a much larger data set than you would with a designed experiment. Still, it's always important to remember the difference in what we can claim as a result of observational studies versus designed experiments.

Types of Observational Studies

There are three major types of observational studies, and they're listed in your text: cross-sectional studies, case-control studies, and cohort studies. Your textbook does a good job describing each, but we'll summarize them again here and give a couple quick examples of each.

Cross-sectional Studies

This first type of observational study involves collecting data about individuals at a certain point in time. A researcher concerned about the effect of working with asbestos might compare the cancer rate of those who work with asbestos versus those who do not.

Cross-sectional studies are cheap and easy to do, but they don't give very strong results. In our quick example, we can't be sure that those working with asbestos who don't report cancer won't eventually develop it. This type of study only gives a bit of the picture, so it is rarely used by itself. Researchers tend to use a cross-sectional study to first determine if their might be a link, and then later do another study (like one of the following) to further investigate.

Case-control Studies

Case-control studies are frequently used in the medical community to compare individuals with a particular characteristic (this group is the *case*)with individuals who do not have that characteristic (this group is the *control*). Researchers attempt to select homogeneous groups, so that on average, all other characteristics of the individuals will be similar, with only the characteristic in question differing.

One of the most famous examples of this type of study is the early research on the link between smoking and lung cancer in the United Kingdom by Richard Doll and A. Bradford Hill. In the 1950's, almost 80% of adults in the UK were smokers, and the connection between smoking and lung cancer had not yet been established. Doll and Hill interviewed about 700 lung cancer patients to try to determine a possible cause.

This type of study is ***retrospective***,because it asks the individuals to look back and describe their habits(regarding smoking, in this case). There are clear weaknesses in a study like this, because it expects individuals to not only have an accurate memory, but also to respond honestly. (Think about a study concerning drug use and cognitive impairment.) Not only that, we discussed previously that such a study may prove ***association***, but it cannot prove **causation**.

Cohort Studies

A cohort describes a group of individuals, and so a cohort study is one in which a group of individuals is selected to participate in a study. The group is then observed over a period of time to determine if particular characteristics affect a response variable.

Based on their earlier research, Doll and Hill began one of the largest cohort studies in 1951. The study was again regarding the link between smoking and lung cancer. The study began with 34,439 male British doctors, and followed them for over 50 years. Doll and Hill first reported findings in 1954 in the British Medical Journal, and then continued to report their findings periodically afterward. Their last report was in 2004,again published in the British Medical Journal. This last report reflected on 50 years of observational data from the cohort.

This last type of study is called **prospective**, because it begins with the group and then collects data over time. As your textbook mentions,cohort studies are definitely the most powerful of the observational studies,particularly with the quantity and quality of data in a study like the previous one.

Let's look at some examples.

Example 4

A recent article in the BBC News Health section described a study concerning dementia and "mid-life ills". According to the article, researches followed more than 11,000 people over a period of 12-14 years. They found that smoking, diabetes, and high blood pressure were all factors in the onset of dementia.

What type of observational study was this? Cross-sectional, case-control,or cohort?

What was it?

Example 5

In 1993, the National Institute of Environmental Health Sciences funded a study in Iowa regarding the possible relationship between radon levels and the incidence of cancer. The study gathered information from 413 participants who had developed lung cancer and compared those results with 614 participants who did not have lung cancer.

What type of study was this?

What was it?

Example 6

In 2004, researchers published an article in the New England Journal of Medicine regarding the relationship between the mental health of soldiers exposed to combat stress. The study collected information from soldiers in four combat infantry units either before their deployment to Iraq or three to four months after their return from combat duty.

What type of study was this?

What was it?

Section 1.3: Simple Random Sampling

1.1 Introduction to the Practice of Statistics
1.2 Observational Studies versus Designed Experiments
1.3 Simple Random Sampling
1.4 Other Effective Sampling Methods
1.5 Bias in Sampling
1.6 The Design of Experiments

Objectives

By the end of this lesson, you will be able to...

1. obtain a simple random sample

The next section we want to discuss is how to pick a "random" sample from a population. Even more-so - what does it mean to be "random"?

Why do we sample?

Let's suppose we want to know what ECC students think about parking on campus. It isn't possible to ask every single student, so instead we try to get a sample of students. One important characteristic that this sample must have is that it must be representative of the entire student body. (In other words, we can't have all Culinary Arts students, or all students that are fresh from high school.)

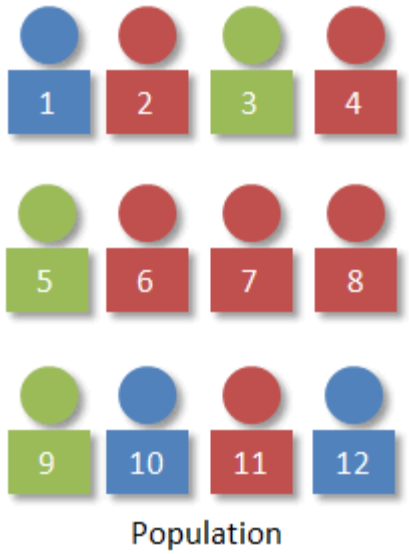
In this section and Section 1.4, we'll introduce several sampling strategies: simple random, stratified, systematic, and cluster.

Simple Random Sampling

The first type of sampling, called *simple random sampling*, is the simplest. Here's the textbook definition:

> A sample of size *n* from a population of size *N* is obtained through **simple random sampling** if every possible sample of size *n* has an equally likely chance of occurring.

OK, so maybe that didn't sound simple. Essentially, in order to qualify as a *simple random sampling* process, each sample must be equally likely. You've probably already used this method without knowing it.

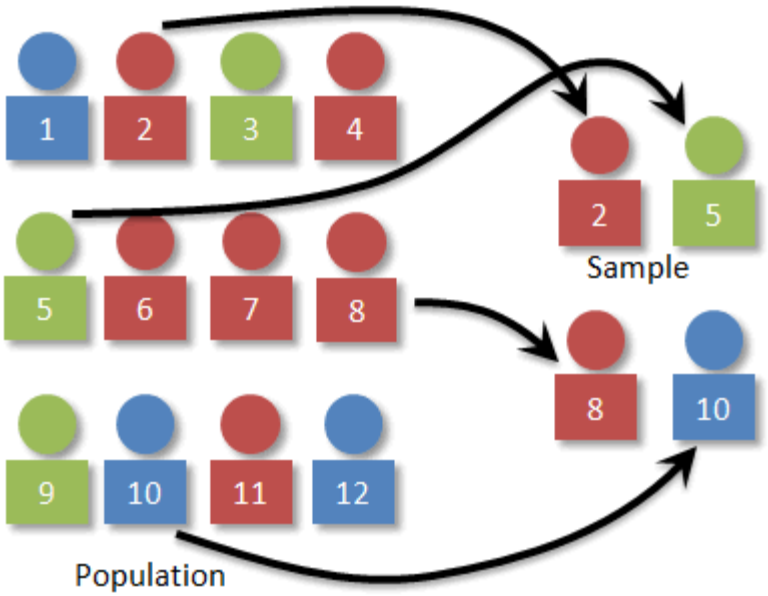


Let's suppose you want to select a sample of 4 people from a group of 12 (see image above). Here are some common ways to select a simple random sample:

- write everyone's name on a slip of paper and draw two from a hat
- write all possible samples of size two on slips of paper and draw one from a hat
- number each individual and use technology to randomly select two integers between 1 and 30

Practically, the first two lost their effectiveness with large groups, so we'll be focusing on the latter method.

With our example of a sample size 4 from a population of 12, we might use technology to select four random integers between 1 and 12. Say we get 2, 5, 8, and 10. Our sample would then look this this:



For another take, watch this YouTube video:

Random

The only thing left to do, then, is to generate a random number. But how do you do that? Just pick a number from your head?

For a good explanation, watch this video from Clive Rix, at the University of Leicester in England:

OK, then how do we actually generate a random number? The "Technology" box below shows how to generate what are called "pseudo random numbers", which is a reasonable enough technique for this course.

To get a true random number, you need something more sophisticated. One solution is random.org. For information about randomness and the difference between *pseudo* random numbers and *true* random numbers, you can visit their page on an Introduction to Randomness and Random Numbers.

For the purposes of this course, feel free to use the instructions below.

Technology

Here's a quick overview of how to generate random integers in StatCrunch.

1. Select Data > Simulate Data > Uniform
2. Enter n for Rows and 1 for Columns
3. Enter the lower and upper limits for a and b.
4. Press Simulate

You can manually round each value, or StatCrunch can do it for you. To round, follow these steps:

1. Select Data > Compute expression
2. Set Y to Uniform1.
3. Select "round(Y)" in the expression dropbox (it's the very last expression).
4. Press Set Expression and press Compute.

Section 1.4: Other Effective Sampling Methods

1.1 Introduction to the Practice of Statistics
1.2 Observational Studies versus Designed Experiments
1.3 Simple Random Sampling
1.4 Other Effective Sampling Methods
1.5 Bias in Sampling
1.6 The Design of Experiments

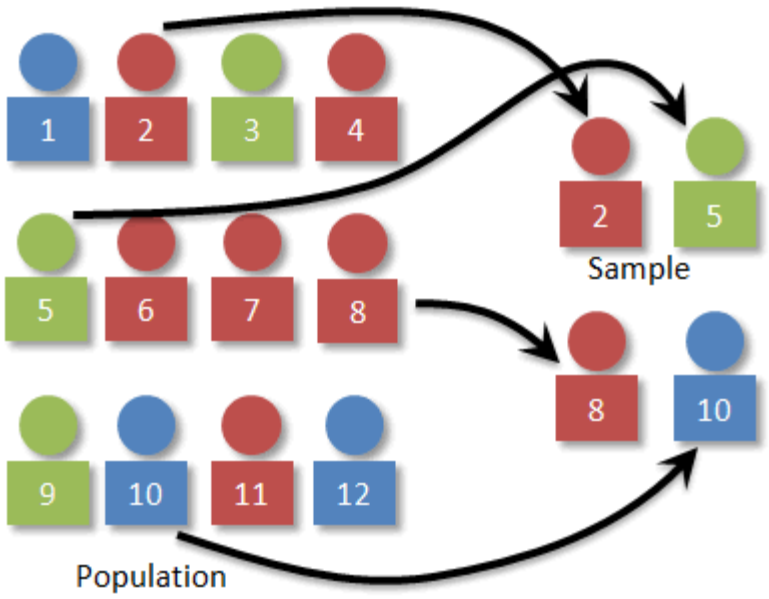
Objectives

By the end of this lesson, you will be able to...

1. describe the difference between the stratified, systematic, and cluster sampling techniques
2. identify which sampling technique was used
3. determine an appropriate sampling technique given a situation
4. obtain a stratified, systematic, or cluster sample

Review: Simple Random Sampling

Do you remember how simple random sampling works? Visually, it's just numbering each individual and randomly selecting a certain number of them. Here's the image we used in the previous section:

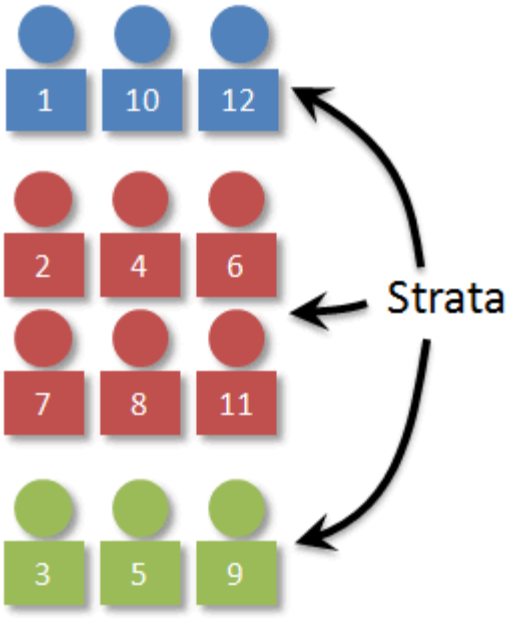


Stratified Sampling

Stratified sampling is different. With this technique, we separate the population using some characteristic, and then take a proportional random sample from each.

A **stratified sample** is obtained by separating the population into non-overlapping groups called *strata* and then obtaining a proportional simple random sample from each group. The individuals within each group should be similar in some way.

Visually, it might look something like the image below. With our population, we can easily separate the individuals by color.



Once we have the strata determined, we need to decide how many individuals to select from each stratum. (Man, that's a weird word!) The key here is that the number selected should be *proportional*. In our case, 1/4 of the individuals in the population are blue, so 1/4 of the sample should be blue as well. Working things out, we can see that a stratified (by color) random sample of 4 should have 1 blue, 1 green, and 2 reds.

For another take, watch this YouTube video:

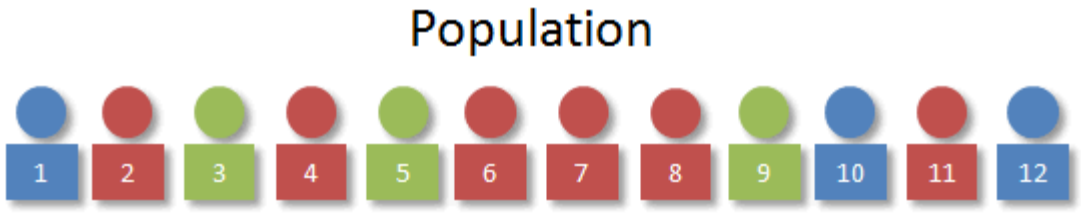
Example 1

One easy example using a stratified technique would be a sampling of people at ECC. To make sure that a sufficient number of students, faculty, and staff are selected, we would stratify all individuals by their status - students, faculty, or staff. (These are the *strata*.) Then, a proportional number of individuals would be selected from each group.

Systematic Sampling

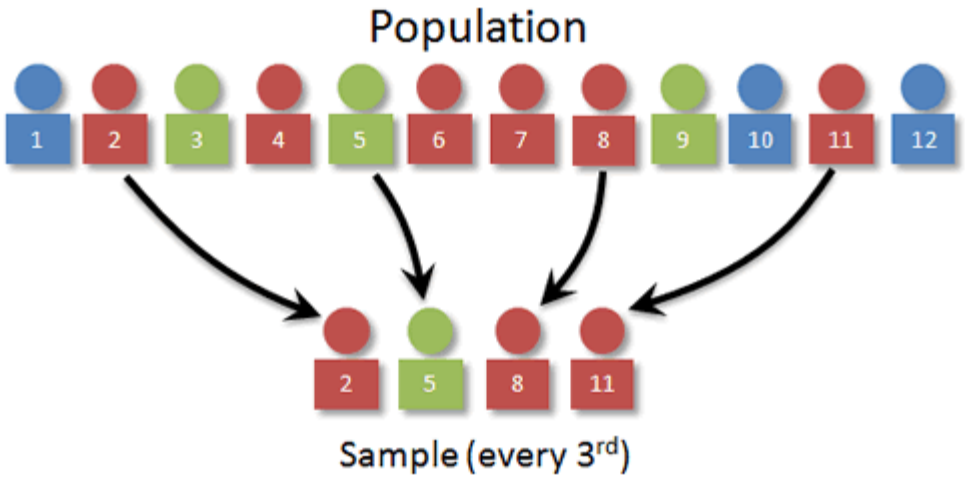
A **systematic sample** is obtained by selecting every *k*th individual from the population. The first individual selected corresponds to a random number between 1 and *k*.

So to use systematic sampling, we need to first order our individuals, then select every *k*th. (More on how to select *k* in a bit.)



In our example, we want to use 3 for *k*? Can you see why? Think what would happen if we used 2 or 4.

For our starting point, we pick a random number between 1 and *k*. For our visual, let's suppose that we pick 2. The individuals sampled would then be 2, 5, 8, and 11.



In general we find k by taking N/n and rounding down to the nearest integer.

For another take, watch this YouTube video:

Example 2

Systematic sampling works well when the individuals are already lined up in order. In the past, students have often used this method when asked to survey a random sample of ECC students. Since we don't have access to the complete list, just stand at a corner and pick every 10th* person walking by.

* Of course, choosing 10 here is just an example. It would depend on the number of students typically passing by that spot and what sample size was needed.

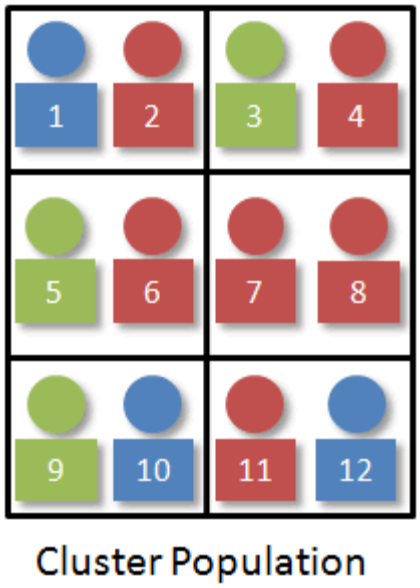
Cluster Sampling

Cluster sampling is often confused with stratified sampling, because they both involve "groups". In reality, they're very different. In stratified sampling, we split the population up into groups (strata) based on some characteristic.

> A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.

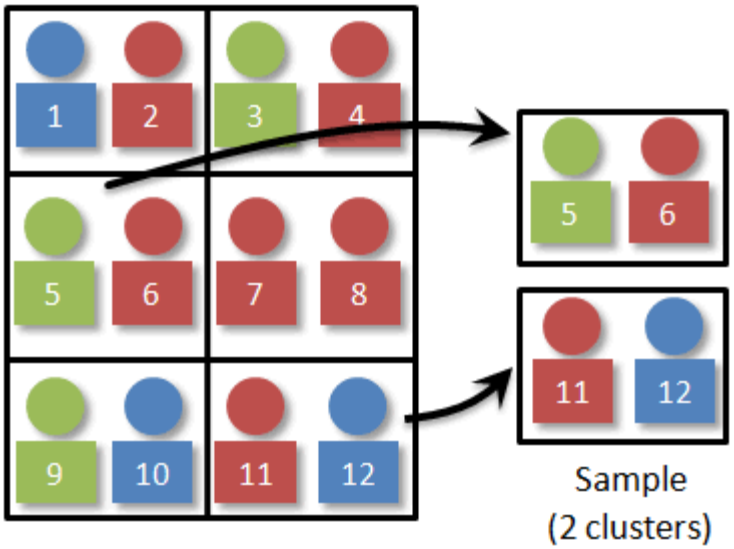
In essence, we use cluster sampling when our population is already broken up into groups (*clusters*), and each cluster represents the population. That way, we just select a certain number of clusters.

With our visual, let's suppose the 12 individuals are paired up just as they were sitting in the original population.



Cluster Population

Since we want a random sample of size four, we just select two of the clusters. We would number the clusters 1-6 and use technology to randomly select two random numbers. It might look something like this:



Cluster Population

Sample
(2 clusters)

For another take, watch this YouTube video:

Example 3

One situation where cluster sampling would apply might be in manufacturing. Suppose your company makes light bulbs, and you'd like to test the effectiveness of the packaging. You don't have a complete list, so simple random sampling doesn't apply, and the bulbs are already in boxes, so you can't order them to use systematic. And all the bulbs are essentially the same, so there aren't any characteristics with which to stratify them.

To use cluster sampling, a quality control inspector might select a certain number of entire boxes of bulbs and test each bulb within those boxes. In this case, the boxes are the *clusters*.

Convenience Sampling

Other methods do exist for finding samples of populations. In fact, you've seen some already. Probably the most common is the so-called **convenience sample**. Convenience samples are just what they sound like - convenient. Unfortunately, they're rarely representative. Think of the radio call-in show, those people in the shopping malls trying to survey you about your purchasing habits, or even the voting on American Idol!

Here's a specific example. It's a poll on beliefnet.com, titled "What Evangelicals Want". All online polls use, by nature, convenience sampling. According to the article, "The poll was promoted on Beliefnet's web site and through its newsletters." Only those evangelicals who visit this particular web site and actually answer the survey are included. Beware any poll result taken with convenience sampling.

Multistage Sampling

Often one technique isn't possible, so many professional polling agencies use a technique called **multistage sampling**. The strategy is relatively self-explanatory - two or more sampling techniques are used.

For example, consider the light-bulb example we looked at earlier with cluster sampling. Let's suppose that the bulbs come off the assembly line in boxes that each contain 20 packages of four bulbs each. One strategy would be to do the sample in two stages:

Stage 1: A quality control engineer removes every 200th box coming off the line. (The plant produces 5,000

boxes daily. (This is *systematic* sampling.)

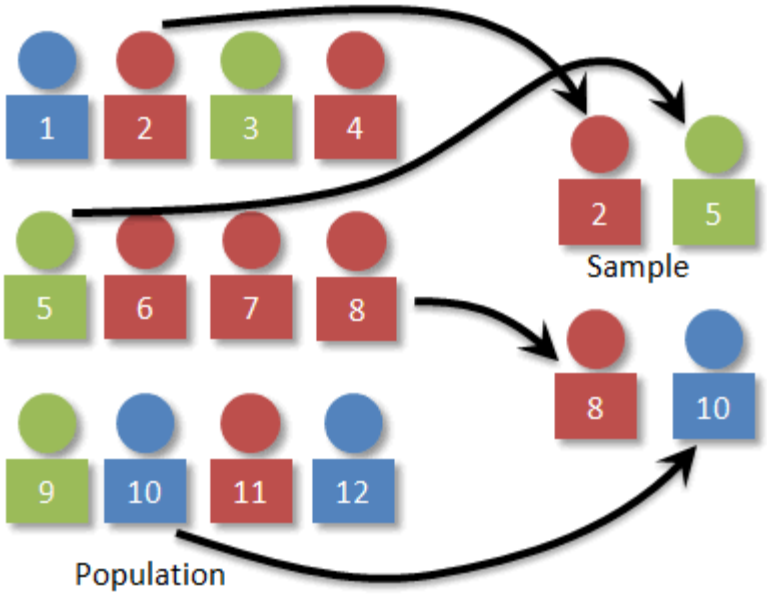
Stage 2: From each box, the engineer then samples three packages to inspect. (This is an example of **cluster** sampling.)

The US Census also uses multistage sampling. If you haven't already (you should have!), read Section 1.4 in your text for more details.

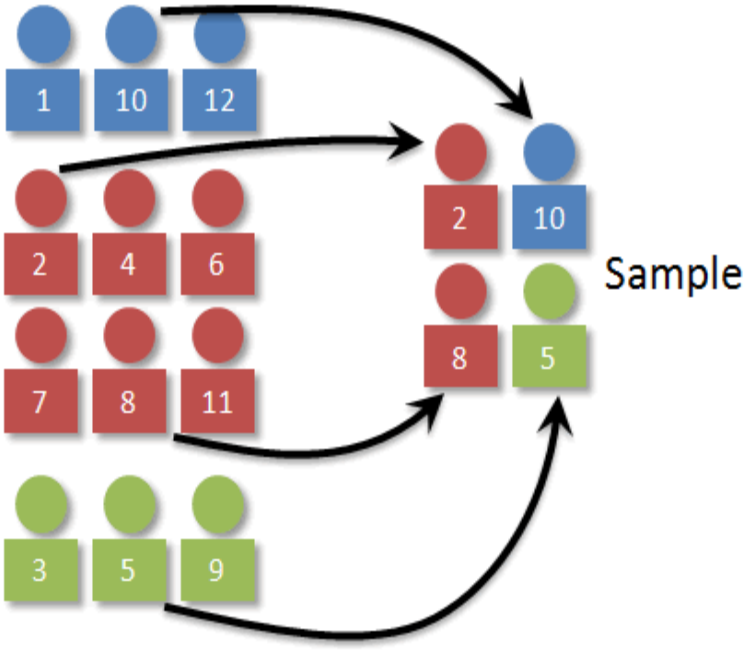
Summary

Here's a visual summary of the four main sampling strategies:

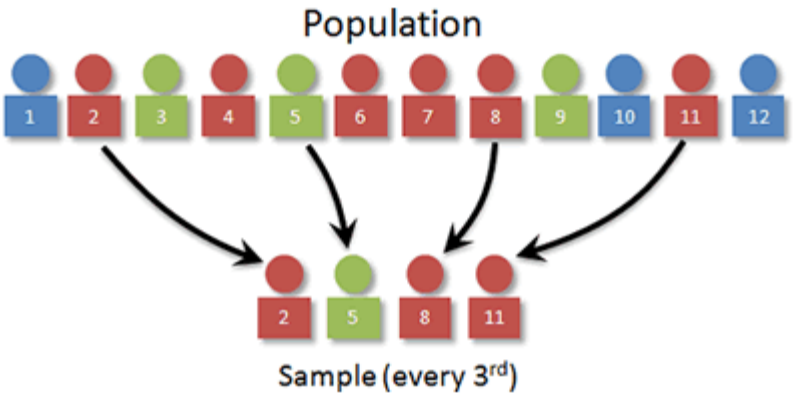
Simple Random:



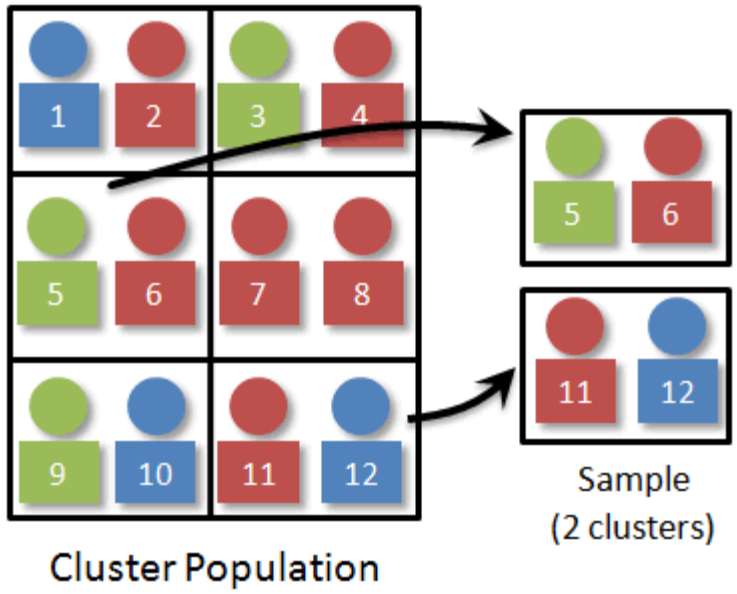
Stratified:



Systematic:



Cluster:



<< previous section | next section >>

1 2 3 4 5 6 7 8 9 10 11 12 13

Section 1.5: Sources of Errors in Sampling

1.1 Introduction to the Practice of Statistics
1.2 Observational Studies versus Designed Experiments
1.3 Simple Random Sampling
1.4 Other Effective Sampling Methods
1.5 Bias in Sampling
1.6 The Design of Experiments

Objectives

By the end of this lesson, you will be able to...

1. understand how error can be introduced during sampling
2. identify which errors have been made given an example

In general, there are two types of errors that can result during sampling.

> **Nonsampling errors** are errors that result from the survey process.

Examples of nonsampling errors might be nonresponses of individuals selected to be in the survey, inaccurate responses, poorly worded questions, poor interviewing technique, etc.

> **Sampling error** is the error that results from using a sample to estimate information regarding a population.

There's really nothing we can do about this second type. Unless we sample every single individual in the sample, there will be some error in our results. Much later in the course, we'll talk about how we can actually get an estimate for how close we are to the true population information we're trying to get at.

Since we can't control the sampling error, we'll focus in this section on the different types of nonsampling errors. There aren't a lot of graphical ways to represent this material, and I don't want to just repeat what's already in your text (pages 14-16), so I'll just summarize each source of error here.

The Frame

As your text says, surveys of voters or even of ECC students require a *complete* list of all the individuals. If an individual isn't on the list, any sample taken won't be representative. A common example of this is surveys over the phone - think of the types of people who either don't have land lines, have caller ID, or maybe change phones so often that they're not on the list. Any survey done via the telephone is clearly suspect. Unfortunately, it's often the only practical option for pollsters.

Nonresponse

Any survey will always have a portion of those sampled who simply don't respond. At ECC, we do an annual employee satisfaction survey. The people in the Institutional Research office are ecstatic with a 40% response rate.

Check out this link from the SuperSurvey Knowledge Base with a more detailed description of some reasons.

Interviewer Error

Have you seen the movie, Kinsey? It's a movie based on the life of Dr. Alfred Kinsey, who formed the Kinsey

Institute, which published the Kinsey Reports about the sexual behaviors of men and women. During research. Kinsey and his colleagues performed countless in-person interviews. Imagine what a difference the quality of the interviewer would make in a context like that!

Misrepresented Answers

A classic example here is a survey I've done in my developmental classes about how often students study. Because of the nature of the variable in question, it has to be self-reported, but many students misjudge or even lie about how much they're really studying.

Data Checks

There's nothing like finishing your research about how many children the typical family has and finding that outlier - 45! Chances are, it was most likely an incorrectly entered 4 or 5, but it may be too late at that point to find out. As your textbook states, "It is imperative that data be checked for accuracy at every stage of the statistical analysis."

Next we'll focus on specifics regarding the design of questionnaires.

Types of Questions

In general, there are two types of survey questions - **open** and **closed**.

An example of an **open question** might be:	An example of a **closed** version of the same question would be this:
What issue is most important to you in determining which political candidate to support?	What issue is most important to you in determining which political candidate to support? a. the economy b. the war in Iraq c. health care d. immigration e. education

Each design has its own limitations - the open question makes compiling the data difficult, while the closed question limits the responses. A good compromise is to first give a "presurvey" with open questions, and then use the most common responses from that survey to form the actual survey with closed questions.

Wording and Ordering of Questions

Your textbook has quite a good summary of some of the issues, so I'll just give a few good links.

- Abortion, the Court and the Public, from the Pew Research Center
- Why Question Order Changes Poll Results, from CBS News
- Question Wording, information from the American Association for Public Opinion Research
- Conducting a Survey In Your Community, the department of Human and Community Development at the University of Urbana Champaign. (Note: The first page asks for a login, but this is optional.)
- What is a Survey?, a pamphlet (downloadable or viewable online) originally created by the American Statistical Association

Take a few minutes and read through these articles. You'll be expected to use the information there and in your text to write your own survey, so read carefully!

Section 1.6: The Design of Experiments

1.1 Introduction to the Practice of Statistics
1.2 Observational Studies versus Designed Experiments
1.3 Simple Random Sampling
1.4 Other Effective Sampling Methods
1.5 Bias in Sampling
1.6 The Design of Experiments

Objectives

By the end of this lesson, you will be able to...

1. describe the characteristics of a designed experiment
2. explain the steps in designing an experiment
3. explain the types of experimental design
4. design your own experiment

Designed Experiments

Before we can talk about what to design an experiment, we first need to know what an experiment is in a statistics context.

A **designed experiment** is a controlled study in which one or more **treatments** are applied to **experimental units** (subjects). The experimenter then observes the effect of varying these treatments on a **response variable**.

You can see already that we've got quite a few terms. You may want to get the definition sheet that we started back in Section 1.1.

experimental unit - person or object upon which the treatment is applied

treatment - condition applied to the experimental unit

response variable - the variable of interest

factors - variables which affect the response variable

To help clarify all this terminology, let's consider a simple example:

Example 1

Consider the study we looked at in Example 3 in Section 1.2. It was from the New England Journal of Medicine and concerned the low-carb Atkins diet. If you need a refresher, here's a link to the summary of the report in the New England Journal of Medicine.

If you'd like more detail, there's a copy of the full article through the New England Journal of Medicine. Focus on the "Methods" section for details on the experimental design and sampling procedure.

Once you've reread the articles, try to determine the experimental units,

response variable, treatment, and factors from the study. When you're ready, click on the links in the table below to reveal the answer.

experimental units
response variable
treatment
factors

As is mentioned in your text, many designed experiments are double-blind. This means that neither the subjects nor the experimenters know who is receiving which treatment. Typically, subjects are assigned to two groups, with one receiving the treatment (like a new medical drug), while the other receives a placebo. This can be key to avoid researcher bias. Suppose, for example, that the previous study was done by the Atkins Institute and researchers new who was on which diet. Don't you think they'd be tempted to try to influence the results somehow?

In some cases, though, a single-blind experiment is preferable. One good example of this might be a study involving a heart medication. In this case, the doctors involved should be aware of who is taking the drug, and who is taking the placebo.

The Steps in Designing an Experiment

Step 1: *Identify the problem or claim to be studied.*
The statement of the problem needs to be as specific as possible. As your text says, it must "identify the response variable and the population to be studied".

Step 2: *Determine the factors affecting the response variable.*
This is best done by an expert in the field, but we'll be able to do this for most examples we'll be looking at.

Step 3: *Determine the number of experimental units.*
In general, more experimental units is better. Unfortunately, time and money will always be limiting factors, so we have to decide an appropriate number. We'll talk more about this later on in the course.

Step 4: *Determine the level(s) of each factor.*
We split factors up into three categories:

1. **Control:** If possible, we try to fix the level of factors that we're not interested in.
2. **Manipulate:** This is the treatment - we manipulate the levels of the variable that we think will affect the response variable.
3. **Randomize:** Often, there are factors we just can't control. To mitigate their effect on the data, we randomize the groups. By randomly assigning experimental units, these factors should be equally spread among all groups.

Step 5: *Conduct the experiment.*

Step 6: *Test the claim.*
We'll focus on this step much later in the course - Chapters 9-12. It uses **inferential statistics**, where we look at information from a sample and try to make a generalization about the population.

OK, now that we have the basic process down, let's look at an example using various designs.

We're going to focus on three particular experimental designs - **completely randomized**, **matched-pairs**, and **randomized block**. Your textbook also goes through all three following an example of the effect of fertilizers on plant growth. I'm going to do something similar, but using a different example.

Example 2

Suppose we want to determine the effect of using the practice exams on student exam scores. If we do a survey of students and determine which have used the practice exam and which haven't, we might not really know if the practice exam made a difference. Can you see why?

OK, I have an idea.

Let's start our design process.

Step 1: *Identify the problem or claim to be studied.*
We want to study the effectiveness of course supplements on student success. For the purpose of this study, we'll specify our population as all Mth120 student at ECC. In addition, we'll characterize "success" based on the 1st exam score.

Step 2: *Determine the factors affecting the response variable.*
There are plenty of factors here, but let's list a few. Obviously, the use of course supplements is a factor. We might also include intelligence, previous knowledge, study habits, sleep, diet, number of hours working, and of course, the instructor! I'm sure you could come up with several more.

Step 3: *Determine the number of experimental units.*
This will depend on which design we use, so let's hold off on this step until later.

Step 4: *Determine the level(s) of each factor.*
I'll take the list of factors we have above, and try to fit them into one of the groups.

1. **Control:** Looking at the list, the only factor I see that we can control would be the instructor - we'll make sure that all the students involved have the same instructor.
2. **Manipulate:** This is the treatment - supplements used. Let's have three levels - reviewing without the video and using the Video Lecture Series.
3. **Randomize:** This is everything else - intelligence, previous knowledge, study habits, sleep, diet, and number of hours working.

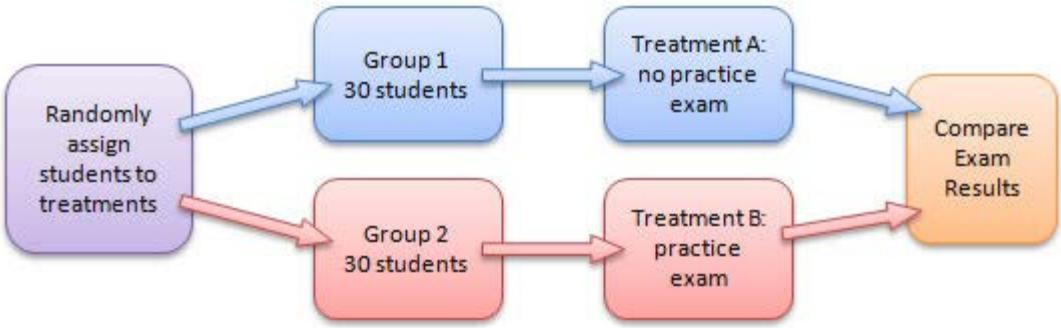
That's the basics. Now on to the experiment itself.

Completely Randomized Design

A **completely randomized design** is when each experimental unit is assigned to a treatment completely at random. (This is similar to simple random sampling.)

In this design, we would randomly select 60 students and randomly split them into two groups with 30 each. One group does not take the practice exam, while the other does. We have the two groups then take the actual exam and we compare results.

Here's a visual:



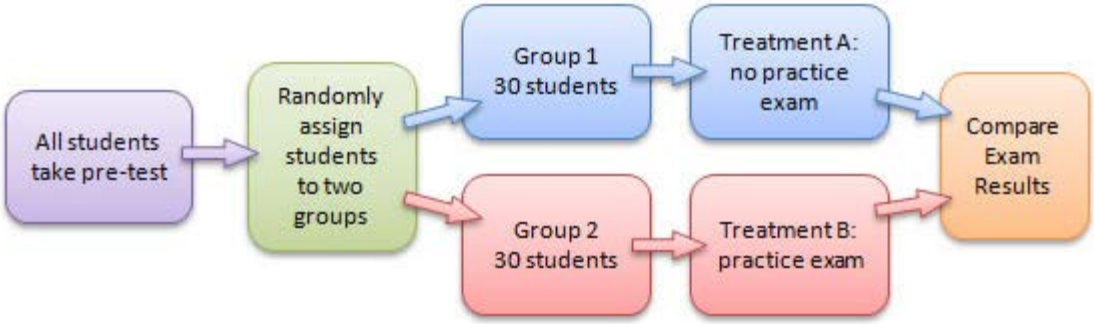
Matched-Pairs Design

A **matched-pair design** is when the experimental units are paired up and each of the pair is assigned to a different treatment.

There are a couple ways to do matched-pairs - we could find people who are very similar somehow, and have one do the practice exam and the other not. Unfortunately, there are so many factors affecting performance on the exam, this pretty impractical.

Another way to do a matched-pair design is to have the same individual before and after the treatment. In this case, we could do just that - give the exam, have students study the practice exam, and then give the exam again. The problem with this design is that we don't know if the improvement (if any) is from the practice exam or just from seeing the material again.

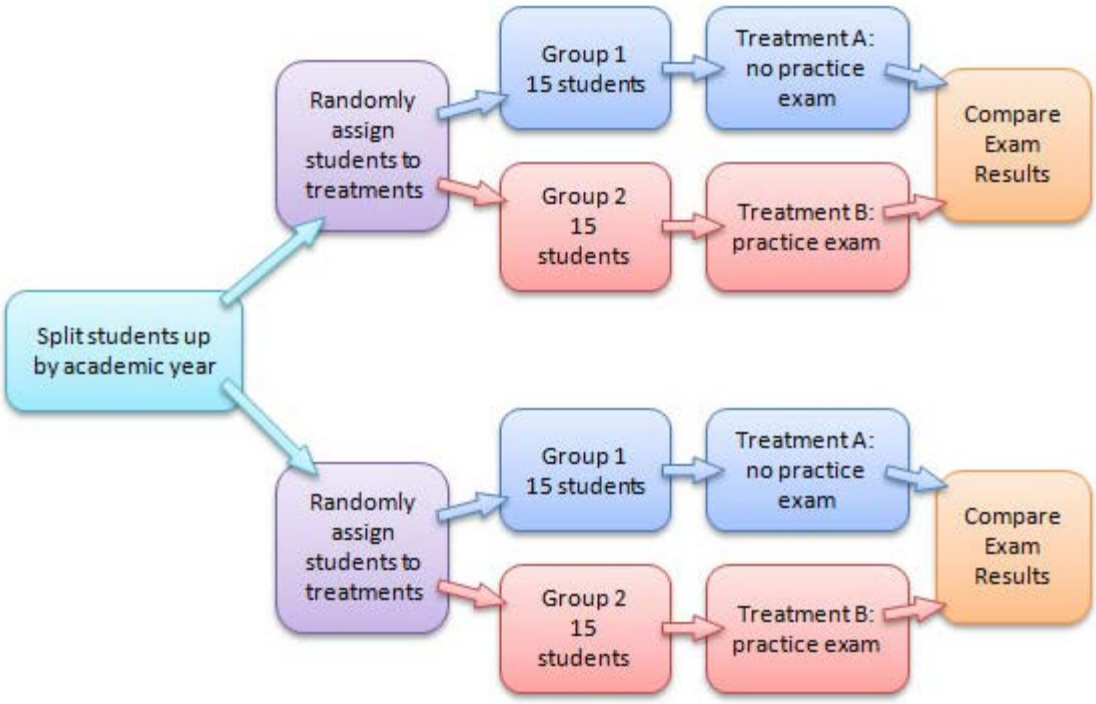
A better plan would be to have all individuals take the exam as a "pre-test", then have 30 students take the practice exam, while the rest do not. Then we have the students all take the exam again, and we compare the "before" and the "after".



Randomized Block Design

A **randomized block design** is used when the experimental units are divided into homogeneous groups called blocks. Within each block, the experimental units are randomly assigned treatments. (This is similar to stratified sampling.)

Student maturity is a *huge* factor in college success. Another idea might be to split our sample by academic year - those in their first year versus those in their second. Essentially, we're stratifying the sample, and then doing a completely randomized design on each of the "strata".



<< previous section | next section >>

Chapter: 1 2 3 4 5 6 7 8 9 10 11 12 13