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.

:: start ::

This work is licensed under a Creative Commons License.
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:

<table>
<thead>
<tr>
<th>Qualitative (or categorical) variables</th>
<th>allow for classification of individuals based on some attribute or characteristic.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative variables</td>
<td>provide numerical measures of individuals. Arithmetic operations such as...</td>
</tr>
</tbody>
</table>
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
This work is licensed under a Creative Commons License.
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.
A recent article in the BBC News Health section described a study concerning dementia and "mid-life ills". According to the article, researchers 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?

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?

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?
Section 1.3: Simple Random Sampling

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 like this:

For another take, watch this YouTube video:
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.
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.)

![Population Diagram]

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.

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:

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:**
- **Cluster Sampling:**
- **Multistage Sampling:**
Section 1.5: Sources of Errors in Sampling

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:

What issue is most important to you in determining which political candidate to support?

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?

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](https://www.pewresearch.org/fact-tank/2018/06/29/abortion-the-court-and-the-public/), from the Pew Research Center
- [Question Wording](https://www.apor.org/standards/guides/question-wording.html), information from the American Association for Public Opinion Research
- [Conducting a Survey In Your Community](https://people.humandev.illinois.edu/human_development/2011/12/survey_in_your_community.html), 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?](https://www.amstat.org/resources/pubs/whatissurvey.cfm), 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

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,
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".
Chapter 2: Organizing and Summarizing Data

2.1 Organizing Qualitative Data
2.2 Organizing Quantitative Data: The Popular Displays
2.3 Additional Displays of Quantitative Data
2.4 Graphical Misrepresentations of Data

Let's review the process of statistics we introduced in Section 1.1:

In Chapter 1, we focused on how to collect data. In this next chapter, we'll talk about how to organize and summarize data using tables in graphs. Section 2.2 will focus on qualitative data, while sections 2.2 and 2.3 will focus on quantitative data. The last section, Section 2.4, talks about various ways that data can be misrepresented.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::
Section 2.1: Organizing Qualitative Data

2.1 Organizing Qualitative Data
2.2 Organizing Quantitative Data: The Popular Displays
2.3 Additional Displays of Quantitative Data
2.4 Graphical Misrepresentations of Data

**Objectives**

By the end of this section, you will be able to...
1. organize qualitative data in tables
2. construct bar graphs
3. construct pie charts

**Frequency and Relative Frequency Tables**

Let's suppose you give a survey concerning favorite color, and the data you collect looks something like the table below.

<table>
<thead>
<tr>
<th>favorite color</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>10</td>
</tr>
<tr>
<td>red</td>
<td>3</td>
</tr>
<tr>
<td>orange</td>
<td>1</td>
</tr>
<tr>
<td>yellow</td>
<td>3</td>
</tr>
<tr>
<td>green</td>
<td>5</td>
</tr>
<tr>
<td>pink</td>
<td>3</td>
</tr>
<tr>
<td>purple</td>
<td>1</td>
</tr>
</tbody>
</table>

Clearly, we need a better way to summarize the data. The most obvious thing to do would be to make a table with the list of favorite colors and the frequency for each.

A **frequency distribution** lists each category of data and the number of occurrences for each category.

The **relative frequency** is the proportion (or percent) of observations within a category and is found using the formula:

\[
\text{relative frequency} = \frac{\text{frequency}}{\text{sum of all frequencies}}
\]
A relative frequency distribution lists each category of data together with the relative frequency of each category.

<table>
<thead>
<tr>
<th>favorite color relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
</tr>
<tr>
<td>red</td>
</tr>
<tr>
<td>orange</td>
</tr>
<tr>
<td>yellow</td>
</tr>
<tr>
<td>green</td>
</tr>
<tr>
<td>pink</td>
</tr>
<tr>
<td>purple</td>
</tr>
</tbody>
</table>

Technology

Here’s a quick overview of how to create frequency and relative frequency tables in StatCrunch.

1. Enter or import the data.
2. Select Stat > Tables > Frequency.
3. Select the column(s) you want to summarize and click Next.
4. Add any modifications for an "Other" category and how to order the categories.
5. Click Calculate and another window with these numbers calculated will pop up.
6. You can then choose Options > Copy to copy the output for use elsewhere.

Bar Graphs

Bar graphs are probably the most commonly used graphs, and one you're already familiar with. I won't mention much more here, except to state a couple keys:

1. heights can be frequency or relative frequency
2. bars must not touch

Using our the data from our previous color example,

<table>
<thead>
<tr>
<th>favorite color frequency relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
</tr>
<tr>
<td>red</td>
</tr>
<tr>
<td>orange</td>
</tr>
<tr>
<td>yellow</td>
</tr>
<tr>
<td>green</td>
</tr>
<tr>
<td>pink</td>
</tr>
<tr>
<td>purple</td>
</tr>
</tbody>
</table>

we could then make both frequency and relative frequency bar graphs.
Here's a quick overview of how to create bar graphs in StatCrunch.

1. Enter or import the data.
2. Select **Graphics > Bar Graph**, then choose **with data** or **with summary**.
3. If you chose **with data**, select the column(s) you wish to use and click **Next**. If you chose **with summary**, set the columns containing the categories and counts and click **Next**.
4. Choose the type (**Frequency** or **Relative Frequency**) and click **Next**.
5. Enter any modifications and/or color schemes and click **Create Graph**!
6. You can then choose **Options > Copy** to copy the box plot for use elsewhere.
Pareto Charts

A Pareto chart is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.

You see Pareto charts fairly often in the newspaper, because often the article is trying to show that one particular category is the highest or lowest. The image below, for example, is from the Chicago Tribune. You can see clearly from the graph that it's attempting to show that the local BP refinery in Whiting, Indiana is the highest-capacity refinery that is considering expansion.

Midwest oil refineries look to grow

REFINERIES CONSIDERING UPGRADES
And crude refining capacity, in barrels per day, as of Jan. 1, 2007:

1. **BP**, Whiting, Ind. 410,000
2. **ConocoPhillips**, Wood River, Ill. 306,000
3. **Marathon**, Robinson, Ill. 192,000
4. **Husky**, Lima, Ohio 146,120
5. **BP**, Toledo, Ohio 131,000
6. **Marathon**, Detroit, Mich. 100,000
7. **Marathon**, St. Paul Park, Minn. 70,000
8. **Murphy**, Superior, Wis. 34,300

SOURCES: Energy Information Administration, Tribune reporting

TRIBUNE GRAPHIC

If you don't remember the issue, you can read up about BP's plan to expand its refinery in [this article from CBS2 Chicago](#).

Here's another one, using the favorite color data from the last section:

Side-by-Side Bar Graphs
Side-by-side bar graphs are used when you want to compare two different populations. The key with side-by-side bar graphs is that you must use relative frequencies. Do you know why?

I think so. But just in case...

Here's a good example of a side-by-side chart, from the Associated Press.

What's shown isn't quite a relative frequency as we've defined it - it's the number per 100,000, where ours as a percent is the number per 100. The reason why the rate per 100,000 is used here is because the percents would all be less than 1% and difficult to read. Still, if frequency was used instead, the "White" category would be the largest, simply because that's the largest segment of the U.S. population.

**Technology**

Here's a quick overview of how to create side-by-side bar graphs in StatCrunch.

1. Enter or import the data.
2. Select **Graphics > Chart > Columns**
3. Select the columns you'll be using.
4. Select the location of the labels (Row labels in).
5. If desired, choose an order.
6. Choose the plot type (vertical bars for a side-by-side bar graph) and click **Next**.
7. Enter any modifications and/or color schemes and click **Create Graph!**
8. You can then choose **Options > Copy** to copy the box plot for use elsewhere.
Pie Charts

Like bar graphs, pie charts are very common. You're probably already aware of these as well. I'll just include a couple comments:

1. should always include the relative frequency
2. also should include labels, either directly or as a legend

Using our the data from our previous color example,

<table>
<thead>
<tr>
<th>favorite color</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>10</td>
<td>$10/26 \approx 0.38$</td>
</tr>
<tr>
<td>red</td>
<td>3</td>
<td>$3/26 \approx 0.12$</td>
</tr>
<tr>
<td>orange</td>
<td>1</td>
<td>$1/26 \approx 0.04$</td>
</tr>
<tr>
<td>yellow</td>
<td>3</td>
<td>$3/26 \approx 0.12$</td>
</tr>
<tr>
<td>green</td>
<td>5</td>
<td>$5/26 \approx 0.19$</td>
</tr>
<tr>
<td>pink</td>
<td>3</td>
<td>$3/26 \approx 0.12$</td>
</tr>
<tr>
<td>purple</td>
<td>1</td>
<td>$1/26 \approx 0.04$</td>
</tr>
</tbody>
</table>

we get this pie chart:. 
Here's a quick overview of how to create pie charts in StatCrunch.

1. Enter or import the data.
2. Select **Graphics > Pie Chart**, then choose **with data** or **with summary**.
3. If you chose **with data**, select the column(s) you wish to use and click **Next**. If you chose **with summary**, set the columns containing the categories and counts and click **Next**.
4. Enter any modifications (labels, title, color scheme, etc) and click **Create Graph!**
5. You can then choose **Options > Copy** to copy the box plot for use elsewhere.
Section 2.2: Organizing Quantitative Data: The Popular Displays

2.1 Organizing Qualitative Data
2.2 Organizing Quantitative Data: The Popular Displays
2.3 Additional Displays of Quantitative Data
2.4 Graphical Misrepresentations of Data

Objectives

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

1. organize quantitative data into tables
2. construct histograms for discrete and continuous data
3. draw stem-and-leaf plots
4. draw dot plots
5. identify the shape of a distribution

Like qualitative data in the last section, quantitative data can (and should) be organized into tables. We'll break this page up into two parts - discrete and continuous.

Organizing Discrete Data into Tables

If you recall from Section 1.2,

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, ...)

Since we can list all the possible values (that's essentially what countable means), one way to make a table is just to list the values along with their corresponding frequency.

Example 1

Here's some data I collected from a previous students Mth120 course. It refers to the number of children in their family (including themselves).

<table>
<thead>
<tr>
<th>children</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3/26 ≈ 0.12</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>8/26 ≈ 0.31</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>10/26 ≈ 0.38</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2/26 ≈ 0.08</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3/26 ≈ 0.12</td>
</tr>
</tbody>
</table>
Sometimes, however, we have too many values to make a row for each one. In that case, we'll need to group several values together.

**Example 2**

A good example might be the scores on an exam, ranging from 1-100. Here are some data from a past Mth120 class.

<table>
<thead>
<tr>
<th>62 87 67 58 95 94 91 69 52</th>
</tr>
</thead>
<tbody>
<tr>
<td>76 82 85 91 60 77 72 83 79</td>
</tr>
<tr>
<td>63 88 79 88 70 75 87</td>
</tr>
</tbody>
</table>

In this case, we'll have to set up intervals of numbers called **classes**. Each class has a **lower class limit** and an **upper class limit**, along with a **class width**. The class width is the difference between successive lower class limits.

To be consistent, the class width should be same for each class. One good option might look something like this:

<table>
<thead>
<tr>
<th>Exam Score</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>50-59</td>
<td>2</td>
</tr>
<tr>
<td>60-69</td>
<td>5</td>
</tr>
<tr>
<td>70-79</td>
<td>7</td>
</tr>
<tr>
<td>80-89</td>
<td>7</td>
</tr>
<tr>
<td>90-99</td>
<td>4</td>
</tr>
</tbody>
</table>

**Organizing Continuous Data into Tables**

Organizing continuous data is similar to organizing multi-valued discrete data. We have to form classes which don't overlap. I usually try to design a class width that's either logical (i.e. 10 points for grades above) or so that I have 5-8 classes when complete.

**Example 3**

For this example, let's consider the average commute for each of the 50 states. The data below show the average daily commute of a random sample of 15 states.

<table>
<thead>
<tr>
<th>23.1 18.3 23.2 19.9 26.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.8 23.1 23.2 22.7 29.4</td>
</tr>
<tr>
<td>22.3 30.0 25.8 21.9 16.7</td>
</tr>
</tbody>
</table>

Source: **US Census**

Do you know why this is a continuous random variable and not discrete? (Hint: It's *not* because of the decimal.)
I think I know!

To make a frequency or relative frequency for continuous data, we use the same strategy we'd use for multi-valued discrete data.

<table>
<thead>
<tr>
<th>average commute</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-17.9</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>18-19.9</td>
<td>2</td>
<td>2/15 ≈ 0.13</td>
</tr>
<tr>
<td>20-21.9</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>22-23.9</td>
<td>6</td>
<td>6/15 = 0.40</td>
</tr>
<tr>
<td>24-25.9</td>
<td>2</td>
<td>2/15 ≈ 0.13</td>
</tr>
<tr>
<td>26-27.9</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>28-29.9</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>30-31.9</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
</tbody>
</table>

Once we have these tables, we'll need to learn how to create some charts to display the information, which is what the next few pages are about.

**Technology**

Here's a quick overview of how to create frequency and relative frequency tables for quantitative data in StatCrunch.

**Discrete Data**

1. Enter or import the data.
2. Select **Stat > Tables > Frequency**.
3. Select the column(s) you want to summarize and click **Next**.
4. Add any modifications for an "Other" category and how to order the categories, and click **Calculate**.

**Continuous or Multi-valued Discrete Data:**

1. Enter or import the data.
2. Select **Data > Bin Column**.
3. Select the column containing the data, select "Use fixed width bins", and set the lowest class limit (**Start bins at:**) and class (**bin**) width.
4. Click **Calculate**.
5. Select **Stat > Tables > Frequency**.
6. Select the newly created bin column and click **Calculate**.*

* Note that these classes *seem* to overlap, but that the class "0-k" does not include **Mk**.

**Stem-and-Leaf Plots**

Stem-and-leaf plots are another way to represent quantitative data. They give more detail because they show the actual data. The idea is to split each data value into two parts - a **stem** and a **leaf**. The **stem** is everything of the right-most digit, and the **leaf** is that right-most digit. Here's an example, using the data from earlier this section regarding exam scores from a previous Mth120 class.

Example 6

| 62 | 87 | 67 | 58 | 95 | 94 | 91 | 69 | 52 |


With these data, the stems are the first digits - 5, 6, 7, 8, and 9. The leafs are all the second digits, 0, 1, ..., 9. The full **stem-and-leaf plot** lists the stems down the left side, a vertical bar between, and then lists the leafs in order to the right. Something like this:

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2 8</td>
</tr>
<tr>
<td>6</td>
<td>2 7 9 0 3</td>
</tr>
<tr>
<td>7</td>
<td>6 7 2 9 9 0 5</td>
</tr>
<tr>
<td>8</td>
<td>7 2 5 3 8 8 7</td>
</tr>
<tr>
<td>9</td>
<td>5 4 1 1</td>
</tr>
</tbody>
</table>

It's interesting that this plot looks very similar to a histogram, only it gives us the actual data. Take a look at this animation to see the relationship:

There are some limitations to stem-and-leaf plots. In particular, we're limited to small data sets - can you imagine the leaves if we had 1,000 test scores? Also, the range in the data needs to be fairly small.

By that, I mean if the data values range from 1-100, our stems can be 0, 10, 20, ..., 90, as they were in this example. On the other hand, if the values range from 1-10,000, the stems would have to be 0, 10, 20, ..., 9,980, 9,990. That's a lot of rows!

**Technology**

Here's a quick overview of how to create stem-and-leaf plots in StatCrunch.

1. Enter or import the data.
2. Select **Graphics > Stem and Leaf**
3. Select the column you wish to use and click **Create Graph!**

**Dot Plots**

Dot pots are similar to single-valued histograms, but rather than placing rectangles above each particular value, a dot plot just places the required number of dots above each value. Looking at our example again with the number of children, the plot would look something like this:
Here’s a quick overview of how to create dot plots in StatCrunch.

1. Enter or import the data.
3. Select the column you wish to use and click Next.
4. Set any options and click Create Graph!

Distribution Shape

A good way to describe a distribution is its shape. In general, we describe a distribution’s shape in one of four ways (though there are others):

1. uniform - frequencies are evenly spread out among all values of the variable
2. symmetric (bell-shaped) - highest value is in the middle, with values tailing off to the right and left
3. left-skewed - highest value is on the right, with a longer left "tail"
4. right-skewed - highest values is on the left, with a longer right "tail"
This work is licensed under a Creative Commons License.
Section 2.3: Additional Displays of Quantitative Data

2.1 Organizing Qualitative Data
2.2 Organizing Quantitative Data: The Popular Displays
2.3 Additional Displays of Quantitative Data
2.4 Graphical Misrepresentations of Data

Objectives

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

1. construct frequency polygons*
2. create cumulative frequency and relative frequency tables
3. construct ogives*
4. draw time-series graphs

* You will not be tested on these objectives.

In addition to histograms, stem-and-leaf plots, and dot plots, there are some other, section common plots. We'll introduce a couple in this section. The first type, frequency polygons, are not a type of plot that will be expected of you on exams, though you will be asked questions about them on homework.

Frequency Polygons

A frequency polygon is drawn by plotting a point above each class midpoint and connecting the points with a straight line. (Class midpoints are found by average successive lower class limits.)

Example 1

To illustrate the idea, let's look at the average commute data from the last section.

<table>
<thead>
<tr>
<th>average commute</th>
<th>midpoint</th>
<th>frequency</th>
<th>relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-17.9</td>
<td>17</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>18-19.9</td>
<td>19</td>
<td>2</td>
<td>2/15 ≈ 0.13</td>
</tr>
<tr>
<td>20-21.9</td>
<td>21</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>22-23.9</td>
<td>23</td>
<td>6</td>
<td>6/15 = 0.40</td>
</tr>
<tr>
<td>24-25.9</td>
<td>25</td>
<td>2</td>
<td>2/15 ≈ 0.13</td>
</tr>
<tr>
<td>26-27.9</td>
<td>27</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>28-29.9</td>
<td>29</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>30-31.9</td>
<td>31</td>
<td>1</td>
<td>1/15 ≈ 0.07</td>
</tr>
</tbody>
</table>

The three images below show the relationship between the histogram and the frequency polygon.
Cumulative tables are just what they imply - they show the sum of values up to and including that particular
Example 2

To illustrate the idea, let's look at the average commute data from the last section.

<table>
<thead>
<tr>
<th>average commute</th>
<th>frequency</th>
<th>cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-17.9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18-19.9</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>20-21.9</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>22-23.9</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>24-25.9</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>26-27.9</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>28-29.9</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>30-31.9</td>
<td>1</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>average commute</th>
<th>relative frequency</th>
<th>cumulative relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-17.9</td>
<td>1/15 ≈ 0.07</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>18-19.9</td>
<td>2/15 ≈ 0.13</td>
<td>3/15 ≈ 0.20</td>
</tr>
<tr>
<td>20-21.9</td>
<td>1/15 ≈ 0.07</td>
<td>4/15 ≈ 0.27</td>
</tr>
<tr>
<td>22-23.9</td>
<td>6/15 = 0.40</td>
<td>10/15 ≈ 0.67</td>
</tr>
<tr>
<td>24-25.9</td>
<td>2/15 ≈ 0.13</td>
<td>12/15 = 0.80</td>
</tr>
<tr>
<td>26-27.9</td>
<td>1/15 ≈ 0.07</td>
<td>13/15 ≈ 0.87</td>
</tr>
<tr>
<td>28-29.9</td>
<td>1/15 ≈ 0.07</td>
<td>14/15 ≈ 0.93</td>
</tr>
<tr>
<td>30-31.9</td>
<td>1/15 ≈ 0.07</td>
<td>15/15 = 1.00</td>
</tr>
</tbody>
</table>

Technology

Unfortunately, there is no easy way to create cumulative tables in StatCrunch. The best method is to create a regular frequency or relative frequency table and compute the cumulative values by hand.

Ogives

Ogives are pretty funky graphs, and rarely used except in specific areas. We'll just give a quick example here, but like frequency polygons, you won't be expected to create these on an exam. (Though it may come up in homework.)

An ogive (read as "oh jive") is a graph that represents the cumulative frequency or cumulative relative frequency for the class. It is constructed by plotting points - the x-coordinates are the upper class limits and the y-coordinate is the corresponding cumulative frequency or cumulative relative frequency.
Example 3

To illustrate the idea, let's again use the average commute data from the last section.

<table>
<thead>
<tr>
<th>average commute</th>
<th>relative frequency</th>
<th>cumulative relative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-17.9</td>
<td>1/15 ≈ 0.07</td>
<td>1/15 ≈ 0.07</td>
</tr>
<tr>
<td>18-19.9</td>
<td>2/15 ≈ 0.13</td>
<td>3/15 ≈ 0.20</td>
</tr>
<tr>
<td>20-21.9</td>
<td>1/15 ≈ 0.07</td>
<td>4/15 ≈ 0.27</td>
</tr>
<tr>
<td>22-23.9</td>
<td>6/15 = 0.40</td>
<td>10/15 ≈ 0.67</td>
</tr>
<tr>
<td>24-25.9</td>
<td>2/15 ≈ 0.13</td>
<td>12/15 = 0.80</td>
</tr>
<tr>
<td>26-27.9</td>
<td>1/15 ≈ 0.07</td>
<td>13/15 ≈ 0.87</td>
</tr>
<tr>
<td>28-29.9</td>
<td>1/15 ≈ 0.07</td>
<td>14/15 ≈ 0.93</td>
</tr>
<tr>
<td>30-31.9</td>
<td>1/15 ≈ 0.07</td>
<td>15/15 = 1.00</td>
</tr>
</tbody>
</table>

Note: No technology section this time, since you won't be asked to do this for exams.

Time-Series Graphs

Time series graphs are much more common than the last couple times we've looked at. It's common to see stock
prices and daily temperature graphs in the news - both are time series plots.

A time series plot is obtained by plotting the time in which a variable is measured on the horizontal axis and the corresponding value of the variable on the vertical axis.

The example above is from the Chicago Tribune and reflects the price of uranium from 2001-2006.

Example 4

Here's another example, using the daily high temperature in Elgin, IL, for the month of June, 2008.

<table>
<thead>
<tr>
<th>date</th>
<th>daily high temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/1</td>
<td>80</td>
</tr>
<tr>
<td>6/2</td>
<td>86</td>
</tr>
<tr>
<td>6/3</td>
<td>72</td>
</tr>
<tr>
<td>6/4</td>
<td>81</td>
</tr>
<tr>
<td>6/5</td>
<td>89</td>
</tr>
<tr>
<td>6/6</td>
<td>89</td>
</tr>
<tr>
<td>6/7</td>
<td>86</td>
</tr>
<tr>
<td>6/8</td>
<td>85</td>
</tr>
<tr>
<td>6/9</td>
<td>73</td>
</tr>
<tr>
<td>6/10</td>
<td>80</td>
</tr>
<tr>
<td>6/11</td>
<td>84</td>
</tr>
<tr>
<td>6/12</td>
<td>91</td>
</tr>
<tr>
<td>6/13</td>
<td>82</td>
</tr>
<tr>
<td>6/14</td>
<td>84</td>
</tr>
<tr>
<td>6/15</td>
<td>81</td>
</tr>
<tr>
<td>6/16</td>
<td>72</td>
</tr>
<tr>
<td>6/17</td>
<td>77</td>
</tr>
<tr>
<td>6/18</td>
<td>78</td>
</tr>
<tr>
<td>6/19</td>
<td>81</td>
</tr>
<tr>
<td>6/20</td>
<td>85</td>
</tr>
<tr>
<td>6/21</td>
<td>82</td>
</tr>
<tr>
<td>6/22</td>
<td>81</td>
</tr>
<tr>
<td>6/23</td>
<td>78</td>
</tr>
<tr>
<td>6/24</td>
<td>81</td>
</tr>
<tr>
<td>6/25</td>
<td>80</td>
</tr>
<tr>
<td>6/26</td>
<td>85</td>
</tr>
<tr>
<td>6/27</td>
<td>82</td>
</tr>
<tr>
<td>6/28</td>
<td>83</td>
</tr>
<tr>
<td>6/29</td>
<td>75</td>
</tr>
<tr>
<td>6/30</td>
<td>81</td>
</tr>
</tbody>
</table>

And the time series plot would look something like this:
Here's a quick overview of how to create a time series plot in StatCrunch.

1. Enter or import the data.
2. Select **Graphics > Index Plot**
3. Select the column(s) you want to plot and click **Next**.
4. Set any desired options and click **Create Graph!**
Section 2.4: Graphical Misrepresentations of Data

2.1 Organizing Qualitative Data  
2.2 Organizing Quantitative Data: The Popular Displays  
2.3 Additional Displays of Quantitative Data  
2.4 Graphical Misrepresentations of Data

Objectives

By the end of this section, you will be able to...
1. describe what can make a graph misleading or deceptive

Misleading and Deceptive Graphs

The author of your text makes an interesting distinction between "misleading" and "deceptive" graphs. It's an important point, so read through that paragraph before continuing on to the examples. (Page 104)

Example 1

This first one was from the Washington Post after the Iowa caucuses in January, 2008. Look carefully at the graphic and try to determine what was misleading about it.

OK, I have an idea.

Example 2

This next graphic is attempting to relate the purchasing power of the Canadian dollar (also known as the "Loonie" - I love that!) in relation to the U.S. dollar. This is a bit more subtle. Can you see what’s misleading about this?
Example 3

Here’s a classic graphic from the Chicago Tribune. This is very typical of graphics representing the stock market. Can you see what’s wrong?

The Fed effect

Some say the Fed’s actions have tempered the market’s volatility, while others say the Fed’s actions aren’t having enough of a sustained positive impact on the Dow.

I think so. Let me see if I’m right.

Look for this error next time whenever you read an article that’s trying to show how quickly something is increasing or decreasing.
This work is licensed under a Creative Commons License.
Chapter 3: Numerically Summarizing Data

3.1 Measures of Central Tendency
3.2 Measures of Dispersion
3.3 Measures of Central Tendency and Dispersion from Grouped Data
3.4 Measures of Position and Outliers
3.5 Measures of Position and Outliers

Let's again review the process of statistics we introduced in Section 1.1:

In Chapter 1, we focused on how to collect data. In Chapter 2, we talked about how to organize and summarize data using tables in graphs. In this chapter, we'll introduce various ways to summarize data numerically, along with one new graphical representation - the box plot. In general, we have three ways to summarize the distribution of a random variable - shape, center, and spread. Shape was discussed back in Section 2.2, but the center and spread will be introduced here in Section 3.1 and Section 3.2, respectively. Section 3.3 talks about estimating measures of center and spread from grouped data.

In Section 3.4, we talk about summarizing information about an individual observation in relation to the rest of the sample/population. (We call these measures of position.)

And finally, in Section 3.5, we introduce a new graphical representation of data called the box plot. We'll be using this plot frequently throughout the semester.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::
Section 3.1: Measures of Central Tendency

3.1 Measures of Central Tendency
3.2 Measures of Dispersion
3.3 Measures of Central Tendency and Dispersion from Grouped Data
3.4 Measures of Position and Outliers
3.5 The Five-Number Summary and Boxplots

Objectives

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

1. determine the arithmetic mean of a variable from raw data
2. determine the median of a variable from raw data
3. explain what it means for a statistics to be resistant
4. determine the mode of a variable from raw data
5. use the mean and median to help identify the shape of a distribution

It's often very helpful to get a sense of what a "typical" individual might be in a population. This is what we mean when we say we're looking at measures of "center" or "central tendency".

Before we get into specifics, we need to clarify whether we're talking about typical individual from a population or from a sample.

A parameter is a descriptive measure of a population.
A statistic is a descriptive measure of a sample.

Arithmetic Mean

You already know the arithmetic mean, though maybe not by name. It's more commonly referred to as the "average". It's calculated just by finding the some of the values and dividing by the number of observations. As mentioned above, we'll have two different means - one for the population and one if we're talking about a sample.

The population arithmetic mean, \( \mu \) (pronounced "mew"), is computed using all the individuals in the population. The sample arithmetic mean, \( \bar{x} \) (pronounced "x-bar"), is computed using sample data.

\[
\text{population arithmetic mean: } \mu = \frac{\sum x_i}{N} = \frac{x_1 + x + ... + x_N}{N}
\]
\[
\text{sample arithmetic mean: } \bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x + ... + x_n}{n}
\]

This is pretty formulaic, but the concept should be relatively familiar.

To give another explanation, I'm going to reference one of my favorite web sites, BetterExplained. The author, Kalid Azad, presents topics in a non-traditional way, and I feel it's much more accessible and easier to understand than traditional texts. Here's what Kalid writes about the arithmetic mean:

The Arithmetic Mean
The arithmetic mean is the most common type of average:

\[
\text{average} = \frac{\text{sum}}{\text{number}}
\]

### Arithmetic Mean

Let’s say you weigh 150 lbs, and are in an elevator with a 100lb kid and 350lb walrus. What’s the average weight?

The real question is “If you replaced this merry group with 3 identical people and want the same load in the elevator, what should each clone weigh?”

In this case, we’d swap in three people weighing 200 lbs each \([(150 + 100 + 350)/3]\), and nobody would be the wiser.

**Pros:**

- It works well for lists that are simply combined (added) together.
- Easy to calculate: just add and divide.
- It’s intuitive — it’s the number “in the middle”, pulled up by large values and brought down by smaller ones.

**Cons:**

- The average can be skewed by outliers — it doesn’t deal well with wildly varying samples. The average of 100, 200 and -300 is 0, which is misleading.

The arithmetic mean works great 80% of the time; many quantities are added together. Unfortunately, there’s always those 20% of situations where the average doesn’t quite fit.

**Source:** BetterExplained, Kalid Azad

**Article:** How to Analyze Data Using the Average

Used with permission.

So, let’s try an example.

### Example 1

Suppose we record the exam scores from a sample of six students from a class of 30 (see table below).

<table>
<thead>
<tr>
<th>student exam score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joseph</td>
</tr>
<tr>
<td>Alicia</td>
</tr>
</tbody>
</table>
Find the sample mean, along with its appropriate symbol.

When necessary, **round the mean to one more digit than the original data.** i.e. If the data are whole numbers, you should round the mean to the tenths place (as in the previous example). If the data are already to the tenths place, you should round to the hundredths place.

You might also consider watching this video regarding rounding (in Quicktime or iPod format).

**Example 2**

One point that should be emphasized again is the effect of outliers on the arithmetic mean. Because it adds all the values together, the arithmetic mean can be skewed by extremely large or extremely small values.

A helpful way to illustrate this is to think of the mean as the center of gravity - like the balance point. Suppose we consider the ages of the six Jackson cousins, Hudson, Abella, Amelia, Jillian, Katelyn, and Jessica. The figure below represents their ages and the corresponding sample mean. *(Sample, in this case, because this isn't all of the Jackson cousins.)*

If we replace Jessica with her father, who is 34 years old, we get something like this:

You can see very clearly here the effect of including the dad. 16 years old does not really represent the "middle" value.

**Technology**

Here's a quick overview of the formulas for finding arithmetic mean in StatCrunch.
2. Select the variable you want to summarize (e.g., "Heights")—leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated.
5. Click "Calculate" and another window with these numbers calculated will pop up.

You can also visit the video page for links to see videos in either Quicktime or iPod format.

Median

As we mentioned at the end of the previous page, we need another measure of center when the data include outliers. The most common choice is called the median.

The median of a variable is the value that lies in the middle of the data when arranged in ascending order. That is, half the data are below the median and half the data are above the median. We use M to represent the median.

Like the previous topic, I really appreciate how Kalid Azad explained the median on his web site, BetterExplained. Here’s what he wrote:

The median is “the item in the middle”. But doesn’t the average (arithmetic mean) imply the same thing? What gives?

Humor me for a second: what’s the “middle” of these numbers?

1, 2, 3, 4, 100

Well, 3 is the middle of the list. And although the average (22) is somewhere in the “middle”, 22 doesn’t really represent the distribution. We’re more likely to get a number closer to 3 than to 22. The average has been pulled up by 100, an outlier.

The median solves this problem by taking the number in the middle of a sorted list. If there’s two middle numbers (even number of items), just take their average. Outliers like 100 only tug the median along one item in the sorted list, instead of making a drastic change: the median of 1 2 3 4 is 2.5.
Example 3

Let's again consider the exam scores from a sample of six students from a class of 30 (see table below).

<table>
<thead>
<tr>
<th>student</th>
<th>exam score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joseph</td>
<td>62</td>
</tr>
<tr>
<td>Alicia</td>
<td>83</td>
</tr>
<tr>
<td>Kendra</td>
<td>77</td>
</tr>
<tr>
<td>Cheryl</td>
<td>92</td>
</tr>
<tr>
<td>Adrian</td>
<td>89</td>
</tr>
<tr>
<td>Brian</td>
<td>75</td>
</tr>
</tbody>
</table>

Find the sample median.

[ reveal answer ]
Example 4

To illustrate how the median deals with outliers, let's again consider the ages of the six Jackson cousins. The figure below represents their ages and the corresponding sample median.

If we replace Jessica with her father, who is 34 years old, we get something like this:

You can immediately see the benefit of using the median - it is not affected by the age of Jessica's father.

Technology

Here's a quick overview of the formulas for finding median in StatCrunch.

1. Select **Stat > Summary Stat > Columns**.
2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated.
5. Click "Calculate" and another window with these numbers calculated will pop up.

You can also visit the video page for links to see videos in either Quicktime or iPod format.

Mode

Often, we just want to know what "most" people think on an issue. We don't call it that, but we're really looking
at is called the **mode**.

The **mode** of a variable is the most frequent observation of the variable.

Look at any poll from the Pew Research Center. Any time an article discusses the "most common" or "most popular" choice, it's talking about the mode.

As with the previous two measures of central tendency, I like Kalid Azad's explanation of the mode on his web site, BetterExplained. Here's what he wrote:

**Mode**

The mode sounds strange, but it just means **take a vote**. And sometimes a vote, not a calculation, is the best way to **get a representative sample** of what people want.

Let's say you're throwing a party and need to pick a day (1 is Monday and 7 is Sunday). The “best” day would be the option that satisfies the most people: an average may not make sense. (“Bob likes Friday and Alice likes Sunday? Saturday it is!”).

Similarly, colors, movie preferences and much more can be **measured with numbers**. But again, the ideal choice may be the mode, not the average: the “average” color or “average” movie could be... unsatisfactory (Rambo meets Pride and Prejudice).

**Pros:**

- Works well for exclusive voting situations (this choice or that one; no compromise)
- Gives a choice that the most people wanted (whereas the average can give a choice that nobody wanted).
- Simple to understand

**Cons:**

- Requires more effort to compute (have to tally up the votes)
- "Winner takes all" — there’s no middle path

The term "mode" isn’t that common, but now you know what button to look for when playing around with your favorite statistics program.

**Source:** BetterExplained, Kalid Azad
To see how to find the mode using technology, open the appropriate video from the list below. These videos include all measures of center included in this section, plus other descriptive statistics.

Visit the [video page](#) for links to see videos in either Quicktime or iPod format.

### Using the Mean and Median to Identify the Distribution Shape

In Section 2.2, we talked about different ways to describe the [distribution shape](#). With these new measures of center, we can now use the mean and median to get an idea of the distribution shape as well.

![Distribution Shape Diagram](#)
mean and median
approximately equal

symmetric


Section 3.2: Measures of Dispersion

### Objectives

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

1. compute the range of a variable from raw data
2. compute the variance of a variable from raw data
3. compute the standard deviation of a variable from raw data
4. use the empirical Rule to describe data that are bell-shaped

Consider the following two sets of exam scores:

<table>
<thead>
<tr>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>48 57 58 65 68 69 71 73 73</td>
<td>58 58 61 62 63 64 66 71 71</td>
</tr>
<tr>
<td>74 75 77 78 78 79 80 85</td>
<td>72 73 75 76 77 77 78 78</td>
</tr>
<tr>
<td>87 88 89 89 95 96 97 99</td>
<td>79 81 82 84 85 88 90 91</td>
</tr>
</tbody>
</table>

![Histograms](image)

Often, we want to compare two data sets and look for differences. In this case, however, the sample mean for the first set is 78.3, with a median of 78, while the second set has a sample mean of 78.7, also with a median of 78.

We can see that the measures of center are not enough to distinguish between the two sets, so we'll need to somehow compare their "dispersion", or spread. The first statistic we'll learn about to help do that is called the **range**.

### Range

The **range**, $R$, of a variable is the difference between the largest and smallest data values.

### Example 1

Looking at our data from above, we see that the range for the first set of exam scores is $99-48 = 51$, while the range for the second set is $91-58 = 33$. As we can see from the histograms, the second set of exam scores
has less dispersion.

Unfortunately, the range isn't always enough to distinguish between two sets of data, which we'll see on the next page.

Let's look at another two sets of exam scores.

```
48 49 52 55 57 58 62 64 65
66 67 72 73 75 78 78 78
79 82 84 86 88 89 93 94 95
48 55 57 61 64 65 68 71 71
72 73 74 75 78 78 79 79
79 79 82 84 85 88 89 92 95
```

In this case, we can see that the range is $R = 95-48 = 47$ for both sets, but they clearly don't have the same dispersion. The second set is much more condensed, with the bulk of the scores in the C range - 70-79.

**Variance**

To describe the dispersion in cases like these, we'll need to somehow describe how far a "typical" observation is from the mean, rather than looking at the extreme values.

An obvious choice would be to just look at the average distance from the mean. The figure below shows the heights of the 2008 US Men's Olympic Basketball team and each player's corresponding difference from the mean. (Source: USA Basketball)
If we try to take the average difference from the mean, we have a problem - we get 0!

\[
\frac{1.5 + 2.5 + 3.5 - 0.5 + 4.5 + 1.5 - 2.5 - 6.5 + 2.5 - 0.5 - 2.5 - 3.5}{12} = 0
\]

Why is this? Well, it's because the mean acts as a balancing point, as we talked about earlier. In fact, this will always happen - the average difference from the mean will equal zero.

So what do we do? The obvious choice would be to take the average distance from the mean. That is, take the absolute value of each of the differences. It's a good thought, but anyone who's taken calculus will tell you that absolute values can be pretty difficult to work with, so instead, we square them.

This creates a new measure of dispersion, called the **variance**.

The **population variance**, \( \sigma^2 \), of a variable is the sum of the squared deviations about the population mean divided by \( N \).

\[
\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + ... + (x_N - \mu)^2}{N} = \frac{\sum (x_i - \mu)^2}{N}
\]

This is pretty complex, but we'll have technology to do most of the work for us.

Like the mean, we also have a sample version of this calculation, but unlike the mean, it's actually different.

The **sample variance**, \( s^2 \), is computed by determining the sum of the squared deviations about the sample mean and dividing the result by \( n-1 \).
The first thing most students ask when they see this (I did, too) is "Why n-1 instead of just n?" It's a good question, and a difficult to answer in plain English. The key is to look at the purpose of using the sample variance (or any sample statistics, for that matter). That purpose is to get an estimate for the true population variance.

Unless we have data for the entire population, our estimate will likely be incorrect. If we look at the average of all possible sample variances, though, that average should be the same as the population variance we're trying to estimate. In other words, we'll be wrong most of the time, but the average of all of our attempts will be correct.

The thing is, if we divide by N in the sample variance formula above, our estimate will, on average, be too low. (We can actually prove this mathematically, but it's pretty heady stuff. It's usually not covered until a graduate course in probability and statistics.) We call an estimate like this unbiased, since it consistently under-estimates the parameter it's trying to predict.

Interestingly enough, dividing by n-1 makes the estimate unbiased. (This can also be proven mathematically.) So it may seem like an odd thing to do, but there's very solid mathematical reasoning behind it.

If you'd like more information on this, you can read a more thorough analysis in your text on pages 141-142.

**Example 2**

Let's refer back to the heights of the players on the US Men's Olympic basketball team, and let's treat this as a sample of all the basketball players in the US.

<table>
<thead>
<tr>
<th>Player</th>
<th>Height</th>
<th>(x_i)</th>
<th>(x_i - \bar{x})^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carmelo Anthony</td>
<td>6'8&quot;</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>Carlos Boozer</td>
<td>6'9&quot;</td>
<td>2.5</td>
<td>6.25</td>
</tr>
<tr>
<td>Chris Bosh</td>
<td>6'10&quot;</td>
<td>3.5</td>
<td>12.25</td>
</tr>
<tr>
<td>Kobe Bryant</td>
<td>6'6&quot;</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>Dwight Howard</td>
<td>6'11&quot;</td>
<td>4.5</td>
<td>20.25</td>
</tr>
<tr>
<td>LeBron James</td>
<td>6'8&quot;</td>
<td>1.5</td>
<td>2.25</td>
</tr>
<tr>
<td>Jason Kidd</td>
<td>6'4&quot;</td>
<td>-2.5</td>
<td>6.25</td>
</tr>
<tr>
<td>Chris Paul</td>
<td>6'0&quot;</td>
<td>-6.5</td>
<td>42.25</td>
</tr>
<tr>
<td>Tayshaun Prince</td>
<td>6'9&quot;</td>
<td>2.5</td>
<td>6.25</td>
</tr>
<tr>
<td>Michael Redd</td>
<td>6'6&quot;</td>
<td>-0.5</td>
<td>0.25</td>
</tr>
<tr>
<td>Dwayne Wade</td>
<td>6'4&quot;</td>
<td>-2.5</td>
<td>6.25</td>
</tr>
<tr>
<td>Deron Williams</td>
<td>6'3&quot;</td>
<td>-3.5</td>
<td>12.25</td>
</tr>
</tbody>
</table>

So the sample variance is then:

\[
s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{117}{12 - 1} \approx 10.6
\]

You may notice that I rounded the variance to the tenths place.

Typically, we round the variance to one more digit than the data

---

When necessary, **round the variance to one more digit than the original data**, i.e. If the data
are whole numbers, you should round the variance to the tenths place (as in the previous example). If the data are already to the tenths place, you should round to the hundredths place.

**Technology**

Here's a quick overview of the formulas for finding variance in StatCrunch.

1. Select **Stat > Summary Stat > Columns**.
2. Select the variable you want to summarize (e.g., "Heights")—leave everything else as is for now.
3. Click **Next**.
4. Deselect any statistics that you do not want calculated.
5. Click **Calculate** and another window with these numbers calculated will pop up.

You can also visit the video page for links to see videos in either Quicktime or iPod format.

One major problem with the variance is that the units don't really make sense. Take the previous example about the heights of the players on the 2008 US Men's Olympic Basketball team. If we look at the units for that variance, it's 10.64 inches squared. What does that have to do with the dispersion of the data? The data are in inches, not inches squared!

**Standard Deviation**

To remedy that, we need another measure of dispersion, called the **standard deviation**.

The **population standard deviation**, \( \sigma \), is obtained by taking the square root of the population variance.

\[
\sigma = \sqrt{\sigma^2}
\]

The **sample standard deviation**, \( s \), is obtained by taking the square root of the sample variance.

\[
s = \sqrt{s^2}
\]

So referring again to our previous example, the sample standard deviation is \( \approx \sqrt{10.6} \approx 3.3 \) inches. So we could then say that the "typical" player is about 3.3 inches different from the average height of the team.

Now that makes more sense!

When necessary, round the **standard deviation to one more digit than the original data**. i.e.

If the data are whole numbers, you should round the standard deviation to the tenths place (as in the previous example). If the data are already to the tenths place, you should round to the thousandths place.

You might also consider watching this video regarding rounding (in Quicktime or iPod format).

**What Does It Mean?**
So what can we tell from the standard deviation? Let's go back to those two sets of exam scores. Which one do you think has a higher standard deviation?

48 49 52 55 57 58 62 64 65
66 67 72 73 75 78 78 78
79 82 84 86 88 89 93 94 95

48 55 57 61 64 65 68 71 71
72 73 73 74 75 78 78 79 79
79 79 82 84 85 88 89 92 95

[ reveal answer ]

[ Why? ]

Technology

Here's a quick overview of the formulas for finding standard deviation in StatCrunch.

2. Select the variable you want to summarize (e.g., "Heights")—leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated.
5. Click "Calculate" and another window with these numbers calculated will pop up.

You can also visit the video page for links to see videos in either Quicktime or iPod format.

One nice benefit of understanding the relationship between the standard deviation and the shape of the distribution is it helps us get a sense of how much of the data should be within a certain number of standard deviations.

In particular, if the distribution is bell-shaped, we can be fairly precise about what percentage of the data should lie within 1, 2, or 3 standard deviations.

The Empirical Rule
If a distribution is roughly bell-shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean.
- Approximately 95% of the data will lie within 2 standard deviations of the mean.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean.

How do we know these percentages so accurately? Well, unfortunately we can't explain it until we get to Chapter 7, but because it fits so well in here with standard deviations, you'll just have to accept it on faith at this point!

It does give us interesting information, though. Let's look at an example to illustrate.

**Example 3**

IQ tests are generally designed to have a mean of 100 and a standard deviation of 15. It's also known that the distribution of EQ scores tends to follow a bell-curve.

With that in mind, approximately what percentage of individuals have IQs between 85 and 115?

(reveal answer)

It's difficult to characterize a "genius" explicitly, but some put it at an IQ of about 145+. About what percent of the population are "geniuses" by this criteria?

(reveal answer)
This work is licensed under a Creative Commons License.
Section 3.3: Measures of Central Tendency and Dispersion from Grouped Data

3.1 Measures of Central Tendency
3.2 Measures of Dispersion
3.3 Measures of Central Tendency and Dispersion from Grouped Data
3.4 Measures of Position and Outliers
3.5 The Five-Number Summary and Boxplots

Objectives

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

1. approximate the mean and standard deviation of a variable from grouped data*
2. compute the weighted mean

* You will not be tested on this objective.

Suppose you wanted to estimate the mean and standard deviation for an exam, but all the professor gave you was curve, maybe something like this one:

<table>
<thead>
<tr>
<th>Exam 1 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>90+</td>
</tr>
<tr>
<td>80-89</td>
</tr>
<tr>
<td>70-79</td>
</tr>
<tr>
<td>60-69</td>
</tr>
<tr>
<td>50-59</td>
</tr>
</tbody>
</table>

Could you do it? How?

**Approximating the Mean from Grouped Data**

The technique we'll use (which you may have already thought of) is to treat each individual as the midpoint of its class. So instead of 13 scores from 80-89, we'll say that there are 13 85's. (This really works best with continuous data - we should probably use a midpoint of 84 for this example. Can you see why?)

From there, we should be able to approximate both the mean and standard deviation. We just have to remember to weight each observation by the number that are in that category.

**Example 1**

Using the Exam 1 data from above,

<table>
<thead>
<tr>
<th>Scores</th>
<th>Freq</th>
<th>Midpoint</th>
<th>Weighted Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>90+</td>
<td>8</td>
<td>95</td>
<td>8*95 = 760</td>
</tr>
<tr>
<td>80-89</td>
<td>13</td>
<td>85</td>
<td>13*85 = 1105</td>
</tr>
<tr>
<td>70-79</td>
<td>6</td>
<td>75</td>
<td>6*75 = 450</td>
</tr>
<tr>
<td>60-69</td>
<td>3</td>
<td>65</td>
<td>3*65 = 195</td>
</tr>
<tr>
<td>50-59</td>
<td>1</td>
<td>55</td>
<td>1*55 = 55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>31</td>
</tr>
</tbody>
</table>

2565
So the sample mean is then:
\[
\frac{2565}{31} \approx 82.7
\]
(Notice again that I’m rounding the mean to one extra decimal place.)

Let’s try the sample standard deviation:

<table>
<thead>
<tr>
<th>Scores</th>
<th>Freq</th>
<th>Mdpt</th>
<th>(x_i \cdot \bar{x})</th>
<th>((x_i - \bar{x})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>90+</td>
<td>8</td>
<td>95</td>
<td>12.3 151.29</td>
<td>8*151.29 = 1210.32</td>
</tr>
<tr>
<td>80-89</td>
<td>13</td>
<td>85</td>
<td>2.3   5.29</td>
<td>13*5.29 = 68.77</td>
</tr>
<tr>
<td>70-79</td>
<td>6</td>
<td>75</td>
<td>-7.7  59.29</td>
<td>6*59.29 = 355.74</td>
</tr>
<tr>
<td>60-69</td>
<td>3</td>
<td>65</td>
<td>-17.7 313.29</td>
<td>3*313.29 = 939.87</td>
</tr>
<tr>
<td>50-59</td>
<td>1</td>
<td>55</td>
<td>-27.7 767.29</td>
<td>1*767.29 = 767.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>31</td>
</tr>
</tbody>
</table>

The approximate sample standard deviation is then:
\[
\sqrt{\frac{111.3997}{31-1}} \approx 10.55
\]

Do you ever wonder how your GPA is calculated? This is it!

**Weighted Mean**

A weighted mean occurs when certain values carry more weight than others. The easiest example is your GPA. An "A" in Statistics counts more than an "C" in Tennis - not because it's more important or carries a higher meaning, but because the 4 credits for Statistics outweigh the 1 credit for Tennis. That's why your GPA will be closer to an "A" than a "C" - the Statistics course counts for more.

Here's how it works:

Each letter grade is assigned a weight. At most schools, this means an A=4, B=3, etc. Some schools do have other point systems, and there are many schools that have partial points with A+, B+, etc.

When calculating your GPA, the point value for each course is weighted by the number of credits. In the quick example above, your GPA for that semester wouldn't be a "B", because the Statistics course was worth 4 credits. The real GPA would be:

\[
\text{GPA} = \frac{\text{weighted points}}{\text{total credits}} = \frac{4*4 + 2*1}{4+1} = \frac{18}{5} \approx 3.6
\]

Let's try one that's more interesting.

**Example 2**

Here's a typical course load for a 1st-year student at ECC, along with some typical grades.

<table>
<thead>
<tr>
<th>Class</th>
<th>Credits</th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Course</td>
<td>Credit</td>
<td>Grade</td>
</tr>
<tr>
<td>------------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Statistics</td>
<td>4</td>
<td>B</td>
</tr>
<tr>
<td>Chemistry</td>
<td>5</td>
<td>A</td>
</tr>
<tr>
<td>Tennis</td>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>English</td>
<td>3</td>
<td>C</td>
</tr>
<tr>
<td>Speech</td>
<td>3</td>
<td>B</td>
</tr>
</tbody>
</table>

What is this student's semester GPA?

[ reveal answer ]
Section 3.4: Measures of Position

### Objectives

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

1. determine and interpret z-scores
2. determine and interpret percentiles
3. determine and interpret quartiles
4. check a set of data for outliers

In Sections 3.1 and 3.2, we discussed ways to describe a "typical" individual in a population or sample. In this next section, we'll talk about ways to describe an individual in relation to the population.

### z-Scores

**Example 1**

It's fairly common for upper-level statistics courses to have both undergraduate and graduate students. Given the exam scores listed below, can you determine which score is better relative to its peers, the undergraduate score of 83 or the graduate score of 88?

<table>
<thead>
<tr>
<th>Undergraduate student scores</th>
<th>Graduate student scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>65 89 84 75</td>
<td>82 90 95 72</td>
</tr>
<tr>
<td>52 78 92 80</td>
<td>78 88 92 89</td>
</tr>
<tr>
<td>76 72 83 79</td>
<td></td>
</tr>
</tbody>
</table>

Actually, to answer this question, we need more information. In particular, we need a new way to describe relative position.

The **z-score** represents the number of standard deviations a data value is from the mean.

\[
\begin{align*}
\text{Population z-Score:} & \quad z = \frac{x - \mu}{\sigma} \\
\text{Sample z-Score:} & \quad z \approx \frac{x - \bar{x}}{s}
\end{align*}
\]

I can't over-emphasize the importance of the meaning behind the z-score. Make a note of this now - you'll be seeing this again later on in the semester - it's very important!
Example 1 (continued)

So let's continue with our previous example. The sample mean of the undergraduate scores is 77.1, with a standard deviation of 10.73. That gives a z-score for the undergraduate 83 of:

\[ z = \frac{83 - 77.1}{10.73} \approx 0.55 \]

With a sample mean of 85.75 and a standard deviation of 7.78, the graduate has a z-score of:

\[ z = \frac{88 - 85.75}{7.78} \approx 0.29 \]

Since the undergraduate is more than 1/2 of a standard deviation above the mean (z = 0.55), that's a better relative score.

Note: You may have noticed that I went to the hundredths place for these z-scores. That's standard practice.

Key: We use z-scores when we want to compare to individuals from different populations, relative to their respective populations.

Percentiles

If you've ever taken a standardized exam like the PSAT, SAT, or ACT, you've seen in the report something about your percentile.

The kth percentile, denoted \( P_k \), of a set of data divides the lower \( k \)% of a data set from the upper \( (100-k)\)%.

Percentile ranks are used in a variety of fields:

- Special Education - students scoring below a certain percentile on specific tests qualify for services.
- Physicians - doctors usually track a child's weight and height and compare the growth to that of other children of the same age.

Unfortunately, there's no universally accepted way to calculate percentiles. Most software packages and calculators use a method similar to the one below (from your text), but you should be aware of the possibility of others.

Determining the kth percentile, denoted \( P_k \)

Step 1: Arrange the data in ascending order.
Step 2: Compute an index \( i \) using the formula

\[ i = \left( \frac{k}{100} \right) (n + 1) \]

Step 3:

a. If \( i \) is an integer, the kth percentile, denoted \( P_k \), is the ith value.

b. If \( i \) is not an integer, the kth percentile is the mean of the observations on either side of \( i \).
Example 2

Let's go back to the Jackson cousins we saw in Example 2 in Section 3.1. Suppose this time we add all the cousins, from little Zander at age 4 to Mae, who at age 18 is entering her first year at college.

Use the strategy above to find the 25th percentile by age.

Technology

Here's a quick overview of the formulas for finding percentiles in StatCrunch.

2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated
5. Enter the percentile you wish to calculate in the "Percentile" box.
6. Click "Calculate" and another window with these numbers calculated will pop up.

Note: Some software like Microsoft Excel interpolates instead of taking a simple average when calculating percentiles, so the results may differ slightly.

Determining the Percentile of a Data Value

The last thing we need to do with percentiles is to figure out the percentile of a particular individual. For example, if your Composite ACT score is a 28, what percentile does that leave you?

As before, there is no universally accepted way to calculate percentiles, but the following (from your text) is very common.

Finding the Percentile that Corresponds to a Data Value

**Step 1**: Arrange the data in ascending order.
**Step 2**: Use the following formula to find the percentile of the value, x.

\[
\text{percentile of } x = \frac{\text{number of data values less than } x}{n} \times 100
\]

Round this number to the nearest integer.
Example 3

Consider again the Jackson cousins we looked at in Example 2 above.

What is the percentile rank of James, the 14-year-old?

[ reveal answer ]

Quartiles

As the name implies, **quartiles** divide the data into four equal parts. Therefore the first quartile, $Q_1$, is the 25th percentile, the second quartile, $Q_2$ is the 50th percentile (or the median), and the third quartile, $Q_3$, is the 75th percentile.

<table>
<thead>
<tr>
<th>Minimum value</th>
<th>$Q_1$</th>
<th>$Q_2$ (median)</th>
<th>$Q_3$</th>
<th>Maximum value</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>57</td>
<td>65</td>
<td>71</td>
<td>99</td>
</tr>
<tr>
<td>69</td>
<td>73</td>
<td>78</td>
<td>79</td>
<td>96</td>
</tr>
<tr>
<td>71</td>
<td>73</td>
<td>78</td>
<td>78</td>
<td>97</td>
</tr>
<tr>
<td>73</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>97</td>
</tr>
<tr>
<td>74</td>
<td>75</td>
<td>77</td>
<td>78</td>
<td>95</td>
</tr>
<tr>
<td>78</td>
<td>77</td>
<td>78</td>
<td>79</td>
<td>96</td>
</tr>
<tr>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>97</td>
</tr>
<tr>
<td>78</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>99</td>
</tr>
</tbody>
</table>

Example 4

Let’s consider one of the sets of hypothetical exam scores we looked at in Section 3.2.

| 48 57 58 65 68 69 71 73 73 |
| 74 75 77 78 78 79 80 85   |
| 87 88 89 89 95 96 97 99   |

Find the quartiles.

[ reveal answer ]

Technology

Here’s a quick overview of the formulas for finding quartiles in StatCrunch.

1. Select **Stat > Summary Stat > Columns**.
2. Select the variable you want to summarize (e.g., "Heights")—leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated
5. Click "Calculate" and another window with these numbers calculated will pop up.

Note: Some software like Microsoft Excel interpolates instead of taking a simple average when calculating percentiles, so the results may differ slightly.
Checking for Outliers

One good use of quartiles is they give us a sense of what values might be extreme. In Statistics, we call these values **outliers**. There are various ways to check for outliers. Most depend on the distribution and often can only characterize observations as **possible** outliers. A common technique used is the following:

### Checking for Outliers by Using Quartiles

**Step 1**: Determine the first and third quartiles  
**Step 2**: Compute the inter-quartile range: \( IQR = Q_3 - Q_1 \)  
**Step 3**: Determine the fences.  
   - Lower fence = \( Q_1 - 1.5(IQR) \)  
   - Upper fence = \( Q_3 + 1.5(IQR) \)  
**Step 4**: If a value is less than the lower fence or greater than the upper fence it is considered an outlier.

### Example 5

Let's look at those same exam scores we used in Example 4.

<table>
<thead>
<tr>
<th>48</th>
<th>57</th>
<th>58</th>
<th>65</th>
<th>68</th>
<th>69</th>
<th>71</th>
<th>73</th>
<th>73</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>75</td>
<td>77</td>
<td>78</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>88</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td>99</td>
</tr>
</tbody>
</table>

Use the above method to determine if there are any outliers.

[ reveal answer ]
Section 3.5: The Five-Number Summary and Boxplots

3.1 Measures of Central Tendency
3.2 Measures of Dispersion
3.3 Measures of Central Tendency and Dispersion from Grouped Data
3.4 Measures of Position and Outliers
3.5 The Five-Number Summary and Boxplots

Objectives

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

1. compute the five-number summary
2. draw and interpret boxplots

The Five-Number Summary

The five-number summary of a set of data consists of the smallest data value, $Q_1$, the median, $Q_3$, and the largest value of the data.

Example 1

To illustrate, let’s again look at those exam scores from Example 4 in Section 3.4.

48 57 58 65 68 69 71 73 73
74 75 77 78 78 79 80 85
87 88 89 89 89 95 96 97 99

Find the five-number summary.

[ reveal answer ]

Boxplots

Using the five-number summary and the fences, we can create a new graph called a boxplot.

Drawing a Boxplot

Step 1: Determine the five-number summary and the lower and upper fences.
Step 2: Draw a horizontal line and label it with an appropriate scale.
Step 3: Draw vertical lines at $Q_1$, $M$, and $Q_3$. Enclose these vertical lines in a box.
Step 4: Draw a line from $Q_1$ to the smallest data value that is within the lower fence. Similarly, draw a line from $Q_3$ to the largest value that is within the upper fence.
Step 5: Any values outside the fences are outliers and are marked with an asterisk (*).
A typical boxplot will look something like this:

![Boxplot diagram](image)

labels on horizontal axis should be distributed evenly

**Example 2**

To illustrate, let's again look at those exam scores from Example 4 in Section 3.4.

<table>
<thead>
<tr>
<th>48</th>
<th>57</th>
<th>58</th>
<th>65</th>
<th>68</th>
<th>69</th>
<th>71</th>
<th>73</th>
<th>73</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>75</td>
<td>77</td>
<td>78</td>
<td>78</td>
<td>78</td>
<td>79</td>
<td>80</td>
<td>85</td>
</tr>
<tr>
<td>87</td>
<td>88</td>
<td>89</td>
<td>89</td>
<td>89</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td>99</td>
</tr>
</tbody>
</table>

Take a moment and try to sketch a boxplot of this data set, following the description above.

[ reveal answer ]

**Technology**

Here's a quick overview of how to create box plots in StatCrunch.

1. Enter or import the data.
2. Select **Graphics > Box Plot**.
3. Select the column(s) you want to create a box plot for.
4. Click **Next**.
5. Check "Use fences to identify outliers" and click **Next**.
6. Enter any modifications and click **Next**.
7. Choose a color scheme, if you wish, and click **Create Graph!**
8. You can then choose **Options > Copy** to copy the box plot for use elsewhere.

You can also visit the video page for links to see videos in either Quicktime or iPod format.

**Boxplots and Distribution Shape**

The last thing we want to talk about in Chapter 3 is the relationship between the shape of a boxplot and the shape of the distribution.

In **Section 2.2**, we talked about distribution shape, showing the following four standards:
Let's now see how these are related to boxplots. Here's some information from your text:

**Symmetric distributions**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁ is equally far from the median as Q₃ is</td>
<td>The median line is in the center of the box</td>
</tr>
<tr>
<td>The minimum is equally far from the median as the maximum is</td>
<td>The left whisker is equal in length to the right whisker</td>
</tr>
</tbody>
</table>

**Skewed left distributions**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min Q₁ M Q₃ Max</td>
<td>Min Q₁ M Q₃ Max</td>
</tr>
</tbody>
</table>
Skewed right distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Boxplot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁ is closer to the median than Q₃ is</td>
<td>The median line is to the left of center in the box</td>
</tr>
<tr>
<td>The minimum is closer to the median as the maximum is</td>
<td>The left whisker is shorter than the right whisker</td>
</tr>
</tbody>
</table>

Source: Instructor Resources; Statistics: Informed Decisions Using Data
Author: Michael Sullivan III
© 2007, All right reserved.
In Chapter 3, we looked at numerically summarizing data from one variable (univariate data), but newspaper articles and studies frequently describe the relationship between two variables (bivariate data). It's this second class that we'll be focusing on in Chapter 4.

There are plenty of variables which seem to be related. The links below are articles from various news sources, all discussing relationships between two variables.

Do SAT Scores Really Predict Success?
Range of Variables Affect How SAT Correlates to College GPA
Proximity to highways affects newborns' health: study
Study: Weight-loss surgery cuts cancer risk in women

Our goal in this chapter will be to find ways to describe relationships like the one between a student's SAT score and his/her GPA, and to describe the strength of that relationship.
Section 4.1: Scatter Diagrams and Correlation

Objectives

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

1. draw and interpret scatter diagrams
2. describe the properties of the linear correlation coefficient (LCC)
3. estimate the LCC based on a scatter diagram
4. compute and interpret the LCC
5. explain the difference between correlation and causation

In Chapter 3, we looked at numerically summarizing data from one variable (univariate data), but newspaper articles and studies frequently describe the relationship between two variables (bivariate data). It's this second class that we'll be focusing on in Chapter 4.

There are plenty of variables which seem to be related. The links below are articles from various news sources, all discussing relationships between two variables.

- Do SAT Scores Really Predict Success?  
- Range of Variables Affect How SAT Correlates to College GPA  
- Proximity to highways affects newborns' health: study  
- Study: Weight-loss surgery cuts cancer risk in women

In each case, there's a response variable (GPA, newborn's health, cancer levels) whose value can be explained at least in part by a predictor variable (SAT score, proximity to highways, weight-loss pill consumption).

Remember, unless we perform a designed experiment, we can only claim an association between the predictor and response variables, not a causation.

Our goal in this chapter will be to find ways to describe relationships like the one between a student's SAT score and his/her GPA, and to describe the strength of that relationship.

First, we need a new type of graph.

Scatter Diagrams

Scatter diagrams are the easiest way to graphically represent the relationship between two quantitative variables. They're just x-y plots, with the predictor variable as the x and the response variable as the y.

Example 1

The data below are heart rates of students from a Statistics I class at ECC
during the Spring semester of 2008. Students measured their heart rates (in beats per minute), then took a brisk walk and measured their heart rates again.

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
<th>before</th>
<th>after</th>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>98</td>
<td>58</td>
<td>128</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>62</td>
<td>70</td>
<td>64</td>
<td>74</td>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>52</td>
<td>56</td>
<td>74</td>
<td>106</td>
<td>66</td>
<td>70</td>
</tr>
<tr>
<td>90</td>
<td>110</td>
<td>76</td>
<td>84</td>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>66</td>
<td>76</td>
<td>56</td>
<td>96</td>
<td>78</td>
<td>116</td>
</tr>
<tr>
<td>80</td>
<td>96</td>
<td>72</td>
<td>82</td>
<td>74</td>
<td>114</td>
</tr>
<tr>
<td>78</td>
<td>86</td>
<td>72</td>
<td>78</td>
<td>90</td>
<td>116</td>
</tr>
<tr>
<td>74</td>
<td>84</td>
<td>68</td>
<td>90</td>
<td>76</td>
<td>94</td>
</tr>
</tbody>
</table>

We can see that the heart rate before going on the walk is the predictor (x), and the heart rate after the walk is the response (y).

Here's an excellent video showing a scatter diagram on steroids created by the BBC:
Technology

Here's a quick overview of the steps for creating scatter diagrams in StatCrunch.

1. Select **Graphics > Scatter plot**
2. Select quantitative variables for the X & Y axes.

You can also go to the video page for links to see videos in either Quicktime or iPod format.

### Types of Relationships

Not all relationships have to be linear, like the before/after heart rate data. The images below show some of the possibilities for the relationship (or lack thereof) between two variables.

![Linear vs. Nonlinear vs. No relation diagrams](image)

The price of a manufactured item and the profit the company gains from it, for example, do not have a linear relationship. When prices are low, sales are high, but profit is still low since very little is made from each sale. As prices increase, profits increase, but at some point, sales will start to drop, until eventually too steep of a price will drive sales down so far as to not be profitable. This might be represented by the third, "Nonlinear" image.

### Positive and Negative Association

The next thing we do is somehow quantify the strength and direction of the relationship between two variables. Here's how we'll describe the direction:
In general, we say two linearly related variables are **positively associated** if an increase in one causes an increase in the other (first "Linear" image). We say two linearly related variables are **negatively associated** if an increase in one causes a decrease in the other (second "Linear" image).

The images below show some examples of what scatter plots might look like for two positively associated variables.

<table>
<thead>
<tr>
<th>positively associated</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Scatter Plot" /></td>
</tr>
<tr>
<td>Predictor</td>
</tr>
</tbody>
</table>

And these are some examples of what scatter plots might look like for two negatively associated variables.

<table>
<thead>
<tr>
<th>negatively associated</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4" alt="Scatter Plot" /></td>
</tr>
<tr>
<td>Predictor</td>
</tr>
</tbody>
</table>

### The Linear Correlation Coefficient

As we can see from these examples, knowing the directions isn't enough - we need to quantify the strength of the relationship as well. What we'll use to do that is a new statistic called the **linear correlation coefficient**. (In this class, we'll be dealing solely with linear relationships, so we usually just call it the **correlation**.)

The **linear correlation coefficient** is a measure of the strength of the linear relationship between two variables.

\[
r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y (n - 1)}
\]

where \( \bar{x} \) is the sample mean of the predictor variable
- \( s_x \) is the sample standard deviation of the predictor variable
- \( \bar{y} \) is the sample mean of the response variable
- \( s_y \) is the sample standard deviation of the response variable
- \( n \) is the sample size
I know that's quite a mouthful, but we'll be using technology to calculate it. Here's a quick summary of some of the properties of the linear correlation coefficient, as described in your text.

**Properties of the Linear Correlation Coefficient**

1. The linear correlation coefficient is always between -1 and 1.
2. If \( r = +1 \), there is a perfect positive linear relation between the two variables.
3. If \( r = -1 \), there is a perfect negative linear relation between the two variables.
4. The closer \( r \) is to +1, the stronger is the evidence of positive association between the two variables.
5. The closer \( r \) is to -1, the stronger is the evidence of negative association between the two variables.
6. If \( r \) is close to 0, there is little or no evidence of a linear relation between the two variables - this does not mean there is no relation, only that there is no linear relation.

**Source:** Statistics: Informed Decisions Using Data  
**Author:** Michael Sullivan III  
© 2007, All right reserved.

Next, I'd like you to visit two web sites that offer Java applets. These will help you interact with data to get a sense of the linear correlation coefficient.

**Example 2**

This first applet was created for use with another textbook, *Introduction to the Practice of Statistics*, by David S. Moore and George P. McCabe.

The applet is designed to allow you to add your own points and watch it calculate the linear correlation coefficient for you. (There are other capabilities as well, but we'll get to those in the next section.)

Applet: Correlation and Regression

**Example 3**

This second applet was designed as part of the Rossman/Chance Applet Collection at California Polytechnic State University.

This applet generates scatter plots for you and asks you to guess the correlation for each. Click on "New Sample" to start, enter your answer, and then "Enter" to see if you're correct.

Applet: Guess the Correlation

**Example 4**

Let's try to calculate a correlation ourselves. To make our data set a bit more manageable, let's use the before/after data from *Example 1* in Section 4.1, but let's just use the first 8 as our sample.

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
<th>( \frac{x_i - \bar{x}}{s_x} )</th>
<th>( \frac{y_i - \bar{y}}{s_y} )</th>
<th>( \frac{y_i - \bar{y}}{s_y} ) (( \frac{x_i - \bar{x}}{s_x} ))</th>
<th>(( \frac{y_i - \bar{y}}{s_y} )) (( \frac{y_i - \bar{y}}{s_y} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>98</td>
<td>0.97865</td>
<td>0.78657</td>
<td>0.76978</td>
<td>0.76978</td>
</tr>
</tbody>
</table>
Using computer software, we find the following values:

\[
\bar{x} = 73.5 \\
s_x \approx 12.77274 \\
\bar{y} = 84.5 \\
s_y \approx 17.16308 
\]

Note: We don't want to round these values here, since they'll be used in the calculation for the correlation coefficient - only round at the very last step.

Since we have a sample size of 8, we divide the sum by 7 and get a correlation factor of 0.99. That seems fairly high, but looking at the scatter plot (below), we can see why it's so strong.

![Scatter plot](image)

**Technology**

Here's a quick overview of the formulas for finding the linear correlation coefficient in StatCrunch.

1. Select **Stat > Regression > Simple Linear**
2. Select the predictor variable for X & the response variable for Y
3. Select **Calculate**

You can also go to the [video page](#) for links to see videos in either Quicktime or iPod format.

Here's one for you to try.
Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality. The dependent variable for analysis is age adjusted mortality (called "Mortality").

The data below show the age adjusted mortality rate (deaths per 100,000) and the sulfur dioxide pollution potential. Use StatCrunch to calculate the linear correlation coefficient. Round your answer to three digits.

<table>
<thead>
<tr>
<th>City</th>
<th>Mortality*</th>
<th>SO₂ potential**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akron, OH</td>
<td>921.87</td>
<td>59</td>
</tr>
<tr>
<td>Albany, NY</td>
<td>997.87</td>
<td>39</td>
</tr>
<tr>
<td>Allentown, PA</td>
<td>962.35</td>
<td>33</td>
</tr>
<tr>
<td>Atlanta, GA</td>
<td>982.29</td>
<td>24</td>
</tr>
<tr>
<td>Baltimore, MD</td>
<td>1071.29</td>
<td>206</td>
</tr>
<tr>
<td>Birmingham, AL</td>
<td>1030.38</td>
<td>72</td>
</tr>
<tr>
<td>Boston, MA</td>
<td>934.7</td>
<td>62</td>
</tr>
<tr>
<td>Bridgeport, CT</td>
<td>899.53</td>
<td>4</td>
</tr>
<tr>
<td>Buffalo, NY</td>
<td>1001.9</td>
<td>37</td>
</tr>
<tr>
<td>Canton, OH</td>
<td>912.35</td>
<td>20</td>
</tr>
<tr>
<td>Chattanooga, TN</td>
<td>1017.61</td>
<td>27</td>
</tr>
<tr>
<td>Chicago, IL</td>
<td>1024.89</td>
<td>278</td>
</tr>
<tr>
<td>Cincinnati, OH</td>
<td>970.47</td>
<td>146</td>
</tr>
<tr>
<td>Cleveland, OH</td>
<td>985.95</td>
<td>64</td>
</tr>
<tr>
<td>Columbus, OH</td>
<td>958.84</td>
<td>15</td>
</tr>
<tr>
<td>Dallas, TX</td>
<td>860.1</td>
<td>1</td>
</tr>
<tr>
<td>Dayton, OH</td>
<td>936.23</td>
<td>16</td>
</tr>
<tr>
<td>Denver, CO</td>
<td>871.77</td>
<td>28</td>
</tr>
<tr>
<td>Detroit, MI</td>
<td>959.22</td>
<td>124</td>
</tr>
<tr>
<td>Flint, MI</td>
<td>941.18</td>
<td>11</td>
</tr>
<tr>
<td>Fort Worth, TX</td>
<td>891.71</td>
<td>1</td>
</tr>
<tr>
<td>Grand Rapids, MI</td>
<td>871.34</td>
<td>10</td>
</tr>
<tr>
<td>Greensboro, NC</td>
<td>971.12</td>
<td>5</td>
</tr>
<tr>
<td>Hartford, CT</td>
<td>887.47</td>
<td>10</td>
</tr>
<tr>
<td>Houston, TX</td>
<td>952.53</td>
<td>1</td>
</tr>
<tr>
<td>Indianapolis, IN</td>
<td>968.67</td>
<td>33</td>
</tr>
<tr>
<td>Kansas City, MO</td>
<td>919.73</td>
<td>4</td>
</tr>
<tr>
<td>Lancaster, PA</td>
<td>844.05</td>
<td>32</td>
</tr>
<tr>
<td>Los Angeles, CA</td>
<td>861.26</td>
<td>130</td>
</tr>
<tr>
<td>Louisville, KY</td>
<td>989.26</td>
<td>193</td>
</tr>
<tr>
<td>Memphis, TN</td>
<td>1006.49</td>
<td>34</td>
</tr>
<tr>
<td>Miami, FL</td>
<td>861.44</td>
<td>1</td>
</tr>
<tr>
<td>Milwaukee, WI</td>
<td>929.15</td>
<td>125</td>
</tr>
<tr>
<td>Minneapolis, MN</td>
<td>857.62</td>
<td>26</td>
</tr>
<tr>
<td>Nashville, TN</td>
<td>961.01</td>
<td>78</td>
</tr>
<tr>
<td>New Haven, CT</td>
<td>923.23</td>
<td>8</td>
</tr>
<tr>
<td>New Orleans, LA</td>
<td>1113.16</td>
<td>1</td>
</tr>
<tr>
<td>City</td>
<td>Mortality Rate</td>
<td>Mortality</td>
</tr>
<tr>
<td>-------------------</td>
<td>----------------</td>
<td>-----------</td>
</tr>
<tr>
<td>New York, NY</td>
<td>994.65</td>
<td>108</td>
</tr>
<tr>
<td>Philadelphia, PA</td>
<td>1015.02</td>
<td>161</td>
</tr>
<tr>
<td>Pittsburgh, PA</td>
<td>991.29</td>
<td>263</td>
</tr>
<tr>
<td>Portland, OR</td>
<td>893.99</td>
<td>44</td>
</tr>
<tr>
<td>Providence, RI</td>
<td>938.5</td>
<td>18</td>
</tr>
<tr>
<td>Reading, PA</td>
<td>946.19</td>
<td>89</td>
</tr>
<tr>
<td>Richmond, VA</td>
<td>1025.5</td>
<td>48</td>
</tr>
<tr>
<td>Rochester, NY</td>
<td>874.28</td>
<td>18</td>
</tr>
<tr>
<td>St. Louis, MO</td>
<td>953.56</td>
<td>68</td>
</tr>
<tr>
<td>San Diego, CA</td>
<td>839.71</td>
<td>20</td>
</tr>
<tr>
<td>San Francisco, CA</td>
<td>911.7</td>
<td>86</td>
</tr>
<tr>
<td>San Jose, CA</td>
<td>790.73</td>
<td>3</td>
</tr>
<tr>
<td>Seattle, WA</td>
<td>899.26</td>
<td>20</td>
</tr>
<tr>
<td>Springfield, MA</td>
<td>904.16</td>
<td>20</td>
</tr>
<tr>
<td>Syracuse, NY</td>
<td>950.67</td>
<td>25</td>
</tr>
<tr>
<td>Toledo, OH</td>
<td>972.46</td>
<td>25</td>
</tr>
<tr>
<td>Utica, NY</td>
<td>912.2</td>
<td>11</td>
</tr>
<tr>
<td>Washington, DC</td>
<td>967.8</td>
<td>102</td>
</tr>
<tr>
<td>Wichita, KS</td>
<td>823.76</td>
<td>1</td>
</tr>
<tr>
<td>Wilmington, DE</td>
<td>1003.5</td>
<td>42</td>
</tr>
<tr>
<td>Worcester, MA</td>
<td>895.7</td>
<td>8</td>
</tr>
<tr>
<td>York, PA</td>
<td>911.82</td>
<td>49</td>
</tr>
<tr>
<td>Youngstown, OH</td>
<td>954.44</td>
<td>39</td>
</tr>
</tbody>
</table>

* Age Adjusted Mortality (deaths per 100,000)
** Sulfer Dioxide pollution potential

Source: StatLib
Section 4.2: Least-Squares Regression

4.1 Scatter Diagrams and Correlation
4.2 Least-Squares Regression
4.3 Diagnostics on the Least-Squares Regression Line
4.4 Contingency Tables and Association

Objectives

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

1. find the least-squares regression (LSR) line
2. use the LSR line to make predictions
3. interpret the slope and y-intercept of the LSR line

Because we'll be talking about the linear relationship between two variables, we need to first do a quick review of lines.

The Slope and Y-intercept

If there's one thing we all remember about lines, it's the slope-intercept form of a line:

The slope-intercept form of a line is

\[ y = mx + b \]

where \( m \) is the slope of the line and \( b \) is the y-intercept.

Knowing the form isn't enough, though. We also need to know what each part means. Let's start with the slope. Most of us remember the slope as "rise over run", but that only helps us graph lines. What we really need to know is what the slope represents in terms of the original two variables. Let's look at an example to see if we can get the idea.
Example 1

The equation $T = 6x + 53$ roughly approximates the tuition per credit at ECC since 2001. In this case, $x$ represents the number of years since 2001 and $T$ represents the tuition amount for that year.

The graph below illustrates the relationship.

![Graph illustrating tuition per credit at ECC from 2001 to 2007.](image)

In this example, we can see that both the 6 and the 53 have very specific meanings:

The 6 is the *increase per year*. In other words, for every additional year, the tuition increases $6.

The 53 represents the *initial tuition*, or the tuition per credit hour in 2001.

As we progress into the relationship between two variables, it's important to keep in mind these meanings behind the slope and $y$-intercept.

Finding the Equation for a Line

Another very important skill is finding the equation for a line. In particular, it's important for us to know how to find the equation when we're given two points.

A very useful equation to know is the point-slope form for a line.

The **point-slope form** of a line is

$$y - y_1 = m(x - x_1)$$

where $m$ is the slope of the line and $(x_1, y_1)$ is a point on the line.

Let's practice using this form to find an equation for the line.
Example 2

In Example 1 from section 4.1, we talked about the relationship between student heart rates (in beats per minute) before and after a brisk walk.

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>98</td>
</tr>
<tr>
<td>62</td>
<td>70</td>
</tr>
<tr>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>90</td>
<td>110</td>
</tr>
<tr>
<td>66</td>
<td>76</td>
</tr>
<tr>
<td>80</td>
<td>96</td>
</tr>
<tr>
<td>78</td>
<td>86</td>
</tr>
<tr>
<td>74</td>
<td>84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>128</td>
</tr>
<tr>
<td>64</td>
<td>74</td>
</tr>
<tr>
<td>74</td>
<td>106</td>
</tr>
<tr>
<td>76</td>
<td>84</td>
</tr>
<tr>
<td>56</td>
<td>96</td>
</tr>
<tr>
<td>72</td>
<td>82</td>
</tr>
<tr>
<td>72</td>
<td>78</td>
</tr>
<tr>
<td>68</td>
<td>90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>66</td>
<td>70</td>
</tr>
<tr>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>78</td>
<td>116</td>
</tr>
<tr>
<td>74</td>
<td>114</td>
</tr>
<tr>
<td>90</td>
<td>116</td>
</tr>
<tr>
<td>76</td>
<td>94</td>
</tr>
</tbody>
</table>

Let's highlight a pair of points on that plot and use those two points to find an equation for a line that might fit the scatter diagram.

Using the points (52, 56) and (90, 116), we get a slope of

\[
m = \frac{116-56}{90-52} = \frac{60}{38} \approx 1.58
\]

So an equation for the line would be:

\[
y - y_1 = m(x - x_1)
\]

\[
y - 56 = 1.58(x - 52)
\]

\[
y = 1.58x - 26.16
\]

It's interesting to note the meanings behind the slope and y-intercept for this example. A slope of 1.58 means that for every additional beat per minute before the brisk walk, the heart rate after the walk was 1.58 faster.

The y-intercept, on the other hand, doesn't apply in this case. A y-intercept of -26.16 means that if you have 0 beats per minute before the walk, you'll have -26.16 beats per minute after the walk.?!?!?

This brings up a very important point - models have limitations. In this case, we say that the y-intercept is outside the scope of the model.

Now that we know how to find an equation that sort of fits the data, we need a strategy to find the best line. Let's work our way up to it.

Residuals

Unless the data line up perfectly, any line we use to model the relationship will have an error. We call this error the residual.
The **residual** is the difference between the observed and predicted values for $y$:

\[
\text{residual} = \text{observed } y - \text{predicted } y \\
\text{residual} = y - \hat{y}
\]

Notice here that we used the symbol $\hat{y}$ (read "$y$-hat") for the predicted. This is standard notation in statistics, using the "hat" symbol over a variable to note that it is a predicted value.

**Example 3**

Let's again use the data from Example 1 from section 4.1. In Example 2 from earlier this section, we found the model:

\[
\hat{y} = 1.58x - 30.16
\]

Let's use this model to predict the "after" heart rate for a particular student, the one whose "before" heart rate was 86 beats per minute.

The predicted heart rate, using the model above, is:

\[
\hat{y} = 1.58(86) - 26.16 = 109.72
\]

Using that predicted heart rate, the residual is then:

\[
\text{residual} = y - \hat{y} = 98 - 109.72 = -11.72
\]

Here's that residual if we zoom in on that particular student:

\[
\text{residual} = 98 - 109.72 = -11.72
\]

Notice here that the residual is negative, since the predicted value was more than the actual observed "after" heart rate.
The Least-Squares Regression (LSR) line

So how do we determine which line is "best"? The most popular technique is to make the sum of the squares of the residuals as small as possible. (We use the squares for much the same reason we did when we defined the variance in Section 3.2.) The method is called the method of least squares, for obvious reasons!

The Equation for the Least-Squares Regression line

The equation of the least-squares is given by
\[ \hat{y} = b_1x + b_0 \]
where
\[ b_1 = r \cdot \frac{s_y}{s_x} \] is the slope of the least-squares regression line
and
\[ b_0 = \bar{y} - b_1 \bar{x} \] is the y-intercept of the least squares regression line

Let's try an example.

Example 4

Let's again use the data from Example 1 in Section 4.1, but instead of just using two points to get a line, we'll use the method of least squares to find the Least-Squares Regression line.

<table>
<thead>
<tr>
<th>before</th>
<th>after</th>
<th>before</th>
<th>after</th>
<th>before</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>86</td>
<td>98</td>
<td>58</td>
<td>128</td>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>62</td>
<td>70</td>
<td>64</td>
<td>74</td>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>52</td>
<td>56</td>
<td>74</td>
<td>106</td>
<td>66</td>
<td>70</td>
</tr>
<tr>
<td>90</td>
<td>110</td>
<td>76</td>
<td>84</td>
<td>80</td>
<td>92</td>
</tr>
<tr>
<td>66</td>
<td>76</td>
<td>56</td>
<td>96</td>
<td>78</td>
<td>116</td>
</tr>
<tr>
<td>80</td>
<td>96</td>
<td>72</td>
<td>82</td>
<td>74</td>
<td>114</td>
</tr>
<tr>
<td>78</td>
<td>86</td>
<td>72</td>
<td>78</td>
<td>90</td>
<td>116</td>
</tr>
<tr>
<td>74</td>
<td>84</td>
<td>68</td>
<td>90</td>
<td>76</td>
<td>94</td>
</tr>
</tbody>
</table>
Using computer software, we find the following values:

\[ \bar{x} \approx 72.16667 \]
\[ s_x \approx 10.21366 \]
\[ \bar{y} = 90.75 \]
\[ s_y \approx 17.78922 \]
\[ r \approx 0.48649 \]

Note: We don't want to round these values here, since they'll be used in the calculation for the correlation coefficient - only round at the very last step.

Using the formulas for the LSR line, we have

\[ y = 0.8473x + 29.6018 \]

(A good general guideline is to use 4 decimal places for the slope and y-intercept, though there is no strict rule.)

One thought that may come to mind here is that this doesn't really seem to fit the data as well as the one we did by picking two points! Actually, it does do a much better job fitting ALL of the data as well as possible - the previous line we did ourselves did not address most of the points that were above the main cluster. In the next section, we'll talk more about how outliers like the (58, 128) point far above the rest can affect a model like this one.
Here's a quick overview of how to find the Least-Squares Regression line in StatCrunch.

1. Select Stat > Regression > Simple Linear
2. Select the predictor variable for X & the response variable for Y
3. Select Calculate

The fourth line shows the equation of the regression line. Note that it will not have x and y shown, but rather the names that you've given for x and y. For example:

Avg. Final Grade = 88.73273 - 2.8272727 Num. Absences

You can also go to the video page for links to see videos in either Quicktime or iPod format.

This work is licensed under a Creative Commons License.
Section 4.3: Diagnostics on the Least-Squares Regression Line

4.1 Scatter Diagrams and Correlation
4.2 Least-Squares Regression
4.3 Diagnostics on the Least-Squares Regression Line
4.4 Contingency Tables and Association

Objectives

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

1. compute and interpret the coefficient of determination
2. perform residual analysis on a regression model
3. determine if a linear regression model is appropriate
4. identify influential observations

Before we can go on, we need to first determine two things:

1. Does our model do a good job of predicting the results?
2. Is a linear model appropriate?

We'll have two key factors to help us answer these questions. The first is called the coefficient of determination.

The Coefficient of Determination

The coefficient of determination, $R^2$, is the percent of the variation in the response variable (y) that can be explained by the least-squares regression line.

Looking at the definition, we can see that a higher $R^2$ is better - the LSR line does a better job of explaining the variation in the response variable.

Your textbook has a relatively detailed explanation of how $R^2$ is calculated, so I won't repeat it here. If you'd like to, you can also view the derivation on this page from Wikipedia. (Not that Wikipedia is a perfect source, but this particular page is accurate.)

Here's the end result, which shouldn't come as too much of a surprise:

$$R^2 = r^2$$

Here are some examples:

- $R^2 = 92.1\%$
- $R^2 = 70.0\%$
The second step in residual analysis is using the residuals to determine if a linear model is appropriate. We do this by creating residual plots. A residual plot is a scatter diagram with the predictor as the x and the corresponding residual as the y.

In general, there are three things to watch out for in a residual plot:

1. a pattern in the residuals
2. increasing or decreasing spread
3. influential observations

**Patterned Residuals**

If a residual plot shows a discernable pattern (like a curve), then the predictor and response variables may not be linearly related.
The LSR line clearly does not fit. The residuals show an obvious pattern.

**Increasing or Decreasing Spread**

If a residual plot shows the spread increasing or decreasing, then a strict requirement of the linear model is related. (This strict requirement is called *constant error variance* - the error must be evenly spread.)

![Residual plot with increasing spread](image)

The LSR line seems to be a great fit. The residuals start very small, but increase as the predictor variable increases - this model does not have constant error variance.

**Outliers and Influential Observations**

The next point we need to consider is the existence of outliers and influential observations. We can think of an **outlier** as an observation that doesn’t seem to fit the rest of the data. **Influential observations** are similar, but with the added quality that their existence significantly affects the slope and/or y-intercept of the line.

Consider the scatter diagram shown below, along with its corresponding residual plot:

![Scatter plot with outliers](image)

Let's consider the three cases indicated.

**Case 1:**

This case is considered an outlier because it's x-value is much lower than all but one of the other observations. To determine if it's an influential observation, we'll need to recalculate the LSR line without that observation included. Here are the results:
We can see that while there are some changes in the slope and y-intercept, both are reasonably similar to what they were with Case 1 included. In this case, we would describe Case 1 as an outlier, but not an influential observation.

An interesting point to note, though, is the decreased $R^2$ value. The implication is that Case 1 actually strengthened the correlation. Think of that point pulling the line “tighter”.

**Case 2:**

Looking back at the original diagram, it seems as though Case 2 should be influential, because there are not many values near it to minimize its effect on the LSR line. Here’s the output from computer software:

Here we can clearly see that both the slope and y-intercept (as well as $R^2$) are significantly different, so we would definitely characterize Case 2 as an influential observation.

**Case 3:**

Unlike in Case 2, this particular observation has others near it to minimize its effect, so it most likely will not be influential. Here’s the output from computer software:
Comparing those to the original values, we can indeed see that the slope and y-intercept are both relatively similar. So while this value is an outlier (as seen very clearly on the earlier residual plot), it is not influential.

If you find your data contain outliers or influential observations, and those observations cannot be removed (because they are due to data entry errors or similar) you have only a couple options. The primary option is to collect more data to minimize their impact. The second is to use analysis methods that minimize the effect of outliers. Unfortunately, those techniques are fairly advanced and outside the scope of this course.

**When a Linear Model is Appropriate**

Sometimes it can be difficult to determine if any of the three above conditions have been violated, but here's a good example of a situation where a linear model does seem appropriate.

The LSR line seems to fit the data. The residuals are evenly spread above and below zero, there is no discernable pattern, and there are no outliers.

**Technology**

Here's a quick overview of how to create a residual plot in StatCrunch.

1. Select **Stat > Regression > Simple Linear**
2. Set the X-Variable and Y-Variable and press **Next**.
3. Select **Save residuals** (optional) and press **Next**.
4. Select the options you want - make sure to select "Residuals vs. X-values" is the residual plot.
5. Press **Calculate**.
6. The output will show your regression analysis. On the bottom, press Next to see any graphics.

You can also go to the video page for links to see videos in either Quicktime or iPod format.
Section 4.4: Contingency Tables and Association

4.1 Scatter Diagrams and Correlation
4.2 Least-Squares Regression
4.3 Diagnostics on the Least-Squares Regression Line
4.4 Contingency Tables and Association

Objectives

By the end of this lesson, you will be able to...
1. compute the marginal distribution of a variable
2. construct a conditional distribution of a variable
3. use the conditional distribution to identify association between categorical data

In sections 4.1-4.3, we studied relationships between two quantitative variables. We learned that we could quantify the strength of the linear relationship with the correlation.

What about qualitative (categorical) variables, though? For example, suppose we consider a survey given to 82 students in a Basic Algebra course at ECC, with the following responses to the statement "I enjoy math."

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>9</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Women</td>
<td>12</td>
<td>18</td>
<td>11</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

How do we study this relationship? Is there a way to tell if gender and whether the student enjoys math? In fact, there is! Like usual, though, we need a bit of background work first.

Contingency Tables

A contingency table relates two categories of data. In the example above, the relationship is between the gender of the student and his/her response to the question.

A marginal distribution of a variable is a frequency or relative frequency distribution of either the row or column variable in the contingency table.

Example 1

If we consider the previous example:

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>9</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Women</td>
<td>12</td>
<td>18</td>
<td>11</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

The entire table is referred to as the contingency table.

The marginal distribution for gender removes the effect of whether or not the student enjoys math:
Whereas, the **marginal distribution** for whether or not the student enjoys math removes the effect of gender:

![Marginal Distribution Table](image)

We can also create a **relative frequency marginal distribution**, which, as expected, is simply relative frequencies rather than frequencies.

**Example 2**

The combined **relative frequency marginal distributions** would look like this:

![Relative Frequency Table](image)

Let's consider the frequency marginal distributions from Example 2.

**Example 3**

![Example 3 Table](image)

We might now be interested in comparing the two variables. For example:

a. What proportion of women strongly agreed with the statement "I enjoy math"?
   - Solution: $\frac{12}{52} \approx 0.23$

b. What proportion of women disagreed?
   - Solution: $\frac{6}{52} \approx 0.12$

c. What proportion of men were neutral?
   - Solution: $\frac{5}{30} \approx 0.17$

d. What proportion of men strongly agreed?
   - Solution: $\frac{9}{30} \approx 0.30$
If we completed the table in this fashion, we get something called a **conditional distribution**.

A **conditional distribution** lists the relative frequency of each category of variable, given a specific value of the other variable in the contingency table.

For another explanation of marginal and conditional distributions, watch this YouTube video:

---

**Example 4**

The conditional distribution of how the students feel about math by gender would be as follows:

<table>
<thead>
<tr>
<th></th>
<th>SA</th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>SD</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men</strong></td>
<td>9/30</td>
<td>13/30</td>
<td>5/30</td>
<td>2/30</td>
<td>1/30</td>
<td>30/30</td>
</tr>
<tr>
<td></td>
<td>≈ 0.30</td>
<td>≈ 0.43</td>
<td>≈ 0.17</td>
<td>≈ 0.07</td>
<td>≈ 0.03</td>
<td>= 1</td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td>12/52</td>
<td>18/52</td>
<td>11/52</td>
<td>6/52</td>
<td>5/52</td>
<td>52/52</td>
</tr>
<tr>
<td></td>
<td>≈ 0.23</td>
<td>≈ 0.35</td>
<td>≈ 0.21</td>
<td>≈ 0.12</td>
<td>≈ 0.10</td>
<td>= 1</td>
</tr>
</tbody>
</table>

Note: The row totals sometimes do not add up to 1 due to rounding.

Another way to think of this distribution is that it's the distribution of how students feel for each gender. That's what the "by gender" indicates.

---

**Example 5**

The conditional distribution of gender by how the student feels would be:

<table>
<thead>
<tr>
<th></th>
<th>SA</th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>SD</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Men</strong></td>
<td>9/21</td>
<td>13/31</td>
<td>5/16</td>
<td>2/8</td>
<td>1/6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>≈ 0.43</td>
<td>≈ 0.42</td>
<td>≈ 0.31</td>
<td>≈ 0.25</td>
<td>≈ 0.17</td>
<td></td>
</tr>
<tr>
<td><strong>Women</strong></td>
<td>12/21</td>
<td>18/31</td>
<td>11/16</td>
<td>6/8</td>
<td>5/6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>≈ 0.57</td>
<td>≈ 0.58</td>
<td>≈ 0.69</td>
<td>≈ 0.75</td>
<td>≈ 0.83</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>21/21</td>
<td>31/31</td>
<td>16/16</td>
<td>8/8</td>
<td>6/6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>= 1</td>
<td>= 1</td>
<td>= 1</td>
<td>= 1</td>
<td>= 1</td>
<td></td>
</tr>
</tbody>
</table>
Using Conditional Distributions to Identify Association

One thing we can use conditional distributions for is to identify an association between qualitative variables. The best way to do this is a side-by-side bar graph. We'll illustrate with the same data we've been using.

Example 6

In Example 5, we found the conditional distribution of gender by how the student feels regarding math:

<table>
<thead>
<tr>
<th></th>
<th>SA</th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>9/21</td>
<td>13/31</td>
<td>5/16</td>
<td>2/8</td>
<td>1/6</td>
</tr>
<tr>
<td></td>
<td>≈ 0.43</td>
<td>≈ 0.42</td>
<td>≈ 0.31</td>
<td>= 0.25</td>
<td>≈ 0.17</td>
</tr>
<tr>
<td>Women</td>
<td>12/21</td>
<td>18/31</td>
<td>11/16</td>
<td>6/8</td>
<td>5/6</td>
</tr>
<tr>
<td></td>
<td>≈ 0.57</td>
<td>≈ 0.58</td>
<td>≈ 0.69</td>
<td>= 0.75</td>
<td>≈ 0.83</td>
</tr>
<tr>
<td>Total</td>
<td>21/21</td>
<td>31/31</td>
<td>16/16</td>
<td>8/8</td>
<td>6/6</td>
</tr>
<tr>
<td></td>
<td>= 1</td>
<td>= 1</td>
<td>= 1</td>
<td>= 1</td>
<td>= 1</td>
</tr>
</tbody>
</table>

Since it's difficult to gain much from this table alone, a good way to analyze this would be to make a side-by-side bar graph.

From the graph, we can see that there definitely appear to be some differences between the different responses. The proportions of responses were similar for both "Strongly Agree" and "Agree", but very different for "Neutral" and "Disagree". As we go down the scale, the proportion of the responses that are by women increases.

We might conclude, then, that men tend to enjoy math more than women.

One thing we can't conclude is that their gender caused them to not enjoy math. We've only done an observational study, so we can only claim association, not causation.

One question you might have as a result of this is, "How do we know when it's different enough from equal to say
that there might be a relationship?" It's a very good question. In order to draw a fine line, we'll need a
hypothesis test, which we won't see until we get to Chapter 12.
Chapter 5: Probability

5.1 Probability Rules
5.2 The Addition Rule and Complements
5.3 Independence and the Multiplication Rule
5.4 Conditional Probability and the General Multiplication Rule
5.5 Counting Techniques
5.6 Putting It Together: Probability

In Chapter 5, we step away from data for a while. We take a look at a new topic for us - **probability**. Most of us have an idea already of what probability is, but we'll spend quite a while exploring different probability experiments (like rolling two dice) and investigating the different outcomes.

We'll learn several different rules, ranging from the probability that at least one of two events occurs in Section 5.2 (the Addition Rule), to the probability that both occur in Section 5.3 (the Multiplication Rule), to the probability that one occurs if we know the first has already occurred in Section 5.4 (conditional probability).

In Section 5.5, we learn some new counting techniques that'll help us answer questions like "How many 4-digit garage door codes are possible if digits can't be repeated once used?"

Section 5.6 brings it all together, and helps you choose which strategy to apply.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::
Section 5.1: Probability Rules

5.1 Probability Rules
5.2 The Addition Rule and Complements
5.3 Independence and the Multiplication Rule
5.4 Conditional Probability and the General Multiplication Rule
5.5 Counting Techniques
5.6 Putting It Together: Probability

Objectives

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

1. apply the rules of probability
2. compute and interpret probabilities using the empirical method
3. compute and interpret probabilities using the classical method
4. recognize and interpret subjective probabilities

Probability

So what is probability? Most of us already have an idea. We all know that the probability of heads when flipping a fair coin is 1/2, but what does that mean?

- One out of every two flips will be heads?
- If we have two heads in a row, the next two must be tails?
- Exactly 50 out of every 100 flips will be heads?

In fact, none of these are correct!

In general, **probability** is a measure of the likelihood of some **outcome**. We use it not to describe what will happen in one particular **event**, but rather, what the long-term proportion that outcome will occur.

So the key here isn’t that 1 out of every 2 coin flips, or 50 out of every 100 coin flips will be heads, but that over the long term, about 1/2 will be.

This concept is called the **Law of Large Numbers**. The image below is from [Wikipedia](https://en.wikipedia.org/wiki/Law_of_Large_Numbers), and shows the idea. It’s demonstrating rolling a fair 6-sided die, and calculating the average number. We know that all 6 are equally likely, so the average should be \((1+2+3+4+5+6)/6 = 3.5\). From the image, we can see that while it isn’t 3.5 initially, it does **tend** toward that as the number of rolls increases.
Probability Experiments

There are some key terms we should outline before going much further.

In probability, an **experiment** is any process where the results are uncertain. We call the **sample space**, S, the collection of all possible outcomes. A probability **event** is any collection of outcomes from the experiment.

**Example 1**

Suppose we have a family with three children, and we consider the sex of those three children. If we let B represent a boy and G represent a girl, here is the sample space:

\[ S = \{ \text{BBB, BBG, BGB, GBB, GGG, GBG, GGB, GGG} \} \]

There are thus 8 **outcomes** in this experiment.

One possible event might be:

\[ E = \text{the family has exactly two girls} = \{ \text{BGG, GBG, GGB} \} \]

**Example 2**

Suppose a fair six-sided die is rolled twice.

What is the sample space?

[ reveal answer ]

Source: stock.xchng
Example 3

If we define the event \( E \) = the sum of the two dice is more than 10, what outcomes are in \( E \)?

[ reveal answer ]

### Unusual Events

Unusual Events

Suppose we consider the previous example about rolling two dice. The probability of having the sum of the two dice be more than 10 would be 3/36 or 1/12. Is this unusual? On average, it will occur about 1 in 12 times. Is that unusual enough? We have to be careful when we characterize an event as unusual.

Typically, we say that an event with a probability less than 5% is unusual, but this isn't a hard cutoff. It depends on the context.

Suppose we're planning on making a decision one way, unless the probability of a particularly "unusual" event is too high.

One example might be the jury in a capital case, punishable by death. In this example, jurors need to be sure "beyond a reasonable doubt" that the defendant is guilty. If they decide to convict if they're 95% sure, this means that the "unusual" event that they're wrong has a probability of 5%. If you're that defendant, that's definitely not "unusual" enough!

On the other hand, suppose we're planning a picnic on a nice summer day. If the risk of a rain shower isn't too high, we'll plan on the picnic. In this case, we might set our cutoff at 20% - anything less than that is too unusual (or unlikely) to happen, so we'll risk it.

### Calculating Probabilities Using the Empirical Method

Calculating Probabilities Using the Empirical Method

There are two primary methods for calculating probabilities. The first is to simply look at what has happened in the past and assume the probability is the same as the relative frequency of that particular outcome. This is called the empirical probability of that event.

\[
P(E) \approx \text{relative frequency of } E = \frac{\text{frequency of } E}{\text{total number of trials}}
\]

#### Example 3

During the Spring semester of 2008, 33 out of 60 students from two sections of Mth096 Basic Algebra at ECC were successful. (Successful here is earning a C or better.)

If we define \( E \) = a Mth096 student is successful, \( P(E) \approx 33/60 = 0.55 \).

#### Example 4

Suppose we give a survey to 52 Basic Algebra students, asking them to rate various statements from Strongly Agree to Strongly Disagree. If 18 of the students responded with either Agree or Strongly Agree to the statement "I enjoy math.", what is the probability that a randomly selected Basic Algebra student will enjoy math?
Computing Probabilities Using the Classical Method

The second primary method for calculating probabilities is the **classical method**. The key for this method is to assume that *all outcomes are equally likely*. This is how we know the probability of rolling a 6 on a fair six-sided die is 1/6, because we assume all of the outcomes (1, 2, 3, 4, 5, and 6) are equally likely.

\[
P(E) = \frac{\text{number of ways } E \text{ can occur}}{\text{total number of possible outcomes}} = \frac{N(E)}{N(S)}
\]

**Example 5**

Let's consider again the probability experiment from Example 2 - rolling two fair six-sided dice.

Let the event \( E \) = the sum of the two dice is 6.

Find \( P(E) \).

[ reveal answer ]
Section 5.2: The Addition Rule and Complements

5.1 Probability Rules
5.2 The Addition Rule and Complements
5.3 Independence and the Multiplication Rule
5.4 Conditional Probability and the General Multiplication Rule
5.5 Counting Techniques
5.6 Putting It Together: Probability

Objectives

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

1. use the Addition Rule for disjoint events
2. use the General Addition Rule
3. use the Complement Rule

The Additional Rule for Disjoint Events

We're going to have quite a few rules in this chapter about probability, but we'll start small. The first situation we want to look at is when two events have no outcomes in common. We call events like this disjoint events.

Two events are disjoint if they have no outcomes in common. (Also commonly known as mutually exclusive events.)

Back in 1881, John Venn developed a great way to visualize sets. As is often the case in mathematics, the diagrams took on his name and have since taken on his name - Venn diagrams. Because events are sets of outcomes, it works well to visualize probability as well. Here's an example of a Venn diagram showing two disjoint outcomes, E and F.

Let's continue this a little further and put points on the chart like this - ● - to indicate outcomes.

Looking at the picture, we can clearly see that \( P(E) = 5/15 = 1/3 \), since there are 5 outcomes in E, and 15 total
Example 1 outcomes. Similarly, \( P(F) \) also is 1/3.

Next, we want to consider all of the events that are in either \( E \) or \( F \). In probability, we call that event \( E \text{ or } F \). So in our example, \( P(E \text{ or } F) = 10/15 = 2/3 \).

But, we could just see that from the picture! Just count the dots that are \( E \) and add to it the number of dots in \( F \).

In general, we can create a rule. We'll call it...

**The Addition Rule for Disjoint Events**

If \( E \) and \( F \) are disjoint (mutually exclusive) events, then

\[
P(E \text{ or } F) = P(E) + P(F)
\]

OK - time for an example. Let's use the example from last section about the family with three children, and let's define the following events:

\( E = \) the family has exactly two boys

\( F = \) the family has exactly one boy

Describe the event "\( E \) or \( F \)" and find its probability.

Of course, there are often cases when two events do have outcomes in common, so we'll need a more robust rule for that case.

**General Addition Rule**

What happens when two events do have outcomes in common? Well, let's consider the example below. In this case, \( P(E) = 4/10 = 2/5 \), and \( P(F) = 5/10 = 1/2 \), but \( P(E \text{ or } F) \) isn't 9/10. Can you see why?
The key here is the two outcomes in the middle where $E$ and $F$ overlap. Officially, we call this region the event $E$ and $F$. It's all the outcomes that are in both $E$ and $F$. In our visual example:

In this case, to find $P(E \text{ or } F)$, we'll need to add up the outcomes in $E$ with the outcomes in $F$, and then subtract the duplicates we counted that are in $E$ and $F$. We call this the General Addition Rule.

**The General Addition Rule**

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

Let's try a couple quick examples.

**Example 2**

Let's consider a deck of standard playing cards.

Suppose we draw one card at random from the deck and define the following events:

- $E =$ the card drawn is an ace
- $F =$ the card drawn is a king

Use these definitions to find $P(E \text{ or } F)$. 

Example 3

Considering the deck of playing cards, where one is drawn at random. Suppose we define the following events:

F = the card drawn is a king
G = the card drawn is a heart

Use these definitions to find P(F or G).

So the key idea and the difference between these two examples - when you're finding P(E or F), be sure to look for outcomes that E and F have in common.

The Complement Rule

I think the best way to introduce the last idea in this section is to consider an example. Let's look at a deck of standard playing cards again:

And let's define event E = a card less than a King is drawn. If I ask you to find P(E), you're not going to count them up. (You weren't going to, were you?!) No - you'll say there are 52 cards all together, and there are 4 kings, so therefore there must be 48 cards less than a King. So P(E) = 48/52 = 12/13.

The idea that you're already using there is called the complement. (That's complement, with an e. Not compliment, as in "My, you look pretty today!")

The complement of E, denoted Ec, is all outcomes in the sample space that are not in E.

So essentially, the complement of E is everything but the outcomes in E. In fact, some texts actually write it as "not E."

How is the complement helpful? Well, you actually already used the key idea in the example above. Let's look at a Venn diagram.
From Section 5.1, we know that \( P(S) = 1 \). Clearly, \( E \) and \( E^c \) are disjoint, so 
\( P(E \text{ or } E^c) = P(E) + P(E^c) \). Combining those two facts, we get:

**The Complement Rule**

\[
P(E) + P(E^c) = 1
\]

Keep this in mind when you're looking at an event that's fairly complicated. Sometimes it's much easier to find the probability of the complement.
Section 5.3: Independence and the Multiplication Rule

5.1 Probability Rules
5.2 The Addition Rule and Complements
5.3 Independence and the Multiplication Rule
5.4 Conditional Probability and the General Multiplication Rule
5.5 Counting Techniques
5.6 Putting It Together: Probability

Objectives

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

1. identify independent events
2. use the Multiplication Rule for independent events
3. compute "at least" probabilities

Independence

One of the most important concepts in probability is that of independent events.

Two events E and F are independent if the occurrence of event E does not affect the probability of event F.

Let's look at a couple examples.

Example 1

Consider the experiment where two cards are drawn without replacement. (Without replacement means one is drawn and then the second is drawn without putting the first one back.) Define events E and F this way:

E = the first card drawn is a King
F = the second card drawn is a King

Are events E and F independent?

[reveal answer]

Example 2

Consider the experiment in which two fair six-sided dice are rolled, and define events E and F as follows:

E = the first die is a 3
F = the second die is a 3

Are events E and F independent?

[reveal answer]
Disjoint vs. Independent

It is very common for students to confuse the concepts of disjoint (mutually exclusive) events with independent events. Recall from the last section:

Two events are **disjoint** if they have no outcomes in common. (Also commonly known as **mutually exclusive** events.)

Here's a Venn diagram of two disjoint events.

Looking at this image, we can see very clearly that if event E occurs (that is, the outcome \( \bigcirc \) is in event E), it cannot possibly be in event F. So E and F are **dependent**, since the occurrence of event E made event F impossible.

The Multiplication Rule for Independent Events

To introduce the next idea, let's look at the experiment from Example 2, in Section 5.1.

**Example 3**

The experiment was rolling a fair six-sided die twice. Suppose define the event E:

\[ E = \text{both dice are 2's} \]

The possible outcomes are (1,1), (1,2), (1,3), ... (6,5), and (6,6). Since only one of these is (2,2), we know \( P(E) = 1/36 \). Let's look at it another way, though.

\[
P(E) = \frac{N(E)}{N(S)} = \frac{1}{36} = \frac{1}{6 \cdot 6} = \frac{1}{6} \cdot \frac{1}{6} = P(E) \cdot P(F)
\]
In fact, this will always be true if E and F are independent.

**Multiplication Rule for Independent Events**

If E and F are independent events, then

\[ P(E \text{ and } F) = P(E) \cdot P(F) \]

---

**Example 4**

According to data from the American Cancer Society, about 1 in 3 women living in the U.S. will have some form of cancer during their lives.

If three women are randomly selected, what is the probability that they will all contract cancer at some point during their lives?

[reveal answer]

---

**At-least Probabilities**

The phrase "at least" can make a seemingly simple problem much more difficult. For example, suppose we're looking at cancer rates in women. And suppose we have a random sample of 5 women. If we're looking for the probability that at least one will have some form of cancer, that's really:

\[ P(1 \text{ will have cancer}) + P(2 \text{ will have cancer}) + \ldots + P(5 \text{ will have cancer}) \]

Instead, it's much easier to use the **Complement Rule**, from Section 5.2.

**The Complement Rule**

\[ P(E) + P(E^c) = 1 \]

---

In our example, the complement of at least one will have cancer is none will have cancer. So P(at least one will have cancer) = 1 - P(none will have it) = 1 - (1/3)^5 \approx 0.9959.

Much easier! Keep this idea of at least probabilities and the Complement Rule in mind when you're looking at cases like this.
This work is licensed under a Creative Commons License.
Section 5.4: Conditional Probability and the General Multiplication Rule

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

1. compute conditional probabilities
2. use the General Multiplication Rule
3. determine the independence of events

Conditional Probability

Remember in Example 3, in Section 5.3, about rolling two dice? In that example, we said that events E (the first die is a 3) and F (the second die is a 3) were independent, because the occurrence of E didn't affect the probability of F. Well, that won't always be the case, which leads us to another type of probability called conditional probability.

Conditional Probability

The notation $P(F|E)$ is read "the probability of F given E" and represent the probability that event F occurs, given that event E has already occurred.

Let's look again at Example 1 from that same section.

Example 1

Consider the experiment where two cards are drawn without replacement. (Without replacement means one is drawn and then the second is drawn without putting the first one back.) Define events E and F this way:

$E =$ the first card drawn is a King

$F =$ the second card drawn is a King

Find $P(F|E)$.

[reveal answer]

It can be helpful again to look at a Venn diagram to illustrate the idea. Let's look at this one that we used back in section 5.2.

Example 2
Find $P(E|F)$.

One more example.

**Example 3**

Let's consider a survey given to 52 students in a Basic Algebra course at ECC, with the following responses to the statement "I enjoy math."

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>6</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>8</td>
<td>14</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>24</td>
<td>10</td>
<td>4</td>
<td>0</td>
<td>52</td>
</tr>
</tbody>
</table>

What is the probability that a student enjoys math (Agree or Strongly Agree) given that the student is a female?

**The Monty Hall Problem**

An interesting example of conditional probability is the classic **Monty Hall Problem**. This is based on an old game show, where the host would show three doors. Behind one was a new car, and behind the others were goats. The twist was, once you made your choice, Monty would open one of the other doors showing a goat. The question then - **should you switch?** The answer is different from what you would think. Here's another video from Clive Rix at the University of Leicester in Leicester, England:
Don't believe it? Try this interactive feature from the New York Times, or watch this video from the show Numb3rs.

Wow - you never know where conditional probability can be applied!

**The General Multiplication Rule**

Let's look again at the experiment from Example 1 in Section 5.3.

**Example 4**

Consider the experiment where two cards are drawn without replacement. (Without replacement means one is drawn and then the second is drawn without putting the first one back.) Define events E and F this way:

E = the first card drawn is a King

F = the second card drawn is a King

How would we find $P(E \text{ and } F)$?

We know from Example 1 that E and F are not independent, so we know we can't use the Multiplication Rule for Independent Events. It's probably not too difficult to see how we might do it, though.

$$P(E \text{ and } F) = P(\text{first is King and second is King})$$

$$= P(\text{first is King}) \cdot P(\text{second is King given first is King})$$

$$= (4/52)(3/51)$$

$$\approx 0.0045$$

Or in other words, $P(E \text{ and } F) = P(E) \cdot P(F|E)$

This idea is actually a version of the Multiplication Rule for Independent Events, and is called the **General Multiplication Rule**.

**General Multiplication Rule**

The probability that two events E and F both occur is

$$P(E \text{ and } F) = P(E) \cdot P(F|E)$$

**Example 5**

Let's try a new probability experiment. This time, consider a bag of marbles, containing 10 red, 20 blue, and 15 green marbles. Suppose that two marbles are drawn without replacement. (The first marble is not put back in the bag before drawing the second.)
What is the probability that both marbles drawn are red?

[reveal answer]

**Checking for Independence**

If you recall, in Section 5.3, we defined what it meant for two events to be independent:

> Two events E and F are **independent** if the occurrence of event E does not affect the probability of event F.

Looking at this in terms of conditional probability, if the occurrence of E doesn't affect the probability of F, then $P(F|E) = P(F)$. This is a good way to test for independence. In fact, we can redefine independence using this concept.

> Two events E and F are **independent** if $P(F|E) = P(F)$.

Let's use this new definition in an example to determine if two events are independent.

**Example 6**

Let's again use the data from Example 3 and the survey given to 52 students in a Basic Algebra course at ECC, with the following responses to the statement "I enjoy math."

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>6</td>
<td>10</td>
<td>3</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>8</td>
<td>14</td>
<td>7</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>14</td>
<td>24</td>
<td>10</td>
<td>4</td>
<td>52</td>
</tr>
</tbody>
</table>

Suppose a student is selected at random from those surveyed and we define the events E and F as follows:

E = student selected is female

F = student enjoys math

Are events E and F independent?

[reveal answer]
Section 5.5: Counting Techniques

5.1 Probability Rules
5.2 The Addition Rule and Complements
5.3 Independence and the Multiplication Rule
5.4 Conditional Probability and the General Multiplication Rule

**5.5 Counting Techniques**

5.6 Putting It Together: Probability

**Objectives**

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

1. solve counting problems using the Multiplication Rule
2. solve counting problems using permutations
3. solve counting problems using combinations
4. solve counting problems involving permutations with non-distinct items
5. compute probabilities involving permutations and combinations

Do you remember the classical method for calculating probabilities from Section 5.1?

\[
P(E) = \frac{\text{number of ways } E \text{ can occur}}{\text{total number of possible outcomes}} = \frac{N(E)}{N(S)}
\]

Well, sometimes counting the "number of ways E can occur" or the "total number of possible outcomes" can be fairly complicated. In this section, we'll learn several counting techniques, which will help us calculate some of the more complicated probabilities.

**The Multiplication Rule of Counting**

Let's suppose you're preparing for a wedding, and you need to pick out tuxedos for the groomsmen. Men's Tuxedo Warehouse has a Build-A-Tux feature which allows you to look at certain combinations and build your tuxedo online. Let's suppose you have the components narrowed down to two jackets, two vest and tie combinations, and three shirt colors. How many total combinations might there be?

A good way to help understand this type of situation is something called a **tree diagram**. We begin with the jacket choices, and then each jacket "branches" out into the two vest and tie combinations, and then each of those then "branches" out into the three shirt combinations. It might look something like this:
In total, it looks like we have 12 possible combinations of jackets, vests, and shirts. (Of course, some may not fit your fashion sense, but that's another question all-together...)

Isn't there an easier way to do this? Why yes, there is! Think of it this way, for each jacket choice, there are two vest and tie choices. That gives us 4 total jacket and vest/tie combinations. Then, for each of those, there are three shirt choices, giving us a total of 12.

In general, we multiply the number of ways to make each choice, so...

\[
\text{total number of outfits} = (\text{number of jackets}) \times (\text{number of vest/ties}) \times (\text{number of shirts})
\]

That leads us to the Multiplication Rule of Counting:
Multiplication Rule of Counting

If a task consists of a sequence of choices in which there are \( p \) ways to make the first choice, \( q \) ways to make the second, etc., then the task can be done in

\[
p \cdot q \cdot r \cdot \ldots
\]
different ways.

Let's try some examples.

Example 1

How many 7-character license plates are possible if the first three characters must be letters, the last four must be digits 0-9, and repeated characters are allowed?

[ reveal answer ]

Example 2

Many garage doors have remote-access keypads outside the door. Let's suppose a thief approaches a particular garage and notices that four particular numbers are well-used. If we assume the code uses all four numbers exactly once, how many 4-digit codes does the thief have to try?

[ reveal answer ]

Example 2, from earlier this section is an example of particular counting technique called a permutation. Rather than giving you formulas and examples myself, I'd like to make another reference to some content from one of my favorite web sites, BetterExplained. Here's what the author, Kalid Azad writes about permutations:

Permutations: The hairy details

Let's start with permutations, or all possible ways of doing something. We’re using the fancy-pants term “permutation”, so we’re going to care about every last detail, including the order of items.

Let’s say we have 8 people:

1. Alice
2. Bob
3. Charlie
4. David
5. Eve
6. Frank
7. George
8. Horatio

How many ways can we pick a Gold, Silver, and Bronze medal for “Best friend in the world”? 
We’re going to use permutations since the order we hand out these medals matter. Here’s how it breaks down:

Gold medal: 8 choices: A B C D E F G H (Clever how I made the names match up with letters, eh?). Let’s say A wins the Gold.
Silver medal: 7 choices: B C D E F G H. Let’s say B wins the silver.
Bronze medal: 6 choices: C D E F G H. Let’s say… C wins the bronze.

We picked certain people to win, but the details don’t matter: we had 8 choices at first, then 7, then 6. The total number of options was $8 \times 7 \times 6 = 336$.

Let’s look at the details. We had to order 3 people out of 8. To do this, we started with all options (8) then took them away one at a time (7, then 6) until we ran out of medals.

We know the factorial is: $8! = 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$

Unfortunately, that does too much! We only want $8 \times 7 \times 6$. How can we “stop” the factorial at 5?

This is where permutations get cool: notice how we want to get rid of $5!4!3!2!1$. What’s another name for this? 5 factorial!

So, if we do $8!/5!$ we get:

$$\frac{8!}{5!} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{5 \times 4 \times 3 \times 2 \times 1} = 8 \times 7 \times 6$$

And why did we use the number 5? Because it was left over after we picked 3 medals from 8. So, a better way to write this would be:

$$\frac{8!}{(8 - 3)!}$$

where $8!/(8-3)!$ is just a fancy way of saying “Use the first 3 numbers of 8!”. If we have $n$ items total and want to pick $k$ in a certain order, we get:

$$\frac{n!}{(n - k)!}$$ just means “Use the first $k$ numbers of $n!$”

And this is the fancy permutation formula: You have $n$ items and want to find the number of ways $k$ items can be ordered:

$$P(n, k) = \frac{n!}{(n - k)!}$$

Source: BetterExplained, Kalid Azad

Article: Easy Permutations and Combinations
Used with permission.

As a side note, your text uses the notation $n_P^k$ rather than Kalid’s $P(n,k)$. I’ve seen both used, though the later tends to be more prevalent in higher-level math classes. We’ll stick with the textbook version, just to be consistent.
Permutations of $n$ Distinct Objects Taken $r$ at a Time

The number of arrangements of $r$ objects chosen from $n$ objects in which

1. the $n$ objects are distinct,
2. repeats are not allowed,
3. order matters,

is given by the formula $\frac{n!}{(n-r)!}$.

OK, let's try a couple.

**Example 3**

Suppose an organization elects its officers from a board of trustees. If there are 30 trustees, how many possible ways could the board elect a president, vice-president, secretary, and treasurer?

[ reveal answer ]

**Example 4**

Suppose you're given a list of 100 desserts and asked to rank your top 3. How many possible "top 3" lists are there?

[ reveal answer ]

In the previous page, we talked about the number of ways to choose $k$ objects from $n$ if the ordered mattered - like giving medals, electing officers, or pickling favorite desserts. What if order doesn't matter, like picking members of a committee?

Again, I'll let Kalid Azad explain.

**Combinations, Ho!**

Combinations are easy going. Order doesn’t matter. You can mix it up and it looks the same. Let’s say I’m a cheapskate and can’t afford separate Gold, Silver and Bronze medals. In fact, I can only afford empty tin cans.

How many ways can I give 3 tin cans to 8 people?

Well, in this case, the order we pick people doesn't matter. If I give a can to Alice, Bob and then Charlie, it's the same as giving to Charlie, Alice and then Bob. Either way, they're going to be equally disappointed.

This raises and interesting point — we’ve got some redundancies here. Alice Bob Charlie = Charlie Bob Alice. For a moment, let’s just figure out how many ways we can rearrange 3 people.

Well, we have 3 choices for the first person, 2 for the second, and only 1 for the last. So we have $3 \times 2 \times 1$ ways to re-arrange 3 people.

Wait a minute... this is looking a bit like a permutation! You tricked me!

Indeed I did. If you have N people and you want to know how many arrangements there are for all of them, it’s just N factorial or $N!$.
So, if we have 3 tin cans to give away, there are 3! or 6 variations for every choice we pick. If we want to figure out how many combinations we have, we just create all the permutations and divide by all the redundancies. In our case, we get 336 permutations (from above), and we divide by the 6 redundancies for each permutation and get 336/6 = 56.

The general formula is

\[ C(n, k) = \frac{P(n, k)}{k!} \]

which means “Find all the ways to pick k people from n, and divide by the k! variants”. Writing this out, we get our combination formula, or the number of ways to combine k items from a set of n:

\[ C(n, k) = \frac{n!}{(n-k)!k!} \]

Source: BetterExplained, Kalid Azad
Article: Easy Permutations and Combinations
Used with permission.

As a side note, your text uses the notation \( _rC_k \) rather than Kalid's \( C(n,k) \). As with permutations, we'll stick with the textbook version, just to be consistent.

**Combinations of \( n \) Distinct Objects Taken \( r \) at a Time**

The number of arrangements of \( n \) objects using \( r \leq n \) of them, in which

1. the \( n \) objects are distinct,
2. repeats are not allowed,
3. order does not matter,

is given by the formula \( _rC_k = \frac{n!}{r!(n-r)!} \).

All right, let's try this new one out.

**Example 5**

Let's consider again the board of trustees with 30 members. In how many ways could the board elect four members for the finance committee?

[ reveal answer ]

**Example 6**

Suppose you're a volleyball tournament organizer. There are 10 teams signed up for the tournament, and it seems like a good idea for each team to play every other team in a "round robin" setting, before advancing to the playoffs. How many games are possible if each team plays every other team once?

[ reveal answer ]
This second type is less common. What if we want to know how many ways to order \( n \) objects, but they're not all distinct? Here's an example to illustrate:

**Example 7**

In how many ways could the letters in the word STATISTICS be rearranged?

The answer is a little tricky. Think of the rearranged words as places for letters to go. Something like this:

___ ___ ___ ___ ___ ___ ___ ___ ___

In STATISTICS, we have the following letters:

- 3 S's
- 3 T's
- 2 I's
- 1 A
- 1 C

We can't really say that there are 4 choices for the first letter and proceed from there, since the number of choices for the second letter depend on which letter was chosen for the first.

Instead, we choose the spots for each of the letters. First, pick 3 of the 10 spots for the S's. We can do that in \( \binom{10}{3} \) ways. Then pick 3 spots for the 3 T's. We can do that in \( \binom{7}{3} \) ways. Similarly, we can pick the spots for the I's, the A, and the C in \( \binom{4}{2}, \binom{2}{1}, \) and \( \binom{1}{1} \) ways, respectively. In total, that means we rearrange the letters in:

\[
\binom{10}{3} \cdot \binom{7}{3} \cdot \binom{4}{2} \cdot \binom{2}{1} \cdot \binom{1}{1}
\]

It may just be me, that is really messy. Oddly enough, writing out the combinations reveals a nice way to simplify it:

\[
\frac{10!}{7!3! \cdot 4!2!1!1!} = \frac{10!}{3!3!2!1!1!1!}
\]

**Permutations with Non-distinct Items**

The number of permutations of \( n \) objects, where there are \( n_1 \) of the 1st type, \( n_2 \) of the 2nd type, etc, is

\[
\frac{n!}{n_1! \cdot n_2! \cdots n_k!}
\]

One quick example:

**Example 8**

In how many ways can the word REARRANGE be rearranged?

[ reveal answer ]
Section 5.6: Putting It Together: Which Method Do I Use?

Objectives

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

1. determine the appropriate probability rule to use
2. determine the appropriate counting technique to use

The last thing we really need to do is connect all these counting techniques to the stated purpose - probability. We'll start by first focusing on choosing the correct probability rule. Once we have that down, we'll continue to the next stage of selecting the correct counting technique.

Probability Rules

Let's start by reviewing the probability rules from the previous sections.

The Basic Principle of Probability (Classical Method)

\[ P(E) = \frac{\text{number of ways } E \text{ can occur}}{\text{total number of possible outcomes}} = \frac{N(E)}{N(S)} \]

The General Addition Rule

\[ P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F) \]

The Complement Rule

\[ P(E) + P(E^c) = 1 \]

General Multiplication Rule

The probability that two events E and F both occur is

\[ P(E \text{ and } F) = P(E) \cdot P(F|E) \]

Our task then, is choosing the correct rule. Here's a simplified version of the decision-making flowchart from your textbook.
Let's try an example together.

Example 1

**Problem:** Suppose a basketball player makes 80% of her free throws. If she shoots five free throws, find the probability that she makes at least one basket.

**Solution:** Let's try following our flowchart:

- Are you finding the probability of a compound event? **Yes.**
- Does the problem involve ‘AND’, ‘OR’, or 'At least'? **At least.**
- Use the Complement Rule.

The Complement Rule is \( P(E) + P(E^c) = 1 \). In our case, we have something like this:

\[
P(\text{she makes at least one}) + P(\text{she misses all five}) = 1
\]

We can find what we want by solving for it:

\[
P(\text{she makes at least one}) = 1 - P(\text{she misses all five})
\]

Now, we need to look at the probability that she misses all five. Let's try the sequence again:

- Are you finding the probability of a compound event? **Yes.**
- Does the problem involve 'AND', 'OR', or 'At least'? **AND.** (We want the probability that she misses the first and the second and the third, etc)
- Use the Multiplication Rule.

If we assume the individual attempts are independent (granted, a big assumption), we can find the probability that she misses all five by multiplying the probability that she misses one by itself five times.
P(she makes at least one)  
= 1 - P(she misses all five)  
= 1 - (0.2)(0.2)(0.2)(0.2)(0.2)  
= 0.99968

So she'll make at least one basket 99.97% of the time!

All right, it's time for you to try some yourself. Here we go...

**Example 2** Consider a standard 52-card deck of playing cards. A single card is drawn at random, with the following events defined:

A = a diamond is drawn  
B = a face card is drawn (face cards are Jacks, Queens, or Kings)

Find P(A or B).

[reveal answer]

**Example 3**

Suppose a fair die is tossed and a fair coin is flipped. Find the probability that the die is even and the coin is heads.

[reveal answer]

**Example 4**

Texas Hold'Em is a form of poker regaining popularity recently due to the exposure of the World Series of Poker on ESPN. The game is fairly complex, but components of it can be relatively easy to understand.

From Wikipedia

Hold 'em is a community card game where each player may use any combination of the five community cards and the player's own two hole cards to make a poker hand, in contrast to poker variants like stud or draw where each player holds a separate individual hand.

The game is played by first dealing every player two cards. Then three card are dealt face-up in the middle of the table (this is called the flop). All players are allowed to use these "community" cards in the middle of the table along with the two best. Later, two more card are dealt, with a total of five "community" cards for all players to use in their hands.

Suppose you are dealt A♥Q♥. If we assume the other 50 cards are all equally likely to be dealt during the flop, what is the probability that the next three cards are also hearts, giving you a flush (five of the same suit) after the flop?

Just focus on choosing the correct probability rule - don't worry about actually finding the probability.
Of course, this is only a small selection of the wide variety of problems possible. Be sure to complete all of the MyMathLab homework and the practice exam questions. You might also consider looking at some of the unassigned odd problems in your text for extra practice, or additional practice through MyMathLab.

Once you've determined which probability rule to apply, you often need some counting techniques in order to complete the problem. For many students, this aspect of probability questions is the most troublesome. Like the previous page (and your text), we'll illustrate the main ideas with a flow chart.

## The Counting Techniques

As with the probability rules, we need to first review some of our different counting techniques, beginning with the most important.

### Multiplication Rule of Counting

If a task consists of a sequence of choices in which there are $p$ ways to make the first choice, $q$ ways to make the second, etc., then the task can be done in

$$p \cdot q \cdot r \cdot \ldots$$

different ways.

### Permutations of $n$ Distinct Objects Taken $r$ at a Time

The number of arrangements of $r$ objects chosen from $n$ objects in which

1. the $n$ objects are distinct,
2. repeats are not allowed,
3. order matters,

is given by the formula

$$n^P_r = \frac{n!}{(n-r)!}.$$

### Combinations of $n$ Distinct Objects Taken $r$ at a Time

The number of arrangements of $n$ objects using $r \leq n$ of them, in which

1. the $n$ objects are distinct,
2. repeats are not allowed,
3. order does not matter,

is given by the formula

$$n^C_r = \frac{n!}{r!(n-r)!}.$$

### Permutations with Non-distinct Items

The number of permutations of $n$ objects, where there are $n_1$ of the 1st type, $n_2$ of the 2nd type, etc, is

$$\frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!}$$

## Choosing the Appropriate Counting Technique
The trouble, then, becomes choosing which one of these (or more than one) applies to a particular problem. As we did with the probability rules, we'll illustrate this with a decision-making flowchart. In this case, it's the same one that's in your text.

Let's try one example together.

**Example 5**

**Questions:** In the Powerball lottery, there are 5 balls numbered 1-55, and an additional "powerball" numbered 1-42. To win the grand prize, you must match all 5 and the powerball. **What is the probability of winning the grand prize with one ticket?**

**Solutions:** To answer this question, we need to remember the basic probability rule. Expanded for this example, it would be something like this:

\[ P(\text{winning}) = \frac{\text{number of ways to win}}{\text{total number of possible outcomes}} \]

The number of ways to win is easy - there's only one!

For the total number of possible outcomes, let's consider our flowchart.

- **Are we making a sequence of choices?**
  - Yes. We're first counting the number of ways for the 5 balls to be drawn, and then we need to count the number of ways for the powerball to be drawn.

- **Are the choices independent of each other?** (In other words, is the powerball independent of what happens in the first 5 balls?)
  - Yes.

- **Use the Multiplication Rule of Counting.**

So we need to multiply the number of ways to do each step.

\[
\text{total # of outcomes} = (\# \text{ of ways for the 5 to be drawn}) \times (\# \text{ of ways for powerball})
\]
Since order doesn't matter for the 5 balls, that part is a combination of 5 from 55. The powerball is simply 42, since there are 42 choices.

The total number of outcomes is thus \( \binom{55}{5} \cdot 42 = 146,107,962 \).

The probability is then \( \frac{1}{146,107,962} \approx 0.000\,000\,007 \)

That means a $1 ticket has only about a 1 in 150 million chance of winning the grand prize!

Here are some examples to try. Some are counting questions and some are actual probability questions, but the probability rule shouldn't be the hard part. (Hint: They can all use the classical method!)

---

**Example 6**

Suppose you are told to create a password that has the following rules:

1. it must consist of exactly 4 characters,
2. one must be a number,
3. one must be a capitalized, and
4. the letters must all differ.

How many possible passwords are there following these rules?

[reveal answer]

**Example 7**

Suppose you have a bag of 20 blue marbles and 40 red marbles. What is the probability that if 5 are drawn without replacement, 2 are blue and 3 are red?

[reveal answer]

---

**Example 8**

In Example 4, you worked on choosing the correct probability rule for a Texas Hold'Em situation. Use your rule to actually calculate the probability. Here's the problem again:

Texas Hold'Em is form of poker regaining popularity recently due to the exposure of the World Series of Poker on ESPN. The game is fairly complex, but components of it can be relatively easy to understand.

*From Wikipedia*

Hold 'em is a community card game where each player may use any combination of the five community cards and the player's own two hole cards to make a poker hand, in contrast to poker variants like stud or draw where each player holds a separate individual hand.

The game is played by first dealing every player two cards. Then three
card are dealt face-up in the middle of the table (this is called the flop). All players are allowed to use these "community" cards in the middle of the table along with the two best. Later, two more card are dealt, with a total of five "community" cards for all players to use in their hands.

Suppose you are dealt A♥Q♥. If we assume the other 50 cards are all equally likely to be dealt during the flop, what is the probability that the next three cards are also hearts, giving you a flush (five of the same suit) after the flop?

[ reveal answer ]

As above, this is only a small selection of the wide variety of problems possible. Be sure to complete all of the MyMathLab homework and the practice exam questions. You might also consider looking at some of the unassigned odd problems in your text for extra practice, or additional practice through MyMathLab.
Chapter 6: Discrete Probability Distributions

6.1 Discrete Random Variables
6.2 The Binomial Probability Distribution

In Chapter 6, we expand on the probability concepts we learned in Chapter 5, and introduce the idea of a random variable. Random variables are useful because they help us determine if playing a game like roulette (shown to the right) is profitable in the long-term. (It isn't.)

Random variables also help us determine how insurance companies set the premiums for their policies. They can also help an investor decide whether or not to invest in a company.

In Section 6.2, we'll introduce a specific type of random variable called a binomial random variable. Binomial random variables will help us answer questions like "What's the probability of getting 3 questions right on a multiple choice test if we're just guessing?" Or "If a basketball player makes 80% of her free throws, how often will she make less than 8 of 10?"

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::
Section 6.1: Discrete Random Variables

6.1 Discrete Random Variables
6.2 The Binomial Probability Distribution

Objectives

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

1. distinguish between discrete and continuous random variables
2. identify discrete probability distributions
3. construct probability histograms
4. compute and interpret the mean of a discrete random variable
5. interpret the mean of a discrete random variable as an expected value
6. compute the variance and standard deviation of a discrete random variable*

* You will not be tested on this objective.

Random Variables

Many probability experiments can be characterized by a numerical result. In Example 1, from Section 5.1, we flipped three coins. Instead of looking at particular outcomes (HHT, HTT, etc.), we might instead be interested in the total number of heads. Something like this:

In this case, the number of heads is called a random variable.
A random variable is a numerical measure of the outcome of a probability experiment whose value is determined by chance.

Example 1
Another example might be when we roll two dice, as in Example 2, from Section 5.1. Rather than looking at the dice individually, we can instead look at the sum of the dice, which would be a random variable.

In this case, if we let X = the sum of the two dice, x = 2, 3, 4, ..., 12. (We usually use a capital X to represent the random variable, and a lower case x to represent the particular values it can take on.)

One common goal with random variables is to know what the probability of each value is. This is called a probability distribution.

The probability distribution of a discrete random variable X provides the possible value of the random variable along with their corresponding probabilities. A probability distribution can be in the form of a table, graph, or mathematical formula.

Let's look at the earlier coin example to illustrate.

Example 2
If we flip three fair coins and let X = the number of heads, we know that x = 0, 1, 2, 3. We also know that since the coin is fair, each of the strands in the tree diagram shown earlier is equally likely. Since there are 8 total outcomes (HHH, HHT, HTH, etc), the probability distribution would look something like this:

<table>
<thead>
<tr>
<th>x</th>
<th>P(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/8</td>
</tr>
<tr>
<td>1</td>
<td>3/8</td>
</tr>
<tr>
<td>2</td>
<td>3/8</td>
</tr>
<tr>
<td>3</td>
<td>1/8</td>
</tr>
</tbody>
</table>

You'll notice that if find the sum of the P(x) values, we get 1.

Why don't you try an example now:

Example 3
Consider again the probability experiment where a fair six-sided die is rolled twice, and X = the sum of the two dice. Find the probability histogram for X.
A probability histogram is similar to a histogram for single-valued discrete data from Section 2.2, except the height of each rectangle is the probability rather than the frequency or relative frequency.

**Example 4**

Looking again at the previous example about rolling two dice, the probability histogram would look something like this:

Since random variables represent numbers, it seems reasonable that we should be able to find the mean of those numbers, much like we did in Section 3.1. In fact, we can, but it has a much different meaning in this new context.

**The Mean of a Random Variable**

Let's consider our example again with the two dice.

We know $x = 2, 3, 4, \ldots, 12$, but there aren't equal numbers of each. If we were to calculate the mean of $x$ like we usually would, we'd get something like this:
What an interesting result! In fact, this isn't coincidental. The mean of a random variable will always be the sum of the values of the random variable multiplied by their corresponding probabilities. More formally:

\[\mu = \frac{2 \cdot 3 + 3 \cdot 3 + 4 \cdot 4 + 4 \cdot 4 + \cdots + 11 \cdot 11 + 12 \cdot 12}{36}\]
\[= \frac{2 \cdot 1 + 3 \cdot 2 + 4 \cdot 3 + \cdots + 11 \cdot 2 + 12 \cdot 1}{36}\]
\[= \frac{2 \cdot 1 \cdot \frac{1}{36} + 3 \cdot 2 \cdot \frac{2}{36} + 4 \cdot 3 \cdot \frac{3}{36} + \cdots + 11 \cdot 2 \cdot \frac{11}{36} + 12 \cdot \frac{1}{36}}{36}\]
\[= 2 \cdot P(X = 2) + 3 \cdot P(X = 3) + 4 \cdot P(X = 4) + \cdots + 11 \cdot P(X = 11) + 12 \cdot P(X = 12)\]
\[= 7\]

All right, let's try one.

**Example 5**

Let's go back to the three coins again. If we flip three fair coins and let \(X\) = the number of heads, what is the mean of \(X\)?

[reveal answer]

**Expected Value**

So what does the mean of a random variable actually ... mean? When we found a mean of 7 in the example above, it certainly didn't mean that we'll get 7 every time. And it doesn't mean that our average after say 20 rolls will be 7.

No - the mean of a random variable is like a *long-term expectation*. Think of the mean being what to expect in the long run. In fact, the mean is usually referred to as the **expected value** of the random variable. If we continue throwing the two dice in the earlier example, our long-term average will get closer and closer to 7. In Example 5, the more times we toss three coins, the closer our long-term average will approach 1.5.

**Expected values** are used in lots of calculations in the business and finance world. Poker players use expected values to help make decisions on whether to continue playing in a hand. Hopefully by the end of this section, you'll use the idea of expected value to *not* play games of chance like roulette!

Let's illustrate with some examples.

**Example 6**

In the game of **roulette**, a wheel consists of 38 slots numbered 0, 00, 1, 2, ..., 36. To play the game, a metal ball is spun around the wheel and
is allowed to fall into one of the numbered slots. A dozen bet is betting that one of a particular dozen numbers hits on the next spin of the wheel. The wheel is divided into 3 different groups; 1-12, 13-24, and 25-36. If the payout for a dozen bet is 2 to 1 (the original bet is returned, along with twice its value), what is the expected value of $1 dozen bet?

Solution: To solve this problem, let's first define a random variable. In a situation like this, we typically let $X = \text{amount won}$. $X$ can then take on values of either -$1 (we lose) or +$2 (we get our $1 plus $2 more).

$P(X = 2) = \frac{12}{38}$, since there are 12 ways to land on our dozen (regardless of which dozen we choose), and

$P(X = -1) = \frac{26}{38}$, since there are 26 other numbers

The expected value of $X$ is then:

$E(X) = (-1) \cdot P(X = -1) + (2) \cdot P(X = 2)$

$= (-1)(26/38) + (2)(12/38)$

$= -2/38 \approx -0.05$

So on average, we expect to lose 5¢ for every $1 bet we make. Of course, this doesn't mean we'll lose 5¢ every time - just that in the long run, we'll average a loss of 5¢ per $1 bet.

You can see examples of other Roulette expected values here.

---

**Example 7**

Consider a car owner who has an 80% chance of no accidents in a year, a 20% chance of being in a single accident in a year, and no chance of being in more than one accident in a year.

For simplicity, assume that there is a 50% probability that after the accident the car will need repairs costing $500, a 40% probability that the repairs will cost $5,000, and a 10% probability that the car will need to be replaced, which will cost $15,000.

What is the expected loss for the car owner per year?

Solution: This one is a little trickier. If we let $X = \text{loss for the year}$, $X$ can be $0, $500, $5,000, or $15,000. $P(X=0) = 0.8$, but $P(X = $500$)$ is actually $(0.2)(0.5)$, since there's a 20% chance of being in an accident, and a 50% chance of that accident causing repair costs of $500. The complete list of $X$ and its corresponding values is:

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(X=x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0$</td>
<td>0.80</td>
</tr>
<tr>
<td>$500$</td>
<td>$(0.2)(0.5) = 0.1$</td>
</tr>
<tr>
<td>$5,000$</td>
<td>$(0.2)(0.4) = 0.08$</td>
</tr>
<tr>
<td>$15,000$</td>
<td>$(0.2)(0.1) = 0.02$</td>
</tr>
</tbody>
</table>
Example 8

Gambling is full of expected value calculations. We already did one earlier about roulette (and saw that the game has a negative expectation for the player). Let's try another one, but this time make it more a game of skill.

There are many variations of poker, but most involve rounds of betting, where players have to choose what to bet, and whether to "call" (pay the amount bet by another player). One relatively simple situation to look at is the final round of betting.

Let's suppose you're in a poker game with the great Doyle Brunson. It's the last play of the hand - Doyle has bet $50 and you have to decide whether or not to call. There's $250 in the middle of the table if you call and win. If you call and lose, you'll lose another $50.

Doyle has been around a long time, so you can't be sure if he's bluffing. If you call and he is, you win the $250 in the middle and his additional $50. If you call and he's not bluffing, you lose your additional $50.

Based on the way the hand has played out, you think there's about a 20% chance he has nothing, and an 80% chance that he has you beat. What is the expected value of calling Doyle's bet?

Solution: There's a lot going on here, but let's start like we've done in the past. We'll make a random variable, call it X, with possible values of $300 (you call and win), and -$50 (you call and lose).

\[ E(X) = (300)(0.2) + (-50)(0.8) = 20 \]

So on average, if we were to replay this hand over and over, we'd expect an average profit of $20 over time. So yes, we should call. Keep in mind, we're still going to lose 80% of the time, but the 20% of the time we win is enough to make it worth the while.

Example 9

Suppose you have an investment opportunity. A new small business in town is looking for investors, and they're asking for a $20,000 investment from you. After some investigating, you determine that with the current economic climate, there's a 40% chance the company will fail in the first year, and you'll lose the full $20,000. There's about a 50% chance the company will struggle but survive, and you'll
have a loss of about $5,000. There is an opportunity in the field, though, and you guess there’s about a 10% chance that the company can make it big, and you’ll quadruple your investment in the first year.

Based on these estimates, should you make the investment?

[ reveal answer ]

The Standard Deviation of a Random Variable

Although we’re usually much more interested in the expected value (mean) of a random variable, there are times (especially later on in the course) that we’ll also be interested in the standard deviation. The formula is similar to the mean in that it weights each value by its corresponding probability.

### The Variance and Standard Deviation of a Discrete Random Variable

The variance of a discrete random variable is given by the formula

$$\sigma_x^2 = \sum [(x - \mu_x)^2 \cdot P(x)]$$

where $x$ is the value of the random variable and $P(x)$ is the probability of observing the random variable $x$.

To find the standard deviation of the discrete random variable, take the square root of the variance.

We'll just do one quick example of standard deviation.

#### Example 10

Let’s consider our example again with the two dice.

We know from earlier this section that the expected value is 7. Using that, we can find the variance and standard deviation.

$$\sigma_x^2 = (2 - 7)^2 \cdot \frac{1}{36} + (3 - 7)^2 \cdot \frac{2}{36} + \ldots$$

$$+ (11 - 7)^2 \cdot \frac{2}{36} + (12 - 7)^2 \cdot \frac{1}{36} \approx 5.83$$

And so the standard deviation is: $\sigma_x = \sqrt{\sigma_x^2} \approx 2.42$
Section 6.2: The Binomial Probability Distribution

6.1 Discrete Random Variables

6.2 The Binomial Probability Distribution

Objectives

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

1. determine whether a probability experiment is a binomial experiment
2. compute probabilities of binomial experiments
3. compute and interpret the mean and standard deviation of a binomial random variable

Binomial Experiments

In the last section, we talked about some specific examples of random variables. In this next section, we deal with a particular type of random variable called a **binomial random variable**. Random variables of this type have several characteristics, but the key one is that the experiment that is being performed has only two possible outcomes - *success* or *failure*.

An example might be a free kick in soccer - either the player scores or goal or she doesn't. Another example would be a flipped coin - it's either heads or tails. A multiple choice test where you're totally guessing would be another example - each question is either right or wrong.

Let's be specific about the other key characteristics as well:

**Criteria for a Binomial Probability Experiment**

A **binomial experiment** is an experiment which satisfies these four conditions:

- A fixed number of trials
- Each trial is independent of the others
- There are only two outcomes
- The probability of each outcome remains constant from trial to trial.

In short: *An experiment with a fixed number of independent trials, each of which can only have two possible outcomes.*

(Since the trials are independent, the probability remains constant.)

If an experiment is a binomial experiment, then the random variable \( X = \) the number of successes is called a **binomial random variable**.

Let's look at a couple examples to check your understanding.

**Example 1**

Consider the experiment where three marbles are drawn without replacement from a bag containing 20 red and 40 blue marbles, and the number of red marbles drawn is recorded. Is this a binomial experiment?

[ reveal answer ]

Source: stock.xchng
The Binomial Distribution

Once we determine that a random variable is a binomial random variable, the next question we might have would be how to calculate probabilities.

Let's consider the experiment where we take a multiple-choice quiz of four questions with four choices each, and the topic is something we have absolutely no knowledge. Say... theoretical astrophysics. If we let $X = \text{the number of correct answer}$, then $X$ is a binomial random variable because

- there are a fixed number of questions (4)
- the questions are independent, since we're just guessing
- each question has two outcomes - we're right or wrong
- the probability of being correct is constant, since we're guessing: $1/4$

So how can we find probabilities? Let's look at a tree diagram of the situation:

Finding the probability distribution of $X$ involves a couple key concepts. First, notice that there are multiple ways to get 1, 2, or 3 questions correct. In fact, we can use combinations to figure out how many ways there are! Since $P(X=3)$ is the same regardless of which 3 we get correct, we can just multiply the probability of one line by 4, since there are 4 ways to get 3 correct.

Not only that, since the questions are independent, we can just multiply the probability of getting each one correct or incorrect, so $P(\bigcirc\bigcirc\bigcirc\bigcirc) = (3/4)^3(1/4)$. Using that concept to find all the probabilities, we get the...
following distribution:

<table>
<thead>
<tr>
<th>x</th>
<th>P(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$P(0) = \left(\frac{1}{4}\right)^4$</td>
</tr>
<tr>
<td>1</td>
<td>$P(1) = 4 \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^3$</td>
</tr>
<tr>
<td>2</td>
<td>$P(2) = 6 \left(\frac{1}{4}\right)^2 \left(\frac{3}{4}\right)^2$</td>
</tr>
<tr>
<td>3</td>
<td>$P(3) = 4 \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^1$</td>
</tr>
<tr>
<td>4</td>
<td>$P(4) = \left(\frac{1}{4}\right)^4$</td>
</tr>
</tbody>
</table>

We should notice a couple very important concepts. First, the number of possibilities for each value of X gets multiplied by the probability, and in general there are $4C_x$ ways to get X correct. Second, the exponents on the probabilities represent the number correct or incorrect, so don't stress out about the formula we're about to show. It's essentially:

$$P(X) = (\text{ways to get } X \text{ successes}) \cdot (\text{prob of success})^{\text{successes}} \cdot (\text{prob of failure})^{\text{failures}}$$

**The Binomial Probability Distribution Function**

The probability of obtaining $x$ successes in $n$ independent trials of a binomial experiment, where the probability of success is $p$, is given by

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Where $x = 0, 1, 2, \ldots, n$

**Technology**

Here's a quick overview of the formulas for finding binomial probabilities in StatCrunch.

Click on **Stat > Calculators > Binomial**

Enter $n$, $p$, the appropriate equality/inequality, and $x$. The figure below shows $P(X \geq 3)$ if $n=4$ and $p=0.25$.

Let's try some examples.

**Example 3**

Consider the example again with four multiple-choice questions of which you have no knowledge. What is the probability of getting exactly 3 questions correct?
Example 4
A basketball player traditionally makes 85% of her free throws. Suppose she shoots 10 baskets and counts the number she makes. What is the probability that she makes less than 8 baskets?

Example 5
Traditionally, about 70% of students in a particular Statistics course at ECC are successful. Suppose 20 students are selected at random from all previous students in this course. What is the probability that more than 15 of them will have been successful in the course?

The Mean and Standard Deviation of a Binomial Random Variable

Let's consider the basketball player again. If she takes 100 free throws, how many would we expect her to make? (Remember that she historically makes 85% of her free throws.)

The answer, of course, is 85. That's 85% of 100.

We could do the same with any binomial random variable. In Example 5, we said that 70% of students are successful in the Statistics course. If we randomly sample 50 students, how many would we expect to have been successful?

Again, it's fairly straightforward - 70% of 50 is 35, so we'd expect 35.

Remember back in Section 6.1, we talked about the mean of a random variable as an expected value. We can do the same here and easily derive a formula for the mean of a binomial random variable, rather than using the definition. Just as we did in the previous two examples, we multiply the probability of success by the number of trials to get the expected number of successes.

Unfortunately, the standard deviation isn't as easy to understand, so we'll just give it here as a formula.

The Mean and Standard Deviation of a Binomial Random Variable

A binomial experiment with n independent trials and probability of success p has a mean and standard deviation given by the formulas

\[ \mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1-p)} \]

Let's try a quick example.

Example 6
Suppose you're taking another multiple choice test, this time covering
The Shape of a Binomial Probability Distribution

The best way to understand the effect of $n$ and $p$ on the shape of a binomial probability distribution is to look at some histograms, so let’s look at some possibilities.

Based on these, it would appear that the distribution is symmetric only if $p=0.5$, but this isn’t actually true. Watch what happens as the number of trials, $n$, increases:

Interestingly, the distribution shape becomes roughly symmetric when $n$ is large, even if $p$ isn’t close to 0.5. This brings us to a key point:

As the number of trials in a binomial experiment increases, the probability distribution becomes bell-shaped. As a rule of thumb, if $np(1-p)\geq 10$, the distribution will be approximately bell-shaped.
In Chapter 7, we bring together much of the ideas in the previous two on probability. We expand the earlier bell-shaped distribution (we introduced this shape back in Section 2.2) to its more formal name of a normal curve.

Many random variables have histograms that follow the normal curve, like an individual's height, the thickness of tree bark, IQs, or the amount of light emitted by a light bulb. This is helpful, because we can use it to answer all kinds of interesting questions.

- What proportion of individuals are geniuses?
- Is a systolic blood pressure of 110 unusual?
- What percentage of a particular brand of light bulb emits between 300 and 400 lumens?
- What is the 90th percentile for the weights of 1-year-old boys?

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::

This work is licensed under a Creative Commons License.
Objectives

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

1. use the uniform probability distribution
2. graph a normal curve
3. state the properties of the normal curve
4. explain the role of area in the normal density function

Probability Density Functions

In Chapter 6, we focused on discrete random variables, random variables which take on either a finite or countable number of values. Continuous random variables, which have infinitely many values, can be a bit more complicated.

Consider the rand() function in the computer software Microsoft Excel. It returns a random number between 0 and 1. There are infinitely many possibilities, so each particular value has a probability of 0!

When we consider continuous random variables, we need to instead consider the probability "density", which might not always be the same for each value. Some ranges might be more likely, and hence the probability would be more "dense" near those values. To make this easier to understand, we need a new concept called a probability density function.

Let's look at Example 4, from Section 6.1, in which two dice were tossed and X = the sum of the two dice. The histogram below highlights P(X<6).

We can see from the histogram that P(X<6) = P(X=2) + P(X=3) + P(X=4) + P(X=5), but let's look at things a little differently. Instead of focusing on the probabilities, let's look at the area that's shaded red. The width of each rectangle is 1, so the area of each is its corresponding probability.

This leads us to another interpretation of P(X<6) - we could think of it as the area from 2 to 5. Extending that idea, we can now give a definition of a probability density function.
A **probability density function** is an equation used to compute probabilities of continuous random variables. The equation must satisfy the following two properties:

1. The total area under the graph of the equation over all possible values of the random variable must equal 1.
2. The height of the graph of the equation must be greater than or equal to 0 for all possible values of the random variable.

If we go back and consider the earlier example of the `rand()` function in Excel. Our probability density function would be fairly simple:

![Graph of a probability density function with area equal to 1]

### Probabilities as Areas

Now that we have the basic connection between area underneath the probability density function and the probability of that random variable, let's do a little further exploration.

In general, the **area under a probability density function** over a particular interval of values can have two interpretations:

- the proportion of the population with the characteristic
- the probability that a randomly selected individual will be within the interval

### The Normal Curve

Many continuous variables follow a **bell-shaped distribution** (we introduced this shape back in Section 2.2), like an individual's height, the thickness of tree bark, IQs, or the amount of light emitted by a light bulb. The more formal name of a histogram of this shape is a **normal curve**.
A continuous random variable is **normally distributed** or has a **normal probability distribution** if its relative frequency histogram has the shape of a normal curve.

In **Section 3.2**, we introduced the Empirical Rule, which said that almost all (99.7%) of the data would be within 3 standard deviations, if the distribution is bell-shaped.

We can extend this idea to the shape of other distributions. If $\mu = 0$ and $\sigma = 1$, almost all of the data should be between -3 and 3, with the center at 0. If $\mu = 0$ and $\sigma = 0.5$, almost all of the data should be between -1.5 and 1.5.

**Exploring the Shape of the Normal Curve**

To do some exploring yourself, go to the Demonstrations Project from Wolfram Research, and download the **Bell Curves demonstration**. If you haven't already, download and install the player by clicking on the image to the right.

Once you have the player installed and the Bell Curves demonstration downloaded,
move the sliders for the mean and standard deviations to get a sense of their effects on the shape.

So what effect did you see from moving the mean and standard deviation? You should have seen that moving the mean simply slides the shape left or right - it changes the center, not the spread. The standard deviation, on the other hand, changes the shape.

The key is area, which we mentioned earlier this section. Since the total area under the curve needs to still be equal to 1, if we make the distribution narrower by decreasing the standard deviation, it needs to get taller to equal the same area.

Fun with Plinko

Have you heard of the game, Plinko, from the game show The Price is Right? In this game, the contestant releases a small disc on a board covered with pegs, which direct the disc left or right. Here's a video showing a particularly successful contestant.

What's interesting is that the distribution of the Plinko chips at the bottom follows a normal distribution! Here's an example of a Java applet, showing the distribution as it might develop over hundreds of Plinko chips.

Drawing Normal Curves Using StatCrunch

Click on Stat > Calculators > Normal

Enter the mean and standard deviation (and x and the direction of the
Example 1

Most tests that gauge one's intelligence quotient (IQ) are designed to have a mean of 100 and a standard deviation of 15. It's also known that IQs are normally distributed. So what would the distribution look like for IQs?

There is no universal agreement on what IQ constitutes a "genius", though in 1916, psychologist Lewis M. Thurman set a guideline of 140 (scaled to 136 in today's tests) for "potential genius".

Suppose the area to the right of 136 is about 0.0082. What are two interpretations of that area?
Example 2

Weights of 1-year-old boys are approximately normally distributed, with a mean of 22.8 lbs and a standard deviation of about 2.15. (Source: About.com)

a. Draw a quick sketch of the normal curve for the weights of 1-year-old boys.
b. Shade the area representing the boys who are at least 20 pounds.
c. The area is approximately 0.9036. Give two interpretations of this result.

The Standard Normal Distribution

Back in Section 3.4, we introduced the idea of a z-score:

The z-score represents the number of standard deviations a data value is from the mean.

\[ Z = \frac{x - \mu}{\sigma} \]

We mentioned then that we'd need to remember the z-score later - this is that moment!

The z-score is important, because if the variable X is normally distributed, Z is as well. This brings us to an important fact:

If X is normally distributed with mean \( \mu \) and standard deviation \( \sigma \), then

\[ Z = \frac{x - \mu}{\sigma} \]

is normally distributed with a mean of 0 and a standard deviation of 1. We say that Z has the standard normal distribution.

Exploring the Standard Normal Distribution

To do some exploring yourself, go to the Demonstrations Project from Wolfram Research, and download the Area of a Normal Distribution demonstration. If you haven't already, download and install the player by clicking on the image to the right.

Once you have the player installed and the Area of a Normal Distribution demonstration downloaded, move the sliders for the mean and standard deviation of X and the value of Z to see the relationship between areas under the general normal curve and the areas under the standard normal curve.
The idea here is that the area under the normal curve on the right is equal to the area under the standard normal curve on the left.
Section 7.2: Applications of the Normal Distribution

7.1 Properties of the Normal Distribution
7.2 Applications of the Normal Distribution
7.3 Assessing Normality

Objectives

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

1. find and interpret the area under a normal curve
2. find the value of a normal random variable

Finding Areas Using a Table

Once we have the general idea of the Normal Distribution, the next step is to learn how to find areas under the curve. We'll learn two different ways - using a table and using technology.

Since every normally distributed random variable has a slightly different distribution shape, the only way to find areas using a table is to standardize the variable - transform our variable so it has a mean of 0 and a standard deviation of 1. How do we do that? **Use the z-score!**

\[ Z = \frac{x - \mu}{\sigma} \]

As we noted in Section 7.1, if the random variable \( X \) has a mean \( \mu \) and standard deviation \( \sigma \), then transforming \( X \) using the z-score creates a random variable with mean 0 and standard deviation 1! With that in mind, we just need to learn how to find areas under the standard normal curve, which can then be applied to any normally distributed random variable.

Finding Area under the Standard Normal Curve to the Left

Before we look a few examples, we need to first see how the table works. Before we start the section, you need a copy of the table. You can download a **printable copy of this table**, or use the table in the back of your textbook. It should look something like this:
It's pretty overwhelming at first, but if you look at the picture at the top (take a minute and check it out), you can see that it is indicating the area to the left. That's the key - the values in the middle represent areas to the left of the corresponding z-value. To determine which z-value it's referring to, we look to the left to get the first two digits and above to the columns to get the hundredths value. (Z-values with more accuracy need to be rounded to the hundredths in order to use this table.)

Say we're looking for the area left of -2.84. To do that, we'd start on the -2.8 row and go across until we get to the 0.04 column. (See picture.)
From the picture, we can see that the area left of -2.84 is 0.0023.

**Finding Areas Using StatCrunch**

Click on **Stat > Calculators > Normal**

Enter the mean, standard deviation, $x$, and the direction of the inequality. Then press Compute. The image below shows $P(Z < 1.23)$.

![StatCrunch Normal Distribution](image)

Let's try some examples.

**Example 1**

a. Find the area left of $Z = -0.72$

[b. Find the area left of $Z = 1.90$]

**Finding Area under the Standard Normal Curve to the Right**

To find areas to the right, we need to remember the complement rule. Take a minute and look back at the rule from *Section 5.2*.

Since we know the entire area is 1,

$$(\text{Area to the right of } z_0) = 1 - (\text{Area to the left of } z_0)$$

**Example 2**

a. Find the area to the right of $Z = -0.72$

[b. Find the area to the right of $Z = 2.68$]
An alternative idea is to use the symmetric property of the normal curve. Instead of looking to the right of $Z=2.68$ in Example 2 above, we could have looked at the area left of $-2.68$. Because the curve is symmetric, those areas are the same.

**Finding Area under the Standard Normal Curve Between Two Values**

To find the area between two values, we think of it in two pieces. Suppose we want to find the area between $Z = -2.43$ and $Z = 1.81$.

What we do instead, is find the area left of 1.81, and then subtract the area left of -2.43. Like this:

So the area between -2.43 and 1.81 = $0.9649 - 0.0075 = 0.9574$

*Note:* StatCrunch is unable to draw or calculate the "between" calculations. You'll need to perform them similarly to the one above.
Example 3

a. Find the area between $Z = 0.23$ and $Z = 1.64$.

b. Find the area between $Z = -3.5$ and $Z = -3.0$.

Finding Areas Under a Normal Curve Using the Table

1. Draw a sketch of the normal curve and shade the desired area.
2. Calculate the corresponding Z-scores.
3. Find the corresponding area under the standard normal curve.

If you remember, this is exactly what we saw happening in the Area of a Normal Distribution demonstration. Follow the link and explore again the relationship between the area under the standard normal curve and a non-standard normal curve.

Finding Areas Under a Normal Curve Using StatCrunch

Even though there's no "standard" in the title here, the directions are actually exactly the same as those from above!

Click on Stat > Calculators > Normal

Enter the mean, standard deviation, $x$, and the direction of the inequality. Then press Compute. The image below shows $P(Z < 1.23)$.

Now we finally get to the real reason we study the normal distribution. We want to be able to answer questions about variables that are normally distributed. Questions like:

- What proportion of individuals are geniuses?
- Is a systolic blood pressure of 110 unusual?
What percentage of a particular brand of light bulb emits between 300 and 400 lumens?

- What is the 90th percentile for the weights of 1-year-old boys?

All of these questions can be answered using the normal distribution!

Example 4

Let's consider again the distribution of IQs that we looked at in Example 1 in Section 7.1.

We saw in that example that tests for an individual's intelligence quotient (IQ) are designed to be normally distributed, with a mean of 100 and a standard deviation of 15.

We also saw that in 1916, psychologist Lewis M. Thurman set a guideline of 140 (scaled to 136 in today's tests) for "potential genius".

Using this information, what percentage of individuals are "potential geniuses"?

**Solution:**

1. Draw a sketch of the normal curve and shade the desired area.

2. Calculate the corresponding Z-scores.

   $$ Z = \frac{X - \mu}{\sigma} = \frac{136 - 100}{15} = 2.4 $$

3. Find the corresponding area under the standard normal curve.

   $$ P(Z>2.4) = P(Z<-2.4) = 0.0082. $$

   Based on this, it looks like about 0.82% of individuals can be characterized as "potential geniuses" according to Dr. Thurman's criteria.

Example 5

In Example 2 in Section 7.1, we were told that weights of 1-year-old boys are approximately normally distributed, with a mean of 22.8 lbs and a standard deviation of about 2.15. (Source: About.com)

If we randomly select a 1-year-old boy, what is the probability that he'll weigh at least 20 pounds?

**Solution:**
Let's do this one using technology. We should still start with a sketch:

![Normal Distribution Diagram]

Using StatCrunch, we get the following result:

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.8</td>
<td>2.15</td>
</tr>
</tbody>
</table>

\[ \text{Prob}(X \geq 20) = 0.9036 \]

According to these results, it looks like there's a probability of about 0.9036 that a randomly selected 1-year-old boy will weigh more than 20 lbs.

Why don't you try a couple?

**Example 6**

Suppose that the volume of paint in the 1-gallon paint cans produced by Acme Paint Company is approximately normally distributed with a mean of 1.04 gallons and a standard deviation of 0.023 gallons.

What is the probability that a randomly selected 1-gallon can will actually contain at least 1 gallon of paint?

[ reveal answer ]

**Example 7**

Suppose the amount of light (in lumens) emitted by a particular brand of 40W light bulbs is normally distributed with a mean of 450 lumens and a standard deviation of 20 lumens.

What percentage of bulbs emit between 425 and 475 lumens?

[ reveal answer ]

**Finding Values**
The next type of question comes from the other direction. Instead of giving values and asking for the probability, we'll now be looking at problems where the probability is known, but the values are not. Questions like:

- What is the 90th percentile for the weights of 1-year-old boys?
- What IQ score is below 80% of all IQ scores?
- What weight does a 1-year-old boy need to be so all but 5% of 1-year-old boys weight less than he does?

As with the previous types of problems, we'll learn how to do this using both the table and technology. **Make sure you know both methods - they're both used in many fields of study!**

### Finding Z-Scores Using the Table

The idea here is that the values in the table represent area to the left, so if we're asked to find the value with an area of 0.02 to the left, we look for 0.02 on the inside of the table and find the corresponding Z-score.

<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.4</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0002</td>
</tr>
<tr>
<td>-3.3</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0004</td>
<td>0.0003</td>
</tr>
<tr>
<td>-3.2</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0006</td>
<td>0.0005</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>-2.2</td>
<td>0.0139</td>
<td>0.0136</td>
<td>0.0132</td>
<td>0.0129</td>
<td>0.0125</td>
<td>0.0122</td>
<td>0.0119</td>
<td>0.0116</td>
<td>0.0113</td>
<td>0.0110</td>
</tr>
<tr>
<td>-2.1</td>
<td>0.0179</td>
<td>0.0174</td>
<td>0.0170</td>
<td>0.0166</td>
<td>0.0162</td>
<td>0.0158</td>
<td>0.0154</td>
<td>0.0150</td>
<td>0.0146</td>
<td>0.0143</td>
</tr>
<tr>
<td>-2.0</td>
<td>0.0228</td>
<td>0.0222</td>
<td>0.0217</td>
<td>0.0212</td>
<td>0.0207</td>
<td>0.0202</td>
<td>0.0197</td>
<td>0.0192</td>
<td>0.0188</td>
<td>0.0183</td>
</tr>
<tr>
<td>-1.9</td>
<td>0.0287</td>
<td>0.0281</td>
<td>0.0274</td>
<td>0.0268</td>
<td>0.0262</td>
<td>0.0256</td>
<td>0.0250</td>
<td>0.0244</td>
<td>0.0239</td>
<td>0.0233</td>
</tr>
</tbody>
</table>

Since we don't have an area of exactly 0.02, we have to think a bit. We have two choices: (1) take the closest area, or (2) average the two values if it's equidistant from the two areas.

In this case, it's almost equidistant, so we'll take the average and say that the Z-score corresponding to this area is the average of -2.05 and -2.06, so **-2.055**.

### Finding Z-Scores Using StatCrunch

Click on **Stat > Calculators > Normal**

Enter the mean, standard deviation, the direction of the inequality, and the probability (leave X blank). Then press Compute. The image below shows the Z-score with an area of 0.05 to the right.

![StatCrunch Z-Score Calculation]

Let's try a few!

### Example 8

a. Find the Z-score with an area of 0.90 to the left.

[reveal answer]

b. Find the Z-score with an area of 0.10 to the right.
c. Find the Z-score such that \( P( Z < z_0 ) = 0.025. \)

So we've talked about how to find a z-score given an area. If you remember, the technology instructions didn't specify that the distribution needed to be the standard normal - we actually find values in any normal distribution that correspond to a given area/probability using those same techniques.

**Example 9**

Referring to IQ scores again, with a mean of 100 and a standard deviation of 15. Find the 90th percentile for IQ scores.

**Solution:**

First, we need to translate the problem into an area or probability. In Section 3.4, we said the **kth percentile** of a set of data divides the lower \( k\% \) of a data set from the upper \((100-k)\%\). So the 90th percentile divides the lower 90% from the upper 10% - meaning it has about 90% below and about 10% above.

Using StatCrunch, we get the following result:

![StatCrunch result image]

Therefore, the 90th percentile for IQ scores is about 119.

**Example 10**

Suppose that the volume of paint in the 1-gallon paint cans produced by Acme Paint Company is approximately normally distributed with a mean of 1.04 gallons and a standard deviation of 0.023 gallons.

What volume can the Acme Paint Company say that 95% of their cans exceed?

[reveal answer]
Referring to the weights of 1-year-old boys again. (The weights of 1-year-old boys are approximately normally distributed, with a mean of 22.8 lbs and a standard deviation of about 2.15.)

What weight does a 1-year-old boy need to be so all but 5% of 1-year-old boys weight less than he does?

Finding $z_\alpha$

The notation $z_\alpha$ ("z-alpha") is the Z-score with an area of $\alpha$ to the right.

The concept of $z_\alpha$ is used extensively throughout the remainder of the course, so it's an important one to be comfortable with. The applications won't be immediately obvious, but the essence is that we'll be looking for events that are unlikely - and so have a very small probability in the "tail".

Let's try some examples.

a. Find $z_{0.01}$  
[b. Find $z_{0.05}$  
[c. Find $z_{0.025}$
Section 7.3: Assessing Normality

7.1 Properties of the Normal Distribution
7.2 Applications of the Normal Distribution
7.3 Assessing Normality

Objectives

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

1. find and interpret the area under a normal curve
2. find the value of a normal random variable

Earlier in the course, in Section 2.2, we learned that we can characterize the distribution shape of a random variable using a histogram. One of those distribution shapes was bell-shaped (symmetric).

Later, in Section 7.1, we defined a normally distributed random variable to be one whose histogram follows the normal (bell-shaped) curve:

So if we have the histogram, we can determine whether or not the random variable follows the normal distribution.
What happens, though, when the sample size is so small that we can’t really see the distribution shape in the histogram? We need another method, which brings us to the topic for this section.

**The Normal Probability Plot**

A normal probability plot is a graph that plots the observed data versus the *normal score*, which is what we would expect if the data actually followed the standard normal distribution.

In other words, if we have 15 observations, the 10th normal score would be the *expected 10th value if the data followed the standard normal distribution*.

We know from earlier this section that

\[ Z = \frac{x - \mu}{\sigma} \]

If we solve this equation for \( X \), we get \( X = \mu + \sigma Z \), which is the equation for a line. This gets us to the key result:

If sample data are taken from a population that is normally distributed, a normal probability plot **should be approximately linear**.

**Constructing a Normal Probability Plot Using Technology**

Unfortunately, StatCrunch doesn't have a method of producing this plot, so we'll instead be doing a Q-Q plot, which is different but offers similar results.
**Q-Q Plots in StatCrunch**

1. Import the data. To copy-paste,
   a. Copy the data from the data file.
   b. In StatCrunch, select **Data > Load Data > from paste**.
   c. Select **paste data from clipboard** and click **OK**.

2. Select **Graphics > QQ Plot**.
3. Select the column you want to plot, and click **Create Graph!**

You can also go to the [video page](#) for links to see videos in either Quicktime or iPod format.

---

**Example 1**

Suppose we wish to know whether the resting heart rates of a sample of Mth120 students are normally distributed.

<table>
<thead>
<tr>
<th>heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>61 63 64 65 65</td>
</tr>
<tr>
<td>67 71 72 73 74</td>
</tr>
<tr>
<td>75 77 79 80 81</td>
</tr>
<tr>
<td>82 83 83 84 85</td>
</tr>
<tr>
<td>86 86 89 95 95</td>
</tr>
</tbody>
</table>

Based on this plot, it does appear as though the resting heart rates are approximately normally distributed. The plot is fairly linear, with just a couple points straying from the line.

---

**Example 2**

Suppose we wish to know whether the number of children that students in a particular Mth120 class have in their family is normally distributed.

<table>
<thead>
<tr>
<th>number of children</th>
</tr>
</thead>
</table>

---

**Q-Q Plots in StatCrunch**

1. Import the data. To copy-paste,
   a. Copy the data from the data file.
   b. In StatCrunch, select **Data > Load Data > from paste**.
   c. Select **paste data from clipboard** and click **OK**.

2. Select **Graphics > QQ Plot**.
3. Select the column you want to plot, and click **Create Graph!**

You can also go to the [video page](#) for links to see videos in either Quicktime or iPod format.
This plot is clearly *not* linear, so the data do not come from a normally distributed population.
Chapter 8: Sampling Distributions

8.1 Distribution of the Sample Mean
8.2 Distribution of the Sample Proportion

Consider the following three news items.

The average price of unleaded regular fell by 1.6 cents to $3.667 a gallon on Saturday, from $3.683 a gallon, according to survey results from the motorist group AAA. (Source: CNNMoney.com)

The Census Bureau on Tuesday released the 2007 American Community Survey, the government's annual estimates of social, economic and housing characteristics for the nation. Among the highlights: 25.3 - In minutes, the average commute to work in 2007, an increase from 25.0 minutes in 2006. (Source: Chicago Tribune)

Barack Obama leads John McCain, 49% to 44%, when registered voters are asked who they would vote for if the election were held today, according to the latest Gallup Poll Daily tracking update. (Source: Gallup)

All three of these are estimates based on samples. In fact, they're probably not correct, due to sampling error. Our goal in this chapter and the next is to characterize how close we think we are. In Section 8.1, we'll talk about sample means (the first two examples above), and in Section 8.2, we'll talk about proportions (the third example).

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.
Objectives

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

1. describe the distribution of the sample mean for samples obtained from normal populations
2. describe the distribution of the sample mean for samples obtained from a population that is not normal

Section 8.1: Distribution of the Sample Mean

Sampling Distributions

Consider the following three news items.

The average price of unleaded regular fell by 1.6 cents to $3.667 a gallon on Saturday, from $3.683 a gallon, according to survey results from the motorist group AAA. (Source: CNNMoney.com)

The Census Bureau on Tuesday released the 2007 American Community Survey, the government’s annual estimates of social, economic and housing characteristics for the nation. Among the highlights: 25.3 - In minutes, the average commute to work in 2007, an increase from 25.0 minutes in 2006. (Source: Chicago Tribune)

Barack Obama leads John McCain, 49% to 44%, when registered voters are asked who they would vote for if the election were held today, according to the latest Gallup Poll Daily tracking update. (Source: Gallup)

All three of these are estimates based on samples. In fact, they’re probably not correct, due to sampling error. From Section 1.4,

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

The idea is this - unless we sample every single individual in the sample, there will be some error in our results. Our goal in this section will be to characterize the distribution of the sample mean.

The Distribution of the Sample Mean

Let’s look again at the definition of a random variable, from Section 6.1.

A random variable is a numerical measure of the outcome of a probability experiment whose value is determined by chance.

Think about the sample mean, $\bar{X}$. Isn’t it’s value determined by chance as well? Since we the individuals in a sample are randomly selected, the sample mean will depend on those individuals selected, so it, too, is a random variable. The big question, then, is the distribution of $\bar{X}$ - in other words, what are its mean (the mean of the sample mean, $\mu_{\bar{X}}$) and its standard deviation (the standard deviation of the sample mean, $\sigma_{\bar{X}}$)?

To investigate these, let’s look at a particular population.
Consider the heights of the players from the starting line-up from the 2008 Men's Olympic Basketball gold medal game - Jason Kidd (76"), LeBron James (80"), Kobe Bryant (78"), Carmelo Anthony (78"), and Dwight Howard (83"). (Source: NBC Sports) The mean of the population is 79", with a standard deviation of 2.37"

First, let's consider the different samples of size 2. There are 10 such samples (5C2 = 10), shown below, along with their corresponding sample means.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Heights</th>
<th>$\bar{X}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidd, James</td>
<td>76, 80</td>
<td>78</td>
</tr>
<tr>
<td>Kidd, Bryant</td>
<td>76, 78</td>
<td>77</td>
</tr>
<tr>
<td>Kidd, Anthony</td>
<td>76, 78</td>
<td>77</td>
</tr>
<tr>
<td>Kidd, Howard</td>
<td>76, 83</td>
<td>79.5</td>
</tr>
<tr>
<td>James, Bryant</td>
<td>80, 78</td>
<td>79</td>
</tr>
<tr>
<td>James, Anthony</td>
<td>80, 78</td>
<td>79</td>
</tr>
<tr>
<td>James, Howard</td>
<td>80, 83</td>
<td>81.5</td>
</tr>
<tr>
<td>Bryant, Anthony</td>
<td>78, 78</td>
<td>78</td>
</tr>
<tr>
<td>Bryant, Howard</td>
<td>78, 83</td>
<td>80.5</td>
</tr>
<tr>
<td>Anthony, Howard</td>
<td>78, 83</td>
<td>80.5</td>
</tr>
</tbody>
</table>

Interestingly, the mean of the sample means of size 2 is 79" as well. This is actually reasonable, though, because we know that the mean of a random variable is also its expected value, and it makes perfect sense that the value we should expect from the sample mean is the same as the population mean!

The standard deviation, though, is very different. It helps to look at things visually. The image below represents all possible sample means for samples of size 1 (individuals), 2, 3, 4, and 5 (the population). Pay particular attention to the standard deviation.

The interesting things to note here are that $\mu_{\bar{X}} = 79$, regardless of the sample size, but the standard deviation decreases as $n$ increases. If we think about this a bit, this too, is reasonable. The more individuals we have in our sample, the more likely we are to be closer to the true mean. Things brings us to our first major point.
The Law of Large Numbers
As \( n \) increases, the difference between \( \bar{x} \) and \( \mu \) approaches zero.

We're now ready to investigate the standard deviation of \( \bar{x} \) a bit more in-depth.

The Central Limit Theorem

The Central Limit Theorem
Regardless of the distribution shape of the population, the sampling distribution of \( \bar{x} \) becomes approximately normal as the sample size \( n \) increases (conservatively \( n \geq 30 \)).

This is very interesting! So it doesn't matter if the distribution shape was left-skewed, right-skewed, uniform, binomial, anything - the distribution of the sample mean will always become normal as the sample size increases. What an amazing result!

Exploring the Distribution of the Sample Mean

To do some exploring yourself, go to the Demonstrations Project from Wolfram Research, and download the Central Limit Theorem demonstration. If you haven't already, download and install the player by clicking on the image to the right.

Once you have the player installed and the Central Limit Theorem demonstration downloaded, move the slider for the sample size to get a sense of its affect on the distribution shape. You can also move the new sample slider to get a different sample.
The Distribution of the Sample Mean

We can even be more specific about the distribution of $\bar{x}$:

**The Sampling Distribution of $\bar{x}$**

If a simple random sample of size $n$ is drawn from a large population with mean $\mu$ and standard deviation $\sigma$, the sampling distribution of $\bar{x}$ will have mean and standard deviation:

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ is the **standard error of the mean**.

Now let's apply this distribution to various problems.

**Using the Central Limit Theorem**

In order to find probabilities about a normal random variable, we need to first know its mean and standard deviation. With the results of the Central Limit Theorem, we now know the distribution of the sample mean, so let's try using that in some examples.

Let's see a couple examples.

**Example 1**

Let's consider again the distribution of IQs that we looked at in Example 1 in Section 7.1.

We saw in that example that tests for an individual's intelligence quotient (IQ) are designed to be normally distributed, with a mean of 100 and a standard deviation of 15.

What is the probability that a randomly selected sample of 20 individuals would have a mean IQ of more than 105?

**Solution:**

To answer this question, we need to find $P(\bar{x} > 105)$, if $n = 20$. Before we can do that, we need to first find the distribution of $\bar{x}$. From the distribution of the sample mean, we know $\mu_{\bar{x}} = \mu = 100$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{20}} \approx 3.35$.

Here's what the distribution of $\bar{x}$ looks like in relation to the distribution of $X$. 

Now that we have the standard deviation, we can find the probability. Using StatCrunch...

![StatCrunch screenshot](source)

**Example 2**

In Example 2 in Section 7.1, we were told that weights of 1-year-old boys are approximately normally distributed, with a mean of 22.8 lbs and a standard deviation of about 2.15. (Source: About.com)

Suppose the sample mean of the 10 1-year-old boys at the Kiddie Care day care center is 22.3 lbs. Is that unusual?

**Solution:**

In order to determine if an event is *unusual*, we need to find its probability. If the probability of the event is less than 5%, we can classify it as an unusual event.

In this case, we want to find the probability of observing a sample mean of 22.3 or less. Using the distribution of the sample mean, $\mu_X = \mu = 22.8$ and $\sigma_X = \sigma/\sqrt{n} = 2.15/\sqrt{10} \approx 0.66$. Using StatCrunch...

![StatCrunch screenshot](source)

So we'd observe a sample mean of 22.3 lbs or less from a sample of 10 1-year-old boys about 23% of the time, which is not very unusual at all.

Here's one for you to try:

**Example 3**

Suppose that a particular professor's Statistics exam on probability traditionally has a mean score of 74, with a standard deviation of 11.

The professor suspects that his current crop of students is very strong. To
compare, he gives them the same exam he has in the past. The sample mean of the 28 students in his current class was a 78.

Was the professor correct? Is his current class of students unusual compared to those from the past?

[ reveal answer ]
Section 8.2: Distribution of the Sample Proportion

8.1 Distribution of the Sample Mean
8.2 Distribution of the Sample Proportion

Objectives

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

1. describe the sampling distribution of a sample proportion
2. compute probabilities of a sample proportion

The Sample Proportion

Consider these recent headlines:

**Hispanics See Their Situation in U.S. Deteriorating**
Half (50%) of all Latinos say that the situation of Latinos in this country is worse now than it was a year ago, according to a new nationwide survey of 2,015 Hispanic adults conducted by the Pew Hispanic Center. (Source: Pew Research)

**Automatic enrollment in 401(k) doesn't take care of everything**
Never got around to signing up for the company retirement plan? The boss may have done it for you. Forty-two percent of employers with 401(k) plans automatically enroll new or existing employees in the plans, nearly double the 23 percent from 2006, according to estimates from the 2008 401(k) Benchmarking Survey by the International Foundation of Employee Benefit Plans and Deloitte Consulting. The survey polled 436 employers with workforces of all sizes. (Source: Chicago Tribune)

**Stem cell, marijuana proposals lead in Mich. poll**
A recent poll shows voter support leading opposition for ballot proposals to loosen Michigan's restrictions on embryonic stem cell research and allow medical use of marijuana. The EPIC-MRA poll conducted for The Detroit News and television stations WXYZ, WILX, WOOD and WJRT found 50 percent of likely Michigan voters support the stem cell proposal, 32 percent against and 18 percent undecided. (Source: Associated Press)

These three articles all have something in common - they're referring to sample proportions - 50% of all Latinos, 42% of employers, and 50% of likely Michigan voters, respectively, in the three articles above. Proportions are the number with that certain characteristics divided by the sample size.

In general, if we let \( x \) = the number with the specific characteristic, then the sample proportion, \( \hat{p} \), (read "p-hat") is given by:

\[
\hat{p} = \frac{x}{n}
\]

Where \( \hat{p} \) is an estimate for the population proportion, \( p \).

Let's focus for a bit on \( x \), the number with that characteristic. If we rephrase that a bit, and consider an individual having that characteristic as a "success", we can see that \( x \) follows the binomial distribution.

From Section 6.2, we know that the distribution of a binomial random variable becomes bell-shaped as \( n \) increases. The three histograms below demonstrate the effect of the sample size on the distribution shape.
As the number of trials in a binomial experiment increases, the probability distribution becomes bell-shaped. As a rule of thumb, if \( np(1-p) \geq 10 \), the distribution will be approximately bell-shaped.

With our new knowledge of the normal distribution, it appears that if \( np(1-p) \geq 10 \), \( x \) is normally distributed, which implies that \( \hat{p} \) is as well. All we need to know, then, is its mean and standard deviation.

**Sampling Distribution of \( \hat{p} \)**

For a simple random sample of size \( n \) such that \( n \leq 0.05N \) (in other words, the sample is less than 5% of the population),

- The shape of the sampling distribution of \( \hat{p} \) is approximately normal provided \( np(1-p) \geq 10 \)
- The mean of the sampling distribution of \( \hat{p} \) is \( \mu_{\hat{p}} = p \).
The standard deviation of the sampling distribution of $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Now that we know how $\hat{p}$ is distributed, we can use that information to find some probabilities.

Let's see a couple examples.

**Example 1**

In a typical class, about 70% of students receive a C or better. Out of a random sample of 100 students, what is the probability that less than 60 receive a C or better?

**Solution:**

Since there are millions of students, 100 is definitely less than 5% of the population.

Because $np(1-p) = 100(0.7)(1-0.7) = 21 \geq 10$, we can say that the distribution of $\hat{p}$ is normal.

We then need to find the mean and standard deviation. From the distribution of the sample proportion, we know

$$\mu_{\hat{p}} = p = 0.7$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.7(1-0.7)}{100}} \approx 0.046$$

Using StatCrunch, the probability of observing a sample proportion of less than $60/100 = 0.6$ is

<table>
<thead>
<tr>
<th>Mean: 0.7</th>
<th>Std. Dev.: 0.046</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prob(X &lt; 0.6) = 0.014855333</td>
<td></td>
</tr>
</tbody>
</table>

So the probability of observing less than 60 out of 100 students is about 0.015 - pretty unusual.

**Example 2**

A basketball player traditionally makes 85% of her free throws. Suppose she shoots 100 free-throws during practice. Would it be unusual for her to make less than 75?

**Solution:**

Since our player can shoot an unlimited number of free throws, we can assume that 100 is less than 5% of the population. Checking the distribution:

$$np(1-p) = 100(0.85)(1-0.85) = 12.75 \geq 10$$

So the distribution should be normally distributed, with mean and
standard deviation

\[ \mu_\hat{p} = p = 0.85 \]

\[ \sigma_\hat{p} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.85(1-0.85)}{100}} \approx 0.036 \]

Using StatCrunch, with \( p = \frac{75}{100} = 0.75 \):

<table>
<thead>
<tr>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>0.036</td>
</tr>
</tbody>
</table>

| Prob(\( X \leq 0.75 \)) = 0.0027366017 |

So the probability that she makes less than 75 out of 100 is about 0.0027. That means she'll make less than 75 out of 100 about 0.27% of the time, which is very unusual.

Here's one for you to try:

**Example 3**

According to data from the American Cancer Society, about 3.86% of women develop breast cancer between the ages of 40-59.

What is the probability that in a random sample of 500 39-year-old women without breast cancer, more than 20 will develop breast cancer by the age of 60?

[ reveal answer ]

**A note on rounding:** We've already seen that rounding is an important topic in this course. In this particular section, rounding can make a significant difference in your calculations. It's important that you take a minute and watch this video regarding the effect of rounding when doing the calculations from this section: The Dangers of Rounding (Quicktime or iPod).
Chapter 9: Estimating the Value of a Parameter Using Confidence Intervals

9.1 Estimating a Population Proportion
9.2 Estimating a Population Mean
9.3 Confidence Intervals for a Population Standard Deviation
9.4 Putting It Together: Which Procedure Do I Use?

In Section 8.2, saw the following excerpt from a report by the Pew Research Foundation:

Hispanics See Their Situation in U.S. Deteriorating
Half (50%) of all Latinos say that the situation of Latinos in this country is worse now than it was a year ago, according to a new nationwide survey of 2,015 Hispanic adults conducted by the Pew Hispanic Center. (Source: Pew Research)

Further down that particular article, there's another interesting line:

The margin of error for the full sample is plus or minus 2.8 percentage points; for registered voters, the margin of error is 4.4 percentage points.

How did they determine that the margin was ±2.8%?

This is the question we'll answer in Chapter 9. In fact, we'll talk about confidence intervals for the population mean, population proportion, and population variance.
Section 9.1: Estimating a Population Proportion

9.1 Estimating a Population Proportion
9.2 Estimating a Population Mean
9.3 Confidence Intervals for a Population Standard Deviation
9.4 Putting It Together: Which Procedure Do I Use?

Objectives

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

1. construct and interpret a CI for $p$
2. determine the sample size necessary for estimating $p$ within a specified margin of error

Perhaps the most common confidence intervals that we see in the news are regarding proportions. Consider the following examples:

**Hispanics See Their Situation in U.S. Deteriorating**
Half (50%) of all Latinos say that the situation of Latinos in this country is worse now than it was a year ago, according to a new nationwide survey of 2,015 Hispanic adults conducted by the Pew Hispanic Center. [...] *The margin of error of the survey is plus or minus 2.8 percentage points at the 95% confidence level.* (Source: Pew Research)

**International Poll: No Consensus On Who Was Behind 9/11**
A new WorldPublicOpinion.org poll of 17 nations finds that majorities in only nine of them believe that al Qaeda was behind the 9/11 terrorist attacks on the United States. [...] On average, 46 percent say that al Qaeda was behind the attacks while 15 percent say the US government, seven percent Israel, and seven percent some other perpetrator. One in four say they do not know. [...] *Margins of error range from +/-3 to 4 percent.* (Source: WorldPublicOpinion)

**Stem cell, marijuana proposals lead in Mich. poll**
A recent poll shows voter support leading opposition for ballot proposals to loosen Michigan's restrictions on embryonic stem cell research and allow medical use of marijuana. The EPIC-MRA poll conducted for The Detroit News and television stations WXYZ, WILX, WOOD and WJRT found 50 percent of likely Michigan voters support the stem cell proposal, 32 percent against and 18 percent undecided. The telephone poll of 602 likely Michigan voters was conducted Sept. 22 through Wednesday. *It has a margin of sampling error of plus or minus 4 percentage points.* (Source: Associated Press)

In Section 8.2, saw the following excerpt from a report by the Pew Research Foundation:

**Hispanics See Their Situation in U.S. Deteriorating**
Half (50%) of all Latinos say that the situation of Latinos in this country is worse now than it was a year ago, according to a new nationwide survey of 2,015 Hispanic adults conducted by the Pew Hispanic Center. (Source: Pew Research)

Further down that particular article, there's another interesting line:

*The margin of error for the full sample is plus or minus 2.8 percentage points; for registered voters, the margin of error is 4.4 percentage points.*

How did they determine that the margin was ±2.8%?

The Logic of Confidence Intervals

The whole point of collecting information from a sample is to gain some information about the population. For
example, when the report says that half of all Latinos say that the situation is worse now than it was a year ago, it's not saying that they actually asked every single Latino living in the United States. Rather, it's based on a sample.

In a similar manner, consider one of the results from the American Time Use Survey:

Employed persons worked an average of 7.6 hours on the days that they worked. They worked longer on weekdays than on weekend days - 7.9 versus 5.6 hours.

The news release isn't saying that the average of time spent for all employed persons is 7.6 hours per day - they're referring to those in the sample of 12,250 individuals in the study.

Both of these examples are called point estimates. 50%, for example, is the point estimate for the percentage of all Latinos who feel that way. Similarly, the average number of hours worked per day of 7.6 is a point estimate the average number of hours worked per day for all employed persons.

A confidence interval estimate is an interval of numbers, along with a measure of the likelihood that the interval contains the unknown parameter.

The level of confidence is the expected proportion of intervals that will contain the parameter if a large number of samples is maintained. The notation we use is \((1 - \alpha)\) for the confidence interval. (This will make more sense a bit later!)

The idea of a confidence interval is this: Suppose we're wondering what proportion of ECC students are part-time. We might take a sample of 100 individuals and find a sample proportion of 65%. If we say that we're 95% confident that the real proportion is somewhere between 61% and 69%, we're saying that if we were to repeat this with new samples, and gave a margin of ±4% every time, our interval would contain the actual proportion 95% of the time.

See, every time we get a new sample, we have a new estimate, and hence a new confidence interval. Sometimes we'll be right. Sometimes we won't. The idea behind a certain % (like 95%) is that we're saying we'll be right 95% of the time. Here's a visual for 20 theoretical samples and corresponding confidence intervals for the proportion who work:
You can see that 1 of the 20 (5%) confidence intervals doesn't contain the actual value of 68.5% (based on data from elgin.edu). But the other 19 (or 95% of them) do contain the real value.

This is the idea for what we'll be doing. We'll be giving an interval of value for where we think the mean, proportion, or standard deviation is, and then a confidence level for what proportion of the intervals we believe will contain the true parameter.

**Exploring Confidence Intervals about the Population Mean**

To do some exploring yourself, go to the Demonstrations Project from Wolfram Research, and download the Confidence Intervals demonstration. If you haven't already, download and install the player by clicking on the image to the right.

Once you have the player installed and the Confidence Intervals demonstration downloaded, move the sliders for the *estimate*, *confidence level*, and *sample size* on the confidence interval and margin of error.

Even though this is about a confidence interval for the mean (and not the proportion), the idea is the same. Hopefully you noticed a few keys:

- The *estimate* simply slides the interval left and right.
- A larger *confidence level* means a wider interval - to be more confident, we need to cast a wider net in order to "catch" the actual population mean.
- A larger *sample size* means a narrower interval - the more we have in our sample, the closer we are to having the actual population mean.

**Constructing Confidence Intervals**

Before we can start constructing confidence intervals, we need to review some of the theoretical framework we set up in Chapter 8. In particular, the information about the distribution of $\hat{p}$.

**Reviewing the Distribution of the Sample Proportion**

In Section 8.2, we introduced the idea of a proportion, along with its distribution.

**Sampling Distribution of $\hat{p}$**

For a simple random sample of size $n$ such that $n \leq 0.05N$ (in other words, the sample is less than 5% of
the population), and \(np(1-p) \geq 10\), \(\hat{p}\) is approximately normally distributed, with

\[
\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}
\]

So if \(np(1-p) \geq 10\), \(\hat{p}\) will be approximately normally distributed, with the mean and standard deviation above. Using the properties of the normal distribution, that means about 95% of all sample proportions will be within 1.96 standard deviations of the mean (\(p\)).

In other words, 95% of all sample proportions will be in the interval:

\[
p - 1.96\sigma_{\hat{p}} < \hat{p} < p + 1.96\sigma_{\hat{p}}
\]

With a little algebraic manipulation, we get the following:

\[
\hat{p} - 1.96\sigma_{\hat{p}} < p < \hat{p} + 1.96\sigma_{\hat{p}}
\]

And since \(\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\), we can further get the 95% confidence interval for the proportion as:

\[
\hat{p} - 1.96 \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96 \sqrt{\frac{p(1-p)}{n}}
\]

This is a problem, though. How can we say that the population proportion, \(p\), is in this interval, when the lower and upper bounds contain \(p\)?! All is not lost! In general, we can say that as long as \(n \leq 0.05N\) and \(np(1-\hat{p}) \geq 10\), we can use \(\hat{p}\) in place of \(p\) in the standard error of the sample proportion, which gives us this for a 95% confidence interval:

\[
\hat{p} - 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

Or rephrased,
Finding a 95% Confidence Interval

**Example 1**

Polls are very common examples of confidence intervals, so let’s look at a controversial topic in Illinois - concealed carry. Illinois was the last state to allow concealed carry in public places, and the issues has been brought up frequently through polling over the years. Two 2011 polls are interesting to consider.

First, a poll sponsored by the Illinois Council Against Handgun Violence conducted March 23-27, 2011 found that 56% of likely voters statewide oppose concealed-carry legislation. **Find a 95% confidence interval for the proportion of likely voters who oppose this type of legislation.**

Interesting, another poll conducted at the same time by the Illinois State Rifle Association found that 47% of voters in four state senate districts support concealed carry legislation. **Find a 95% confidence interval for the proportion of likely voters who oppose this type of legislation.**

**Solution:**

Before we can do any analysis, we need to consider if we are able to find a confidence interval. To do that, we need the sample sizes for these polls. From the links, we can see that 600 likely voters were included in the first sample, with 957 in the second. In both cases, \( np(1-\hat{p}) \geq 10 \) and \( n \leq 0.05N \), so we can perform the confidence intervals.

For the ICAHV poll, \( \hat{p} = 56\% \), so a 95% confidence interval is:

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

So we can be 95% confident that the proportion of likely Illinois voters who oppose concealed carry legislation is between 52% and 60%.

For the ISRA poll, \( \hat{p} = 47\% \), so a 95% confidence interval is:

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

So we can be 95% confident that the proportion of voters in these districts who support concealed carry legislation is between 44% and 50%.
**Wait... what is going on?** Something is clearly wrong here. If these polls were similar, then they should have opposite results, not relatively similar for the opposite side of the issue!

So what happened? Well, the first issue is the samples - they're not drawn from the same population. Of greater concern is the questioning in the second survey. Look at the link and the question wording in general. What concerns would you have?

---

**Constructing Confidence Intervals about a Population Proportion**

What if we want to be more confident? Well, we can just replace the 1.96 with a different Z corresponding to a different area in the "tails". With that, we have the following result:

\[
\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

Note: We must have \( n\hat{p}(1-\hat{p}) \geq 10 \) and \( n \leq 0.05N \) in order to construct this interval.

---

**The Margin of Error**

Most of the time (but not always), confidence intervals look roughly like:

point estimate ± margin of error

So in the case of a confidence interval for the population proportion shown above, the margin of error is the portion after the ±, or...

The margin of error, \( E \), in a \((1-\alpha)100\%\) confidence interval for \( p \) is

\[
E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

where \( n \) is the sample size.

For more on the margin of error, watch this YouTube video, from David Longstreet:
Let's look at one of the polls above and see how they found the margin of error for their confidence interval.

**Example 2**

Consider the excerpt shown below from a poll conducted by Pew Research:

*Stem cell, marijuana proposals lead in Mich. poll*

A recent poll shows voter support leading opposition for ballot proposals to loosen Michigan's restrictions on embryonic stem cell research and allow medical use of marijuana. The EPIC-MRA poll conducted for The Detroit News and television stations WXYZ, WILX, WOOD and WJRT found 50 percent of likely Michigan voters support the stem cell proposal, 32 percent against and 18 percent undecided. The telephone poll of 602 likely Michigan voters was conducted Sept. 22 through Wednesday. *It has a margin of sampling error of plus or minus 4 percentage points.* (Source: Associated Press)

Using the confidence interval formula above, let's see if we can get the ±4% they got.

Since we want a 95% confidence level, \( \alpha = 0.05 \). The margin of error is then:

\[
E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 1.96 \cdot \sqrt{\frac{0.5(1 - 0.5)}{602}} \approx 0.04\%
\]

And so the margin of error is ±4.0%.

Why don't you try one:

**Example 3**

Read the following excerpt and calculate the margin of error for the indicated sample proportion.

*Tories lead but voter volatility on the rise*

OTTAWA–The Conservatives still hold a strong lead but shifting allegiances in Quebec and a sharp upsurge in ABC (Anybody But Conservative) thinking nationally could put a Tory majority victory out of reach, a new poll shows.

[...]

The Toronto Star/Angus Reid survey, conducted after the televised leaders’ debates, shows support for Stephen Harper’s Conservative party remains unchanged at 40 per cent, while the Liberals are slightly closing the gap, with their backing among decided voters rising to 25 per cent, according to the survey.
The poll on voting intentions, approval ratings and strategic voting surveyed 1,176 adult Canadians on Oct. 2-3 and has a margin of error of plus-or-minus [xx] percentage points, 19 times out of 20.

Finding Confidence Intervals Using StatCrunch

1. Go to **Stat** > **Proportions** > **One sample** > **with summary**.
2. Enter the number of successes and the number of observations, and press **Next**.
3. Click the **Confidence Interval** button and set the **Confidence Level**. Then press **Calculate**.

   The results should appear.

An interesting consequence of margins of errors in polls is the concept of a "statistical tie". Check out the following excerpt from a Wall Street Journal Article:

**Financial Crisis Has Little Sway in Presidential Poll**

WASHINGTON -- The race between Barack Obama and John McCain remains very tight, despite financial turmoil that has turned the nation's attention to economic issues that tend to favor the Democrats, according to a new Wall Street Journal/NBC News poll.

[...] Overall, the race remains a statistical tie, with 48% favoring Sen. Obama and his running mate, Sen. Joe Biden, and 46% favoring Sen. McCain and his vice-presidential choice, Alaska Gov. Sarah Palin. In the latest Journal poll, two weeks ago, Sen. Obama had a one-point edge. The new poll had a margin of error of plus or minus three percentage points. (Source: Wall Street Journal)

In this case, with a margin of error of ±3%, the confidence intervals actually overlap.

![Overlap of Confidence Intervals](image)

In fact, they'd overlap even if they were 49% and 45%.

![Overlap of Confidence Intervals](image)

It wouldn't be until 50% and 44% that we could say Obama would be statistically ahead of McCain.
Determining the Sample Size Needed

We sometimes need to know the sample size necessary to get a desired margin of error. The way we answer these types of questions is to go back to the margin of error:

The margin of error, \( E \), in a \((1-\alpha)\)100\% confidence interval \( p \) is

\[
E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

where \( n \) is the sample size.

If we're given the margin of error, we can solve for the sample size and get the following result:

The sample size required to obtain a \((1-\alpha)\)100\% confidence interval for \( p \) with a margin of error \( E \) is:

\[
n = \hat{p}(1-\hat{p}) \left( \frac{z_{\alpha/2}}{E} \right)^2
\]

where \( n \) is rounded up to the next integer and \( \hat{p} \) is a prior estimate of \( p \). If no prior estimate is available, use \( \hat{p} = 0.5 \).

Let's try one.

**Example 4**

Suppose you want to know what proportion of Elgin, IL residents are registered to vote. You'd like your results to be accurate to within 2 percentage points with 95\% confidence. What sample size is necessary..

a. if you use results from the US Census stating that about 72.1\% of all citizens were registered for the 2004 election? (Source: US Census)
b. you don't use a prior estimate?

[ reveal answer ]
Section 9.2: Estimating a Population Mean

9.1 Estimating a Population Proportion
9.2 Estimating a Population Mean
9.3 Confidence Intervals for a Population Standard Deviation
9.4 Putting It Together: Which Procedure Do I Use?

Objectives

By the end of this lesson, you will be able to...
1. state properties of Student’s t-distribution
2. determine t-values
3. construct and interpret a confidence interval for a population mean
4. find the sample size needed to estimate the population mean within a given margin of error

Similar to confidence intervals about proportions from Section 9.1, we can also find confidence intervals about population means. Using the same general set-up, we should have something like this:

point estimate ± margin of error

What's our point estimate for the population mean? Why, the sample mean, of course! The margin of error will be similar as well:

\[ \bar{x} \pm z_{a/2} \cdot \sigma_{\bar{x}} \]

If you recall, we discussed the distribution of \( \bar{x} \) in Chapter 8:

The Central Limit Theorem

Regardless of the distribution shape of the population, the sampling distribution of \( \bar{x} \) becomes approximately normal as the sample size n increases (conservatively n≥30).

The Sampling Distribution of \( \bar{x} \)

If a simple random sample of size n is drawn from a large population with mean \( \mu \) and standard deviation \( \sigma \), the sampling distribution of \( \bar{x} \) will have mean and standard deviation:

\[ \mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \]

where \( \sigma_{\bar{x}} \) is the standard error of the mean.

So substituting into the above formula, we get:

\[ \bar{x} \pm z_{a/2} \cdot \frac{\sigma}{\sqrt{n}} \]

There's another problem here, similar to the previous section - if we're looking for a confidence interval for the population mean, \( \mu \), how would we know \( \sigma \)? So like before, we can introduce an estimate there - in this case, the sample standard deviation, s. This introduces a lot of variability, though, since s can differ quite a bit from \( \sigma \). As a result, we need to introduce a new distribution that's similar to the standard normal, but takes into consideration some of this variability. It's called the Student's t-Distribution.
The so-called Student's distribution has an interesting history. Here's a quick summary taken from Wikipedia:

The "student's" distribution was actually published in 1908 by W. S. Gosset. Gosset, however, was employed at a brewery that forbade the publication of research by its staff members. To circumvent this restriction, Gosset used the name "Student", and consequently the distribution was named "Student t-distribution".


Gosset was trying to do research dealing with small samples. He found that even when the standard deviation was not known, the distribution of the sample means was still symmetric and similar to the normal distribution. In fact, as the sample size increases, the distribution approaches the standard normal distribution.

Student's t-Distribution

Suppose a simple random sample size n is taken from a population. If the population follows a normal distribution, then

\[ t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \]

follows a Student's t-distribution with n-1 degrees of freedom.*

* The concept of degrees of freedom is pretty abstract. One way to think of it is like this: suppose five students are choosing from five different lollipops. Each of the first four students has a choice, but the last student does not, so there are 5-1=4 degrees of freedom.

The key idea behind the t-distribution is that it has a similar shape to the standard normal distribution, but has more variability and is affected by n.

Exploring the t-Distribution

To do some exploring yourself, go to the Demonstrations Project from Wolfram Research, and download the Student's t-Distribution demonstration. If you haven't already, download and install the player by clicking on the image to the right.

Once you have the player installed and the Student's t-Distribution demonstration downloaded, uncheck the show t-cdf box (see below) and move the slider for the degrees of freedom to see the relationship between the standard normal distribution and the t-distribution.
You should notice that as the degrees of freedom increase, the distributions become more and more similar.

**Finding Critical Values**

Find critical values in the t-distribution using a table will be a bit different from finding critical values for the standard normal distribution.

Before we start the section, you need a copy of the table. You can download a printable copy of this table, or use the table in the back of your textbook. It should look something like the image below (trimmed to make it more viewable).
Example 1

Notice that unlike with the standard normal table, the t-table has probabilities along the top and critical values in the middle. This is because we primarily use the t-table to find \( t_\alpha \) - the value with \( \alpha \) area (probability) to the right.

Let's try an example.

**Example 1**

Find \( t_{0.05} \) with 21 degrees of freedom. You can use the table above, or print one out yourself. Your textbook should also come with a copy you can use.

[ reveal answer ]

**Finding Critical Values Using StatCrunch**

Click on **Stat > Calculators > T**

Enter the degrees of freedom, the direction of the inequality, and the probability (leave X blank). Then press Compute. The image below
Example 2

Use the technology of your choice to find $t_{0.01}$ with 14 degrees of freedom.

[ reveal answer ]

Constructing Confidence Intervals

Before we can start constructing confidence intervals, we need to review some of the theoretical framework we set up in Chapter 8. In particular, the information about the distribution of $\bar{x}$.

The Central Limit Theorem

Regardless of the distribution shape of the population, the sampling distribution of $\bar{x}$ becomes approximately normal as the sample size $n$ increases (conservatively $n \geq 30$).

The Sampling Distribution of $\bar{x}$

If a simple random sample of size $n$ is drawn from a large population with mean $\mu$ and standard deviation $\sigma$, the sampling distribution of $\bar{x}$ will have mean and standard deviation:

$$\mu_\bar{x} = \mu \quad \text{and} \quad \sigma_\bar{x} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_\bar{x}$ is the standard error of the mean.

Constructing a $(1-\alpha)100\%$ Confidence Interval about $\mu$

In general, a $(1-\alpha)100\%$ confidence interval for $\mu$ when $\sigma$ is unknown is

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where $t_{\alpha/2}$ is computed with $n-1$ degrees of freedom.

Note: The sample size must be large ($n \geq 30$) with no outliers or the population must be normally distributed.

Example 3

Suppose we’d like to know how many hours per week online students at ECC work. If we take a sample of 20 students and find a mean of 16.3
Example 4

In Example 1 in Section 7.4, we looked the resting heart rates of 25 Statistics students.

<table>
<thead>
<tr>
<th>heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>61  63  64  65  65</td>
</tr>
<tr>
<td>67  71  72  73  74</td>
</tr>
<tr>
<td>75  77  79  80  81</td>
</tr>
<tr>
<td>82  83  83  84  85</td>
</tr>
<tr>
<td>86  86  89  95  95</td>
</tr>
</tbody>
</table>

(Click here to view the data in a format more easily copied.)

Use the data to construct a 90% confidence interval for the true average resting heart rate of the students in this class.

Be sure to check that the conditions for creating confidence intervals are met.

Determining the Sample Size Needed

Suppose instead of trying to calculate a confidence interval given a sample mean and sample size, you are targeting a specific accuracy.

For example, say you'd like to know the average IQ of ECC students within 3 points. What sample size you would need?

The way we answer these types of questions is to go back to the margin of error definition:

The margin of error, E, in a \((1-\alpha)100\%\) confidence interval for \(\mu\) is
If we're given the margin of error, we can solve for the sample size and get \( n = \left( \frac{t_{a/2} \cdot s}{E} \right)^2 \).

Uh-oh... another problem. How can we figure out the \( t \)-value, when we need the degrees of freedom... which depends on the sample size?! The key here is that for larger sample sizes, the \( t \)-distribution approaches the standard normal distribution (think about it - as \( n \) increases, the sample standard deviation gets closer to the population standard deviation). So what we can do instead is use \( z \) to approximate \( t \).

The sample size required to estimate \( \mu \) with a \((1 - \alpha)\)100% level of confidence and a margin of error, \( E \), is:

\[
 n = \left( \frac{Z_{a/2} \cdot s}{E} \right)^2
\]

where \( n \) is rounded up to the nearest whole number.

Example 5

Let's again refer to the IQs of ECC students. How many students would we need to sample if we wanted a 95% confidence interval for the average IQ of ECC students to be within 3 points of the true population mean? (Recall that \( \sigma = 15 \) for IQs.)

[ reveal answer ]
Section 9.3: Confidence Intervals for a Population Standard Deviation

9.1 Estimating a Population Proportion
9.2 Estimating a Population Mean
9.3 Confidence Intervals for a Population Standard Deviation
9.4 Putting It Together: Which Procedure Do I Use?

Objectives

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

1. find critical values for the $\chi^2$ distribution
2. construct and interpret CIs about $\sigma^2$ and $\sigma$

The last parameters we need to find confidence intervals for are the population variance ($\sigma^2$) and standard deviation ($\sigma$).

There are many instances where we might be interested in knowing something about the spread of a population based on a sample. For example, we might believe that a particular group of students seems to have a wider variation in their grades than those from the past. Or a part manufacturer may be concerned that one of the parts it’s manufacturing is too inconsistent, even though the mean may be at specifications.

Before we can develop a confidence interval for the variance, we need another distribution.

The Chi-Square ($\chi^2$) distribution

Note: "chi-square" is pronounced "kai" as in sky, not "chai" like the tea.

The Chi-Square ($\chi^2$) distribution

If a simple random sample size $n$ is obtained from a normally distributed population with mean $\mu$ and standard deviation $\sigma$, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with $n-1$ degrees of freedom.
Properties of the $X^2$ distribution

1. It is not symmetric.
2. The shape depends on the degrees of freedom.
3. As the number of degrees of freedom increases, the distribution becomes more symmetric.
4. $X^2 \geq 0$

Finding Critical Values

Find critical values in the $X^2$ distribution using a table is done in the same manner we found critical values for the $t$-distribution.

Before we start the section, you need a copy of the table. You can download a printable copy of this table, or use the table in the back of your textbook. It should look something like the image below (trimmed to make it more viewable).

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>0.99</th>
<th>0.975</th>
<th>0.95</th>
<th>0.90</th>
<th>0.10</th>
<th>0.05</th>
<th>0.025</th>
<th>0.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>—</td>
<td>0.001</td>
<td>0.004</td>
<td>0.016</td>
<td>2.706</td>
<td>3.841</td>
<td>5.024</td>
<td>6.635</td>
</tr>
<tr>
<td>2</td>
<td>0.200</td>
<td>0.051</td>
<td>0.103</td>
<td>0.211</td>
<td>4.605</td>
<td>5.991</td>
<td>7.378</td>
<td>9.210</td>
</tr>
<tr>
<td>3</td>
<td>0.115</td>
<td>0.216</td>
<td>0.352</td>
<td>0.584</td>
<td>6.251</td>
<td>7.815</td>
<td>9.348</td>
<td>11.345</td>
</tr>
<tr>
<td>4</td>
<td>0.297</td>
<td>0.484</td>
<td>0.711</td>
<td>1.064</td>
<td>7.779</td>
<td>9.488</td>
<td>11.143</td>
<td>13.277</td>
</tr>
<tr>
<td>5</td>
<td>0.554</td>
<td>0.831</td>
<td>1.145</td>
<td>1.610</td>
<td>9.236</td>
<td>11.071</td>
<td>12.833</td>
<td>15.086</td>
</tr>
<tr>
<td>6</td>
<td>0.872</td>
<td>1.237</td>
<td>1.635</td>
<td>2.204</td>
<td>10.645</td>
<td>12.592</td>
<td>14.449</td>
<td>16.812</td>
</tr>
<tr>
<td>7</td>
<td>1.239</td>
<td>1.690</td>
<td>2.167</td>
<td>2.833</td>
<td>12.017</td>
<td>14.067</td>
<td>16.013</td>
<td>18.475</td>
</tr>
<tr>
<td>8</td>
<td>1.646</td>
<td>2.180</td>
<td>2.733</td>
<td>3.490</td>
<td>13.362</td>
<td>15.507</td>
<td>17.535</td>
<td>20.090</td>
</tr>
<tr>
<td>13</td>
<td>4.107</td>
<td>5.009</td>
<td>5.892</td>
<td>7.042</td>
<td>19.812</td>
<td>22.362</td>
<td>24.736</td>
<td>27.688</td>
</tr>
<tr>
<td>15</td>
<td>5.229</td>
<td>6.262</td>
<td>7.261</td>
<td>8.547</td>
<td>22.307</td>
<td>24.996</td>
<td>27.488</td>
<td>30.578</td>
</tr>
<tr>
<td>16</td>
<td>5.812</td>
<td>6.908</td>
<td>7.962</td>
<td>9.312</td>
<td>23.542</td>
<td>26.296</td>
<td>28.845</td>
<td>32.000</td>
</tr>
<tr>
<td>17</td>
<td>6.408</td>
<td>7.564</td>
<td>8.672</td>
<td>10.085</td>
<td>24.769</td>
<td>27.587</td>
<td>30.191</td>
<td>33.409</td>
</tr>
<tr>
<td>19</td>
<td>7.633</td>
<td>8.907</td>
<td>10.117</td>
<td>11.651</td>
<td>27.204</td>
<td>30.144</td>
<td>32.852</td>
<td>36.191</td>
</tr>
<tr>
<td>21</td>
<td>8.897</td>
<td>10.283</td>
<td>11.591</td>
<td>13.240</td>
<td>29.615</td>
<td>32.671</td>
<td>35.479</td>
<td>38.932</td>
</tr>
<tr>
<td>23</td>
<td>10.196</td>
<td>11.689</td>
<td>13.091</td>
<td>14.848</td>
<td>32.007</td>
<td>35.172</td>
<td>38.076</td>
<td>41.638</td>
</tr>
<tr>
<td>24</td>
<td>10.856</td>
<td>12.401</td>
<td>13.848</td>
<td>15.659</td>
<td>33.196</td>
<td>36.415</td>
<td>39.364</td>
<td>42.980</td>
</tr>
<tr>
<td>25</td>
<td>11.524</td>
<td>13.120</td>
<td>14.611</td>
<td>16.473</td>
<td>34.382</td>
<td>37.652</td>
<td>40.646</td>
<td>44.314</td>
</tr>
<tr>
<td>26</td>
<td>12.198</td>
<td>13.844</td>
<td>15.379</td>
<td>17.292</td>
<td>35.563</td>
<td>38.885</td>
<td>41.923</td>
<td>45.642</td>
</tr>
<tr>
<td>27</td>
<td>12.879</td>
<td>14.573</td>
<td>16.151</td>
<td>18.114</td>
<td>36.741</td>
<td>40.113</td>
<td>43.194</td>
<td>46.963</td>
</tr>
<tr>
<td>28</td>
<td>13.565</td>
<td>15.308</td>
<td>16.928</td>
<td>18.939</td>
<td>37.916</td>
<td>41.337</td>
<td>44.461</td>
<td>48.278</td>
</tr>
</tbody>
</table>

Notice that the table is similar to the $t$-table, with probabilities along the top and critical values in the middle. This is because we primarily use the chi-square-table to find critical values.

Let's try an example.

Example 1
Find the critical values in the $X$ distribution which separate the middle 95% from the 2.5% in each tail, assuming there are 12 degrees of freedom.

You can use the table above, or print one out yourself. Your textbook should also come with a copy you can use.

### Finding Critical Values Using StatCrunch

Click on **Stat > Calculators > Chi-Square**

Enter the degrees of freedom, the direction of the inequality, and the probability (leave X blank). Then press **Compute**.

### Example 2

Use the technology of your choice to find $\chi^2_{0.01}$ with 20 degrees of freedom.

### Constructing Confidence Intervals about $\sigma^2$ and $\sigma$

Now that we have the basics of the distribution of the variable $\chi^2$, we can work on constructing a formula for the confidence interval.

From the distribution shape on the previous page, we know that $(1 - \alpha)100\%$ of the $\chi^2$ values will be between the two critical values shown below.

This gives us the following inequality:

$$\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}$$

If we solve the inequality for $\sigma^2$, we get the formula for the confidence interval:

A $(1-\alpha)100\%$ confidence interval for $\sigma^2$ is
\[
\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}
\]

Note: The sample must be taken from a normally distributed population.

Note #2: If a confidence interval for \( \sigma \) is desired, we can take the square root of each part.

Now that we have the confidence interval formula, let's try a couple examples.

**Example 3**

Suppose a sample of 30 ECC students are given an IQ test. If the sample has a standard deviation of 12.23 points, find a 90% confidence interval for the population standard deviation.

**Solution:** We first need to find the critical values:

\[\chi^2_{1-\alpha/2} = \chi^2_{0.95,29} \approx 17.706\text{ and } \chi^2_{\alpha/2} = \chi^2_{0.05,29} \approx 42.557\]

Then the confidence interval is:

\[
\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}
\]

\[
\frac{(30-1)12.23^2}{42.557} < \sigma^2 < \frac{(30-1)12.23^2}{17.706}
\]

\[101.9249 < \sigma^2 < 244.9472\]

\[10.10 < \sigma < 15.65\]

So we are 90% confident that the standard deviation of the IQ of ECC students is between 10.10 and 15.65 bpm.

**Finding Confidence Intervals Using StatCrunch**

1. Click on **Stat > Variance** > **One sample**
2. Select **with data** if you have the data, or **with summary** if you only have the summary statistics.
3. If you chose **with data**, click on the variable that you want for the confidence interval. Otherwise, enter the sample statistics.
4. Click on **Next**.
5. Click the Confidence Interval radio button
6. Enter the desired level of confidence and press **Calculate**

The confidence interval should be displayed.

Note: If you need a confidence interval about the population standard deviation, take the square root of the values in the resulting confidence interval.

Here's one for you to try:

**Example 4**

In **Example 3** in Section 9.1, we assumed the standard deviation of the
resting heart rates of students was 10 bpm.

<table>
<thead>
<tr>
<th>heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>61 63 64  65 65</td>
</tr>
<tr>
<td>67 71 72  73 74</td>
</tr>
<tr>
<td>75 77 79  80 81</td>
</tr>
<tr>
<td>82 83 83  84 85</td>
</tr>
<tr>
<td>86 86 89  95 95</td>
</tr>
</tbody>
</table>

(Click here to view the data in a format more easily copied.)

Use StatCrunch to find a 95% confidence interval for the standard deviation of the resting heart rates for students in this particular class.

[ reveal answer ]
Section 9.4: Putting It Together: Which Procedure Do I Use?

9.1 Estimating a Population Proportion
9.2 Estimating a Population Mean
9.3 Confidence Intervals for a Population Standard Deviation
9.4 Putting It Together: Which Procedure Do I Use?

**Objectives**

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

1. determine the appropriate CI to construct

In this chapter, we’ve learned three different confidence intervals. In order for them to be of any value, we need to know which one to apply. Let’s do a quick review of the three.

**Confidence Intervals about \( p \)**

This is probably the easiest one - whenever we’re looking at a proportion (percent), this is the confidence interval we want.

A \((1-\alpha)100\%\) confidence interval for \( p \) is

\[
\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

Note: We must have \( np(1-p) \geq 10 \) and \( n \leq 0.05N \) in order to construct this interval.

**Confidence Intervals about \( \mu \)**

This is the typical confidence interval for a mean. Use this when you’re given information from a sample or if you’re only given data. (In that case, you calculate the sample mean and standard deviation yourself, so you can’t possibly know the population standard deviation.)

In general, a \((1-\alpha)100\%\) confidence interval for \( \mu \) when \( \sigma \) is unknown is

\[
\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}
\]

where \( t_{\alpha/2} \) is computed with \( n-1 \) degrees of freedom.

Note: The sample size must be large (\( n \geq 30 \)) or the population must be normally distributed.

**Confidence Intervals about \( \sigma^2 \) and \( \sigma \)**

The last confidence interval is for either the variance or standard deviation. You should be able to key on those words to help you recognize this confidence interval.
A \((1-\alpha)100\%\) confidence interval for \(\sigma^2\) is
\[
\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}
\]
Note: The sample must be taken from a normally distributed population.

Some Examples

Let's try some examples. For each situation, simply determine which confidence interval should be used. You should specify the parameter and state the appropriate interval formula (because there are two for the population mean).

Example 1
We would like to know the fraction of ECC students who commute to school from their parents' homes. We send emails to students using their ECC email account until 100 have responded; 62 of the responders were commuters.

Find a 95% confidence interval for the fraction of ECC students who commute to school from their parents' homes.

[ reveal answer ]

Example 2
Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in °C) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 on 6 different samples of the liquid. He calculates the sample mean to be 101.82. If he knows that the standard deviation for this procedure is 1.2°, what is the confidence interval for the population mean at a 95% confidence level?

[ reveal answer ]

Example 3
To assess scholastic performance, a state administers an achievement test to a simple random sample of 100 high school seniors. The mean score of the students who took the exam is 99.7 points, with a standard deviation of 7.9 points. Find an approximate 90% confidence interval for the average of the population scores that would have been obtained had every high school senior in the state been administered the achievement test.

[ reveal answer ]

Example 4
We know from previous examples that the standard deviation of IQs is normally distributed with a standard deviation of 15. Suppose we wonder if the IQs of ECC students have more variation. To answer this question, we collect the IQs from a random sample of ECC students and find a standard deviation of 16.2.
Based on this information, do you believe with 95% confidence that the IQs of ECC students have more variation than the population?
Chapter 10: Testing Claims Regarding a Parameter

10.1 The Language of Hypothesis Testing
10.2 Hypothesis Tests for a Population Proportion
10.3 Hypothesis Tests for a Population Mean
10.4 Hypothesis Tests for a Population Standard Deviation
10.5 Putting It Together: Which Method Do I Use?

Let's take a minute to refresh the process of statistics we introduced earlier this semester.

Last chapter, we finally entered the "meat" of the course, where we use our knowledge of calculated statistics (like the sample mean and standard deviation) and our knowledge of probability (the normal distribution, et al) to come to conclusions - confidence intervals.

This chapter, we'll be expanding that knowledge and answering questions about population parameters based on sample data. For example:

- Do online students work more than traditional students?
- Are online students younger than the average ECC student?
- Do ECC students vote at a higher rate than the general population?
Section 10.1: The Language of Hypothesis Testing

10.1 The Language of Hypothesis Testing
10.2 Hypothesis Tests for a Population Proportion
10.3 Hypothesis Tests for a Population Mean
10.4 Hypothesis Tests for a Population Standard Deviation
10.5 Putting It Together: Which Method Do I Use?

Objectives

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

1. determine the null and alternative hypotheses from a claim
2. explain Type I and Type II errors
3. identify whether an error is Type I or Type II
4. state conclusions to hypothesis tests

The Nature of Hypothesis Testing

According to the US Census Bureau, 64% of US citizens age 18 or older voted in the 2004 election. Suppose we believe that percentage is higher for ECC students for the 2008 presidential election. To determine if our suspicions are correct, we collect information from a random sample of 500 ECC students. Of those, 460 were citizens and 18 or older in time for the election. (We have some students still in high school, and some who do not yet have citizenship.) Of those who were eligible to vote, 326 (or about 71%) say that they did vote.

Problem: Based on this random sample, do we have enough evidence to say that the percentage of ECC students who were eligible to vote and did vote in the 2008 presidential election was higher than the proportion of US citizens who voted in the 2004 election?

Solution: 71% of the students in our sample voted. Obviously, this is higher than the national average. The thing to consider, though, is that this is just a sample of ECC students - it isn't every student. It's possible that the students who just happened to be in our sample were those who voted. Maybe this could just happen randomly. In order to determine if it really is that different from the national proportion, we need to find out how probable a sample proportion of 71% would be if the true proportion was really 64%.

To answer this, we consider the distribution of the proportion of eligible voters who did vote. From Section 8.2, we know that this proportion is approximately normally distributed if np(1-p) ≥ 10. Using the techniques from that section, we also know that the mean would be 64%, with a standard deviation of about 2.2%. With this information, the probability of observing a random sample with a proportion of 71% if the true proportion is 64% is about 0.001.

This means that about 1 in 1000 random samples will have a proportion that large. So we have two conclusions:

1. We just observed an extremely rare event, or
2. The proportion of ECC students who vote is actually higher than 64%.

This is the idea behind hypothesis testing. The general process is this:

Steps in Hypothesis Testing

1. A claim is made. (More than 64% of ECC students vote, in our case.)
2. Evidence is collected to test the claim. (We found that 326 of 460 voted.)
3. The data are analyzed to assess the plausibility of the claim. (We determined that the proportion is most likely higher than 64% for ECC students.)
Null and Alternative Hypotheses

In statistics, we call these claims *hypotheses*. We have two types of hypotheses, a **null hypothesis** and an **alternative hypothesis**.

The **null hypothesis**, denoted $H_0$ ("H-naught"), is a statement to be tested. It is usually the status quo, and is assumed true until evidence is found otherwise.

The **alternative hypothesis**, denoted $H_1$ ("H-one"), is a claim to be tested. We will try to find evidence to support the alternative hypothesis.

There are three general ways in this chapter that we'll set up the null and alternative hypothesis.

1. **two tailed**
   - $H_0$: parameter = some value
   - $H_1$: parameter ≠ the value

2. **left-tailed**
   - $H_0$: parameter = some value
   - $H_1$: parameter < the value

3. **right-tailed**
   - $H_0$: parameter = some value
   - $H_1$: parameter > the value

Let's look an example of each.

**Example 1**

In the introduction of this section, we were considering the proportion of ECC students who voted in the 2008 presidential election. We assumed that it was the same as the national proportion in 2004 (64%), and tried to find evidence that it was higher than that. In that case, our null and alternative hypotheses would be:

$H_0: p = 0.64$

$H_1: p > 0.64$

**Example 2**

According to the Elgin Community College website, the average age of ECC students is 28.2 years. We might claim that the average is less for online Mth120 students. In that case, our null and alternative hypotheses would be:

$H_0: \mu = 28$

$H_1: \mu < 28$

**Example 3**

It's fairly standard knowledge that IQ tests are designed to be normally distributed, with an average of 100. We wonder whether the IQ of ECC students is different from this average. Our hypotheses would then be:

$H_0: \mu = 100$

$H_1: \mu \neq 100$
The Four Outcomes of Hypothesis Testing

Unfortunately, we never know with 100% certainty what is true in reality. We always make our decision based on sample data, which may or may not reflect reality. So we'll make our decision, but it may not always be correct.

In general, there are four possible outcomes from a hypothesis test when we compare our decision with what is true in reality - which we will never know!

- We could decide to not reject the null hypothesis when in reality the null hypothesis was true. This would be a correct decision.
- We could reject the null hypothesis when in reality the alternative hypothesis is true. This would also be a correct decision.
- We could reject the null hypothesis when it really is true. We call this error a **Type I error**.
- We could decide to not reject the null hypothesis, when in reality we should have, because the alternative was true. We call this error a **Type II error**.

<table>
<thead>
<tr>
<th>reality</th>
<th>H₀ true</th>
<th>H₁ true</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>decis</strong></td>
<td>do not reject H₀</td>
<td>correct decision</td>
</tr>
<tr>
<td><strong>reject H₀</strong></td>
<td>Type I error</td>
<td>correct decision</td>
</tr>
</tbody>
</table>

To help illustrate the idea, let's look at an example.

**Example 4**

Let's consider a pregnancy test. The tests work by looking for the presence of the hormone human chorionic gonadotropin (hCG), which is secreted by the placenta after the fertilized egg implants in a woman's uterus.

If we consider this in the language of a hypothesis test, the null hypothesis here is that the woman is not pregnant - this is what we assume is true until proven otherwise.

The corresponding chart for this test would look something like this:

<table>
<thead>
<tr>
<th>pregnancy test</th>
<th>reality</th>
</tr>
</thead>
<tbody>
<tr>
<td>not pregnant</td>
<td>H₀ true</td>
</tr>
<tr>
<td>not pregnant</td>
<td>do not reject H₀</td>
</tr>
<tr>
<td>pregnant</td>
<td>reject H₀</td>
</tr>
</tbody>
</table>

In this case, we would call a Type I error a "false positive" - the test was positive for pregnancy, when in reality the mother was not pregnant.

The Type II error in this context would be a "false negative" - the test did not reveal the pregnancy, when the woman really was pregnant.

When a test claims that it is "99% Accurate at Detecting Pregnancies", it is referring to Type II errors. The tests claim to detect 99% of pregnancies, so it will make a Type II error (not detecting the pregnancy) only 1% of the time.

Note: The "99% Accurate" claim is not entirely correct. Many tests do not
have this accuracy until a few days after a missed period. (Source: US Dept. of Health and Human Services)

**Level of Significance**

As notation, we assign two Greek letters, $\alpha$ ("alpha") and $\beta$ ("beta"), to the probability of Type I and Type II errors, respectively.

$$\alpha = P(\text{Type I error}) = P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true})$$

$$\beta = P(\text{Type II error}) = P(\text{not rejecting } H_0 \text{ when } H_1 \text{ is true})$$

Unfortunately, we can't control both errors, so researchers choose the probability of making a Type I error before the sample is collected. We refer to this probability as the **level of significance** of the test.

**Choosing $\alpha$, the Probability of a Type I Error**

One interesting topic is the choice of $\alpha$. How do we choose? The answer to that is to consider the consequences of the mistake. Consider two examples:

---

**Example 5**

Let's consider again a murder trial, with a possibility death penalty if the defendant is convicted. The null and alternative hypotheses again are:

- $H_0$: the defendant is innocent
- $H_1$: the defendant is guilty

In this case, a Type I error (rejecting the null when it really is true) would be when the jury returns a "guilty" verdict, when the defendant is actually innocent.

In our judicial system, we say that we must be sure "beyond a reasonable doubt". And in this case in particular, the consequences of a mistake would be the death of an innocent defendant.

Clearly in this case, we want the probability of making this error very small, so we might assign to $\alpha$ a value like 0.00001. (So about 1 out of every 100,000 such trials will result in an incorrect guilty verdict. You might choose an even smaller value - especially if you are the defendant in question!)

---

**Example 6**

As an alternative, suppose we're considering the average age of ECC students. As stated in Example 2, we assumed that average age is 28, but we think it might be lower for online Mth120 students.

- $H_0$: $\mu = 28$
- $H_1$: $\mu < 28$

In this example, $\alpha = P(\text{concluding that the average age is less than 28, when it really is 28})$. The consequences of that mistake are... that we're wrong about the average age. Certainly nothing as dire as sending an innocent defendant to jail - or worse.

Because the consequences of a Type I error or not nearly as dire, we might assign a value of $\alpha = 0.05$. 

Example 7

Let's look at a different example. Suppose we have a test for the deadly disease statisticitis, which affects approximately 5% of the general population.

**Problem:** If a test is developed that claims to be 90% accurate, what is the probability that an individual with a positive test result for statisticitis actually has the disease?

**Solution:** One good way to do this is to start with 1000 theoretical individuals. Since we know 5% of the population has the disease, 5% of the 1000, or 50, must have the disease. We can illustrate this in a chart like the ones we did earlier:

<table>
<thead>
<tr>
<th>Reality</th>
<th>No Disease</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Result</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>950</td>
<td>50</td>
<td>1000</td>
</tr>
</tbody>
</table>

We're also told that the test is "90% accurate", so 90% of those 50 with the disease will have positive test results. That gives us something like the following:

<table>
<thead>
<tr>
<th>Reality</th>
<th>No Disease</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Result</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>0.1*50 = 5</td>
<td>0.9*50 = 45</td>
<td></td>
</tr>
<tr>
<td>950</td>
<td>50</td>
<td>1000</td>
</tr>
</tbody>
</table>

Now, we also have to consider the test being "90% accurate" and that it should return a negative result for 90% of the 950 who do not have the disease.

<table>
<thead>
<tr>
<th>Reality</th>
<th>No Disease</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Result</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>0.9*950 = 855</td>
<td>0.1*950 = 95</td>
<td></td>
</tr>
<tr>
<td>950</td>
<td>50</td>
<td>1000</td>
</tr>
</tbody>
</table>

So in summary, we have the following table:

<table>
<thead>
<tr>
<th>Reality</th>
<th>No Disease</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Result</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>855</td>
<td>5</td>
<td>860</td>
</tr>
<tr>
<td>95</td>
<td>45</td>
<td>140</td>
</tr>
<tr>
<td>950</td>
<td>50</td>
<td>1000</td>
</tr>
</tbody>
</table>

We were asked originally for the probability that an individual with a positive test actually has the disease. In other words, we want:
\[ P(\text{has disease | positive test}) = \frac{45}{140} \approx 0.32 \]

What does this mean? Well, if this test is "90% accurate", then a positive test result actually only means you have the disease about 32% of the time!

On the other hand, it's interesting to note that a negative result is fairly accurate - \( \frac{855}{860} \approx 0.994 = 99.4\% \). So you can feel pretty confident of a negative result.

This may seem unreal, but some tests really do work this way. What is done to balance this out is to treat a "positive" test result as not definitive - it simply warrants more study.

The test for gestational diabetes is a perfect example. The test is performed in two stages - an initial "screening", which is extended to a second test if the result is positive.

---

**Stating Conclusions to Hypothesis Tests**

We have to be very careful when we state our conclusions. The way this type of hypothesis testing works, we look for evidence to support the alternative claim. If we find it, we say we have enough evidence to support the alternative. If we don't find that evidence, we say just that - we don't have enough evidence to support the alternative hypothesis.

We **never support the null hypothesis**. In fact, the null hypothesis is probably not *exactly* true. There are some very interesting discussions of this type of hypothesis testing (see links at the end of this page), and one criticism is that large enough sample will probably show a difference from the null hypothesis.

That said, we look for evidence to support the alternative hypothesis. If we don't find it, then we simply say that we don't have enough evidence to support the alternative hypothesis.

To illustrate, consider the three examples from earlier this section:

**Example 8**

Let's consider Example 1, introduced earlier this section. The null and alternative hypotheses were as follows:

- \( H_0: p = 0.64 \)
- \( H_1: p > 0.64 \)

Suppose we decide that the evidence supports the alternative claim (it does). Our conclusion would then be:

*There is enough evidence to support the claim that the proportion of ECC students who were eligible to vote and did was more than the national proportion of 64% who voted in 2004.*

**Example 9**

In Example 2, we stated null and alternative hypotheses:

- \( H_0: \mu = 28 \)
- \( H_1: \mu < 28 \)

Suppose we find an average age of 20.4, and we decide to reject the null hypothesis. Our conclusion would then be:

*There is enough evidence to support our claim that the average age of online Mth120 students is less than 28 years.*
Example 10

In Example 3, we stated null and alternative hypotheses:

\[ H_0: \mu = 100 \]
\[ H_1: \mu \neq 100 \]

Suppose we find an average IQ from a sample of 30 students to be 101, which we determine is not significantly different from the assumed mean. Our conclusion would then be:

*There is not enough evidence to support the claim that the average IQ of ECC students is not 100.*

The Controversy Regarding Hypothesis Testing

Unlike other fields in mathematics, there are many areas in statistics which are still being debated. Hypothesis Testing is one of them. There are several concerns that any good statistician should be aware of.

- Tests are significantly affected by sample size
- Not rejecting the null hypothesis does not mean it is true
- Statistical significance does not imply practical significance
- Repeated tests can be misleading (see this illuminating comic from XKCD)

Much more subtly, the truth is likely that the null hypothesis is *never* true, and rejecting it is only a matter of getting a large enough sample. Consider Example 6, in which we assumed the average age was 28. Well... to how many digits? Isn't it likely that the average age is actually 27.85083 or something of the sort? In that case, all we need a large enough sample to get a sample mean that is statistically less than 28, even though the difference really has no practical meaning.

If you're interested in reading more about some of the weaknesses of this method for testing hypotheses, visit these links:

- Commentaries on Significance Testing
- The Concept of Statistical Significance Testing
- Null-Hypothesis Controversy
- The Controversy Over How to Present Research Findings

So Why Are We Studying Hypothesis Testing?

The reality is that we need some way to analytically make decisions regarding our observations. While it is simplistic and certainly contains errors, hypothesis testing still does have value, provided we understand its limitations.

This is also an introductory statistics course. There are limits to what we can learn in a single semester. There are more robust ways to perform a hypothesis test, including effect size, power analysis, and Bayesian inference. Unfortunately, many of these are simply beyond the scope of this course.

The point of this discussion is to be clear that hypothesis testing has weaknesses. By understanding them, we can make clear statements about the results of a hypothesis test and what we can actually conclude.
Section 10.2: Hypothesis Tests for a Population Proportion

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

1. explain the logic of hypothesis testing
2. test hypotheses about a population proportion
3. test hypotheses about a population proportion using the binomial probability distribution

The Logic of Hypothesis Testing

Once we have our null and alternative hypotheses chosen, and our sample data collected, how do we choose whether or not to reject the null hypothesis? In a nutshell, it's this:

If the observed results are unlikely assuming that the null hypothesis is true, we say the result is statistically significant, and we reject the null hypothesis. In other words, the observed results are so unusual, that our original assumption in the null hypothesis must not have been correct.

Your textbook references three different methods for testing hypotheses:

- the classical approach
- *P*-values
- confidence intervals

Because *P*-values are so much more widely used, we will be focusing on this method. You will be required to include *P*-values for your homework and exams.

If you're interested in learning any of these other methods, feel free to read through the textbook.

**P-Values**

In general, we define the *P*-value this way:

The *P*-value is the probability of observing a sample statistic as extreme or more extreme than the one observed in the sample assuming that the null hypothesis is true.

**The Sample Proportion**

In Section 8.2, we learned about the distribution of the sample proportion, so let's do a quick review of that now.

In general, if we let *x* = the number with the specific characteristic, then the sample proportion, \( \hat{p} \), (read "p-hat") is given by:
\[ \hat{p} = \frac{x}{n} \]

Where \( \hat{p} \) is an estimate for the population proportion, \( p \).

We also learned some information about how the sample proportion is distributed:

**Sampling Distribution of \( \hat{p} \)**

For a simple random sample of size \( n \) such that \( n \leq 0.05N \) (in other words, the sample is less than 5% of the population),

- The shape of the sampling distribution of \( \hat{p} \) is approximately normal provided \( np(1-p) \geq 10 \)
- The mean of the sampling distribution of \( \hat{p} \) is \( \mu_{\hat{p}} = p \).
- The standard deviation of the sampling distribution of \( \hat{p} \) is

\[ \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \]

Why are these important? Well, suppose we take a sample of 100 online students, and find that 74 of them are part-time. You might recall that based on data from elgin.edu, 68.5% of ECC students in general are part-time. So is observing 74% of our sample unusual? How do we know - we need the distribution of \( \hat{p} \)!

So what we do is create a **test statistic** based on our sample, and then use a table or technology to find the probability of what we observed. Here are the details.

**Testing Claims Regarding the Population Proportion Using \( P \)-Values**

In this first section, we assume we are testing some claim about the population proportion. As usual, the following two conditions must be true:

1. \( np(1-p) \geq 10 \), and
2. \( n \leq 0.05N \)

**Step 1:** State the null and alternative hypotheses.

**Two-Tailed**

\[ H_0: p = p_0 \quad H_1: p \neq p_0 \]

**Left-Tailed**

\[ H_0: p = p_0 \quad H_1: p < p_0 \]

**Right-Tailed**

\[ H_0: p = p_0 \quad H_1: p > p_0 \]

**Step 2:** Decide on a level of significance, \( \alpha \), depending on the seriousness of making a Type I error. (\( \alpha \) will often be given as part of a test or homework question, but this will not be the case in the outside world.)

**Step 3:** Compute the test statistic, \( z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \)

**Step 4:** Determine the \( P \)-value.

**Step 5:** Reject the null hypothesis if the \( P \)-value is less than the level of significance, \( \alpha \).

**Step 6:** State the conclusion. (You should also include a measure of the **strength** of the results, based on the \( P \)-value.)
Calculating $P$-Values

Right-Tailed Tests

Left-Tailed Tests

Two-Tailed Tests

In a two-tailed test, the $P$-value = $2P(Z \gt |z_o|)$.

It may seem odd to multiply the probability by two, since "or more extreme" seems to imply the area in the tail only. The reason why we do multiply by two is that even though the result was on one side, we didn't know before collecting the data, on which side it would be.

The Strength of the Evidence
Since the $P$-value represents the probability of observing our result or more extreme, the smaller the $P$-value, the more unusual our observation was. Another way to look at it is this:

The smaller the $P$-value, the stronger the evidence supporting the alternative hypothesis. We can use the following guideline:

- $P$-value $< 0.01$: **very strong** evidence supporting the alternative hypothesis
- $0.01 \leq P$-value $< 0.05$: **strong** evidence supporting the alternative hypothesis
- $0.05 \leq P$-value $< 0.1$: **some** evidence supporting the alternative hypothesis
- $P$-value $\geq 0.1$: **weak to no** evidence supporting the alternative hypothesis

These values are not hard lines, of course, but they can give us a general idea of the strength of the evidence.

**But wait!** There is an important caveat here, which was mentioned earlier in the section about The Controversy Regarding Hypothesis Testing. The problem is that it's relatively easy to get a large $p$-value - just get a really large sample size! So the chart above is really with the caveat "assuming equal sample sizes in comparable studies, ..."

This isn't something every statistics text will mention, nor will every instructor mention, but it's important.

**Example 1**

According to the Elgin Community College website, approximately 56% of ECC students are female. Suppose we wonder if the same proportion is true for math courses. If we collect a sample of 200 ECC students enrolled in math courses and find that 105 of them are female, do we have enough evidence at the 10% level of significance to say that the proportion of math students who are female is different from the general population?

Note: Be sure to check that the conditions for performing the hypothesis test are met.

**Hypothesis Testing Regarding $p$ Using StatCrunch**

1. Go to Stat > Proportions > One sample > with summary.
2. Enter the number of successes and the number of observations.
3. Enter $p_0$ and $H_1$, then press Calculate.

The results should be displayed.

**Example 2**

Consider the excerpt shown below (also used in Example 1, in Section 9.3) from a poll conducted by Pew Research:

*Stem cell, marijuana proposals lead in Mich. poll*

A recent poll shows voter support leading opposition for ballot proposals to loosen Michigan's restrictions on embryonic stem cell research and allow medical use of marijuana. The EPIC-MRA poll conducted for The Detroit News and television stations WXYZ, WILX, WOOD and WJRT found 50 percent of
likely Michigan voters support the stem cell proposal, 32 percent against and 18 percent undecided. The telephone poll of 602 likely Michigan voters was conducted Sept. 22 through Wednesday. It has a margin of sampling error of plus or minus 4 percentage points. (Source: Associated Press)

Suppose we wonder if the percent of Elgin Community College students who support stem cell research is different from this. If 61 of 100 randomly selected ECC students support stem cell research, is there enough evidence at the 5% level of significance to support our claim?

Note: Be sure to check that the conditions for performing the hypothesis test are met.

One question you might have is, "What do we do if the conditions for the hypothesis test about \( p \) aren't met?"

Great question!

In that case, we can no longer say that sample proportion, \( \hat{p} \), is approximately normally distributed. What we do instead is return to the binomial distribution, and just consider \( x \), the number of successes. Let's do a quick review of binomial probabilities.

### A Binomial Refresher

#### The Binomial Probability Distribution Function

The probability of obtaining \( x \) successes in \( n \) independent trials of a binomial experiment, where the probability of success is \( p \), is given by

\[
P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}
\]

Where \( x = 0, 1, 2, \ldots , n \)

#### Using Technology to Calculate Binomial Probabilities

Here's a quick overview of the formulas for finding binomial probabilities in StatCrunch.

Click on **Stat > Calculators > Binomial**

Enter \( n \), \( p \), the appropriate equality/inequality, and \( x \). The figure below shows \( P(X \geq 3) \) if \( n=4 \) and \( p=0.25 \).

<table>
<thead>
<tr>
<th>( n ): 4</th>
<th>( p ): 0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(X \geq 3) = \binom{4}{3} 0.25^3 (1 - 0.25)^{4-3} ) = 0.05078125</td>
<td></td>
</tr>
</tbody>
</table>

#### Hypothesis Testing Using the Binomial Distribution

**Example 3**

Traditionally, about 70% of students in a particular Statistics course at ECC are successful. If only 15 students in a class of 28 randomly selected
students are successful, is there enough evidence at the 5% level of significance to say that students of that particular instructor are successful at a rate less than 70%?

[ reveal answer ]
Section 10.3: Hypothesis Tests for a Population Mean

In Section 10.2, we tested hypotheses regarding a population proportion. In this section, we’ll consider claims regarding $\mu$, the population mean.

As we did in the previous section, we have some conditions that need to be true in order to perform the test (based on the Central Limit Theorem from Chapter 8).

1. the sample is obtained using simple random sampling, and
2. the sample has no outliers and the population from which the sample is drawn is normally distributed, or the sample size is large ($n \geq 30$)

The steps in performing the hypothesis test are nearly identical, with only a couple changes.

Performing a Hypothesis Test Regarding $\mu$

**Step 1**: State the null and alternative hypotheses.

- **Two-Tailed**: $H_0: \mu = \mu_0$ \hspace{1cm} $H_1: \mu \neq \mu_0$
- **Left-Tailed**: $H_0: \mu = \mu_0$ \hspace{1cm} $H_1: \mu < \mu_0$
- **Right-Tailed**: $H_0: \mu = \mu_0$ \hspace{1cm} $H_1: \mu > \mu_0$

**Step 2**: Decide on a level of significance, $\alpha$, depending on the seriousness of making a Type I error. ($\alpha$ will often be given as part of a test or homework question, but this will not be the case in the outside world.)

**Step 3**: Compute the test statistic, $t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$.

**Step 4**: Determine the $P$-value (see below).

**Step 5**: Reject the null hypothesis if the $P$-value is less than the level of significance, $\alpha$.

**Step 6**: State the conclusion. (You should also include a measure of the strength of the results, based on the $P$-value.)

Finding $P$-values using the $t$-distribution

Because the $t$-table does not give probabilities, we will need to use technology to find probabilities in the $t$-distribution.
Probabilities in the t-Distribution Using StatCrunch

We can also use either of the software packages we’ve been using so far to find probabilities in the t-distribution. Just click on the name of the software you plan to use.

Click on Stat > Calculators > T

Enter the degree of freedom, the direction of the inequality, and x. Then press Compute.

Example 1

Problem: In Example 1, in Section 10.2, we considered a survey about student work habits. The students who performed the survey found that ECC transfer students work, on average, about 17 hours per week.

In that example, we theorized that Mth120 online students chose the online format because they tend to work full-time, or at least more than the average student. We also assumed that the standard deviation for the hours worked per week was 5 hours.

Suppose instead that we don’t have any prior information, and we simply use the sample data. If we collect a random sample of 30 Mth120 online students, and find a sample mean of 19.3 hours per week with a standard deviation of 6.2 hours, is there enough evidence at the 5% level of significance to support our claim that Mth120 online students tend to work more than the average ECC transfer student? (Note: Assume that the sample data are approximately normally distributed with no outliers.)

Solution:

Step 1:

\[ H_0: \mu = 17 \]
\[ H_1: \mu > 17 \]

Step 2: \( \alpha = 0.05 \) (given)

Step 3:

\[ t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{19.3 - 17}{6.2/\sqrt{30}} \approx 2.03. \]

Step 4: \( P\)-value = \( P(t > 2.03, df=29) \approx 0.0258. \)

Step 5: Since the \( P\)-value < \( \alpha \), we reject the null hypothesis.

Step 6: Yes, there is enough evidence at the 5% level of significance to support our claim that the average hours worked for Mth120 online students is more than 17 hours per week.

Hypothesis Testing Using StatCrunch

1. Enter the data.
2. Go to Stat > t-Statistics > One Sample
3. Select with data if you have the data, or with summary if you only have the summary statistics.
4. If you chose with data, click on the variable that you want for the confidence interval and enter the population standard deviation. Otherwise, enter the sample statistics.
5. Click on Next.
6. Enter $\mu_0$ and the alternative hypothesis.
7. Click on Calculate.

The results should be displayed.

Example 2

In Example 2 in Section 10.2, we looked the resting heart rates of 25 Statistics students.

<table>
<thead>
<tr>
<th>heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>61 63 64 65 65</td>
</tr>
<tr>
<td>67 71 72 73 74</td>
</tr>
<tr>
<td>75 77 79 80 81</td>
</tr>
<tr>
<td>82 83 83 84 85</td>
</tr>
<tr>
<td>86 86 89 95 95</td>
</tr>
</tbody>
</table>

(Click here to view the data in a format more easily copied.)

In that problem, we assumed that the standard deviation of all resting heart rates was 12 beats per minute. Let's suppose we don't have information about the standard deviation.

According to the National Library of Medicine, the average resting heart rate for Americans is about 72 bpm. Is there evidence at the 5% level of significance to support a claim that this particular class of Statistics students has a heart rate different from the national average?

Be sure to check that the conditions for performing hypothesis tests are met.

[ reveal answer ]

Statistical Significance vs. Practical Significance

There's one last point we need to discuss in this section, and it's an important one. We need to talk about the difference between a result being statistically significant versus a result that is practically significant.

The idea is this - suppose we're comparing the performance of men and women in mathematics. To compare, we use a particular exam, given randomly to a certain number of students from across the country. We assume the difference between the two groups is zero, but we wonder whether it might be different. In that case, our null and alternative hypotheses might look something like this:

$$H_0: \mu = 0$$
$$H_1: \mu \neq 0$$

Let's suppose that women are only slightly better than men on this exam - say a difference of 0.5 (if we assume the scale is 0-100). If we have a sample of 30 students - 15 men and 15 women - it'll be very hard to get a sample mean very far from zero, and so it'd be pretty rare that we'd actually find enough evidence to reject the null hypothesis (meaning we'd be making Type II error - not rejecting $H_0$ when we should).

On the other hand, if we had 10,000 students, it'd be very likely to get sample mean very close to the real mean, and a sample size of 10,000 would give us an extremely small $P$-value - definitely small enough to reject $H_0$.

Here's the thing - in that second case, we'd reject the null hypothesis and say that the evidence supports the claim that men and women are different, but the difference in practical terms is insignificant. So statistical
significance (enough evidence to reject the null hypothesis) doesn't necessarily imply practical significance.

Beware of studies with very large sample sizes that claim statistical significance. It may be that the differences have no practical meaning.

As an illustration, check out this comic from XKCD:
This work is licensed under a Creative Commons License.
Section 10.4: Hypothesis Tests for a Population Standard Deviation

10.1 The Language of Hypothesis Testing
10.2 Hypothesis Tests for a Population Proportion
10.3 Hypothesis Tests for a Population Mean
10.4 Hypothesis Tests for a Population Standard Deviation
10.5 Putting It Together: Which Method Do I Use?

Objectives

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

1. test hypotheses about a population standard deviation

Before we begin this section, we need a quick refresher of the $\chi^2$ distribution.

The Chi-Square ($\chi^2$) distribution

Reminder: "chi-square" is pronounced "kai" as in sky, not "chai" like the tea.

The Chi-Square ($\chi^2$) distribution

If a simple random sample size $n$ is obtained from a normally distributed population with mean $\mu$ and standard deviation $\sigma$, then

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

has a chi-square distribution with $n-1$ degrees of freedom.

Properties of the $X^2$ distribution

1. It is not symmetric.
2. The shape depends on the degrees of freedom.
3. As the number of degrees of freedom increases, the distribution becomes more symmetric.
4. $\chi^2 \geq 0$
Example 1

In Example 2, in Section 10.2, we assumed that the standard deviation for the resting heart rates of ECC students was 12 bpm. Later, in Example 2 in Section 10.3, we considered the actual sample data below.

<table>
<thead>
<tr>
<th>heart rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>61 63 64 65 65</td>
</tr>
<tr>
<td>67 71 72 73 74</td>
</tr>
<tr>
<td>75 77 79 80 81</td>
</tr>
<tr>
<td>82 83 83 84 85</td>
</tr>
<tr>
<td>86 86 89 95 95</td>
</tr>
</tbody>
</table>

(Click here to view the data in a format more easily copied.)

Based on this sample, is there enough evidence to say that the standard deviation of the resting heart rates for students in this class is different from 12 bpm?

Note: Be sure to check that the conditions for performing the hypothesis test are met.
1. Go to **Stat > Variance > One sample**
2. Select **with data** if you have the data, or **with summary** if you only have the summary statistics.
3. If you chose **with data**, click on the variable that you want for the hypothesis test and press **Next**. Otherwise, enter the sample statistics and press **Next**.
4. Enter the population **variance** (not standard deviation!) and the H₁.
5. Press **Calculate**.

The results should be displayed.

---

**Example 2**

Let's look at Example 1 again, and try the hypothesis test with technology.

[ reveal answer ]
Section 10.5: Putting It Together: Which Method Do I Use?

10.1 The Language of Hypothesis Testing
10.2 Hypothesis Tests for a Population Proportion
10.3 Hypothesis Tests for a Population Mean
10.4 Hypothesis Tests for a Population Standard Deviation
10.5 Putting It Together: Which Method Do I Use?

**Objectives**

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

1. determine the appropriate hypothesis test to perform

---

**Hypothesis Test Summary**

So far this semester, we've learned three different hypothesis tests, based on the parameter of interest, and what information is given. Those three are:

---

**Tests Regarding the Population Proportion**

In order to perform a hypothesis test regarding the population proportion, all of the following must be true:

1. the sample is a simple random sample, and
2. the sample is less than 5% of the population (n≤0.05N), and
3. np(1-p)≥10

The sample statistic for this test is:

\[
z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}
\]

---

**Tests Regarding the Population Mean**

In order to perform a hypothesis test regarding the population mean, the sample must be a simple random sample, and one of the following must be true:

1. the sample comes from a normally distributed population, or
2. the sample size is more than 30 and there sample contains no outliers.

**Two-Tailed Left-Tailed Right-Tailed**

H\(_0\): µ = µ\(_0\)  
H\(_0\): µ = µ\(_0\)  
H\(_0\): µ = µ\(_0\)  
H\(_1\): µ ≠ µ\(_0\)  
H\(_1\): µ < µ\(_0\)  
H\(_1\): µ > µ\(_0\)

The sample statistic for this test is:

\[
t = \frac{\bar{x} - \mu}{s/\sqrt{n}}
\]

with n-1 degrees of freedom.

---

**Tests Regarding the Population Standard Deviation**
In order to perform a hypothesis test regarding the population standard deviation, the sample \textit{must} come from a normally distributed population. In this case, the sample statistic is:

\[
\chi^2_0 = \frac{(n - 1)s^2}{\sigma^2}
\]

\section*{Choosing the Appropriate Hypothesis Test}

So now the big question is which test to apply. Here's a simplified version of the flowchart in your textbook:

Unfortunately, like most problems where you need to choose which technique to apply, there's no handy blueprint that you can always follow. The key, then, is to try as many examples as possible (like those on the next page) and doing all the assigned homework as soon as possible.

For each example, state null and alternative hypotheses and choose the appropriate test statistic.

\section*{Example 1}

According to the \textbf{US Census Bureau}, approximately 42\% of Americans ages 18-24 voted in the 2004 presidential election. We wonder if the percentage of ECC students is different from this. If we collect a simple random sample of 200 ECC students and find that 103 of them voted in the 2008 presidential election, is there enough evidence at the 10\% level of significance to support our claim?

[ reveal answer ]

\section*{Example 2}

According to the ECC website, the average age of ECC students is 28.2 years. Suppose we collect data from 30 Mth120 online students and find an average age of 20.3 years, with a standard deviation of 2.3 years.

Is there enough evidence at the 5\% level of significance to support the claim that the average age of Mth120 online students is less than 28.2 years?

[ reveal answer ]
We know from previous examples that the standard deviation of IQs is normally distributed with a standard deviation of 15. Suppose we wonder if the IQs of ECC students have more variation. To answer this question, we collect the IQs from a random sample of ECC students and find a standard deviation of 16.2.

Based on this information, is there enough evidence at the 5% level of significance to say that the IQs of ECC students have more variation than the general population?

[ reveal answer ]
Chapter 11: Inference on Two Samples

11.1 Inference about Two Proportions
11.2 Inference about Two Means: Dependent Samples
11.3 Inference about Two Means: Independent Samples
11.4 Inference about Two Standard Deviations
11.5 Putting It Together: Which Method Do I Use?

In Chapters 9 and 10, we studied inferential statistics (confidence intervals and hypothesis tests) regarding population parameters of a single population - the average rest heart rate of students in a class, the proportion of ECC who voted, etc.

In Chapter 11, we'll be considering the relationship between two populations - means, proportions and standard deviations. One frequent comparison we want to make between two populations is concerning the proportion of individuals with certain characteristics. For example, suppose we want to determine if college faculty voted at a higher rate than ECC students in the 2008 presidential election. Since we don't have any information from either population, we would need to take samples from each. This isn't an example of a hypothesis test from Section 10.4, about one proportion, it'd be comparing two proportions.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::

1 2 3 4 5 6 7 8 9 10

This work is licensed under a Creative Commons License.
Inferential Statistics Regarding Two Proportions

Objectives

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

1. test hypotheses regarding two population proportions
2. construct and interpret confidence intervals for the difference between two population proportions
3. determine the sample size necessary for estimating the difference between two population proportions within a specific margin of error

In Chapters 9 and 10, we studied inferential statistics (confidence intervals and hypothesis tests) regarding population parameters of a single population - the average rest heart rate of students in a class, the proportion of ECC who voted, etc.

In Chapter 11, we'll be considering the relationship between two populations - means, proportions and standard deviations.

A frequent comparison we want to make between to populations is concerning the proportion of individuals with certain characteristics. For example, suppose we want to determine if college faculty voted at a higher rate than ECC students in the 2008 presidential election. Since we don't have any information from either population, we would need to take samples from each. This isn't an example of a hypothesis test from Section 10.4, about one proportion, it'd be comparing two proportions, so we need some new background.

The information that follows is a bit heavy, but it shows the theoretical background for testing claims and finding confidence intervals for the difference between two population proportions.

The Difference Between Two Population Proportions

In Section 8.2, we discussed the distribution of one sample proportion, \( \hat{p} \). What we'll need to do now is develop some similar theory regarding the distribution of the difference in two sample proportions, \( \hat{p}_1 - \hat{p}_2 \).

The Sampling Distribution of the Difference between Two Proportions

Suppose simple random samples size \( n_1 \) and \( n_2 \) are taken from two populations. The distribution of \( \hat{p}_1 - \hat{p}_2 \) where \( \hat{p}_1 = \frac{x_1}{n_1} \) and \( \hat{p}_2 = \frac{x_2}{n_2} \), is approximately normal with mean \( \mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2 \) and standard deviation

\[
\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}, \text{ provided:}
\]

1. \( n_1 \hat{p}_1(1-\hat{p}_1) \geq 10 \)
2. \( n_2 \hat{p}_2(1-\hat{p}_2) \geq 10 \)
3. both sample sizes are less than 5% of their respective populations.

The standardized version is then
which has an approximate standard normal distribution.

The thing is, in most of our hypothesis testing, the null hypothesis assumes that the proportions are the same ($p_1 = p_2$), so we can call $p = p_1 = p_2$.

Since $p_1 = p_2$, we can substitute 0 for $p_1 - p_2$, and substitute $p$ for both $p_1$ and $p_2$. In that case, we can rewrite the above standardized $z$ the following way:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n_1} + \frac{\hat{p}(1 - \hat{p})}{n_2}}}$$

Which leads us to our hypothesis test for the difference between two proportions.

**Performing a Hypothesis Test Regarding $p_1 - p_2$**

**Step 1:** State the null and alternative hypotheses.

<table>
<thead>
<tr>
<th>Two-Tailed</th>
<th>Left-Tailed</th>
<th>Right-Tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: $p_1 - p_2 = 0$</td>
<td>$H_0$: $p_1 - p_2 = 0$</td>
<td>$H_0$: $p_1 - p_2 = 0$</td>
</tr>
<tr>
<td>$H_1$: $p_1 - p_2 \neq 0$</td>
<td>$H_1$: $p_1 - p_2 &lt; 0$</td>
<td>$H_1$: $p_1 - p_2 &gt; 0$</td>
</tr>
</tbody>
</table>

**Step 2:** Decide on a level of significance, $\alpha$.

**Step 3:** Compute the test statistic, $z = \frac{\hat{p}_1 - \hat{p}_2}{\hat{p}(1 - \hat{p})\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$.

**Step 4:** Determine the $P$-value.

**Step 5:** Reject the null hypothesis if the $P$-value is less than the level of significance, $\alpha$.

**Step 6:** State the conclusion.

**A note about the difference between two proportions:** As with the previous two sections, the order in which the proportions are placed is not important. The important thing is to note clearly in your work what the order is, and then to construct your alternative hypothesis accordingly.

**Hypothesis Testing Regarding $p_1 - p_2$ Using StatCrunch**

1. Go to **Stat > Proportions > Two Sample > with summary**
2. Enter the number of successes and observations for each sample and press **Next**.
3. Set the null proportion difference and the alternative hypothesis.
4. Click on **Calculate**.

The results should appear.
Example 1

Problem: Suppose a researcher believes that college faculty vote at a higher rate than college students. She collects data from 200 college faculty and 200 college students using simple random sampling. If 167 of the faculty and 138 of the students voted in the 2008 Presidential election, is there enough evidence at the 5% level of significance to support the researcher’s claim?

Solution:
First, we need to check the conditions. Both sample sizes are clearly less than 5% of their respective populations. In addition,
\[ n_1 \hat{p}_1 (1 - \hat{p}_1) = 200 \left( \frac{167}{200} \right) \left( 1 - \frac{167}{200} \right) \approx 27.6 \geq 10 \]
\[ n_2 \hat{p}_2 (1 - \hat{p}_2) = 200 \left( \frac{138}{200} \right) \left( 1 - \frac{138}{200} \right) \approx 42.8 \geq 10 \]
So our conditions are satisfied.

Step 1:
Let’s take the two portions in the order we receive them, so \( p_1 = p_f \) (faculty) and \( p_2 = p_s \) (students)
Our hypotheses are then:
\( H_0: p_f - p_s = 0 \)
\( H_1: p_f - p_s > 0 \) (since the researcher claims that faculty vote at a higher rate)
Step 2: \( \alpha = 0.05 \) (given)
Step 3: (we’ll use StatCrunch)
Step 4: Using StatCrunch:

Hypothesis test results:
- \( p_1 \): proportion of successes for population 1
- \( p_2 \): proportion of successes for population 2
- \( p_1 - p_2 \): difference in proportions
- \( H_0: p_1 - p_2 = 0 \)
- \( H_A: p_1 - p_2 > 0 \)

<table>
<thead>
<tr>
<th>Difference</th>
<th>Count1</th>
<th>Total1</th>
<th>Count2</th>
<th>Total2</th>
<th>Samp</th>
<th>Z-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_1 - p_2 )</td>
<td>167</td>
<td>200</td>
<td>138</td>
<td>200</td>
<td>3.4073462</td>
<td>0.0003</td>
<td></td>
</tr>
</tbody>
</table>

(Trimmed to fit on this page.)

Step 5: Since the \( P \)-value < \( \alpha \), we reject the null hypothesis.

Step 6: Based on these results, there is very strong evidence (certainly enough at the 5% level of significance) to support the researcher’s claim.

Confidence Intervals about the Difference Between Two Proportions

We can also find a confidence interval for the difference in two population proportions.

In general, a \((1-\alpha)100\%\) confidence interval for \( p_1 - p_2 \) is...
Example 2

**Problem**: Considering the data from Example 1, find a 99% confidence interval for the difference between the proportion of faculty and the proportion of students who voted in the 2008 Presidential election.

**Solution**: From Example 1, we know that the conditions for performing inference are met, so we'll use StatCrunch to find the confidence interval.

99% confidence interval results:
- $\hat{p}_1$: proportion of successes for population 1
- $\hat{p}_2$: proportion of successes for population 2
- $p_1 - p_2$: difference in proportions

<table>
<thead>
<tr>
<th>Difference</th>
<th>Count1</th>
<th>Total1</th>
<th>Count2</th>
<th>Total2</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1 - p_2$</td>
<td>168</td>
<td>200</td>
<td>137</td>
<td>20</td>
<td>0.047218394</td>
<td>0.26278162</td>
</tr>
</tbody>
</table>

(Timmed to fit on this page.)

So we can say that we're 99% confident that the difference between the proportion of faculty who vote and the proportion of students who vote is between 3.7% and 25.3%.

### Determining the Sample Size Needed

In Section 9.3, we learned how to find the necessary sample size if a specific margin of error is desired. We can do a similar analysis for the difference in two proportions. From the confidence interval formula, we know that the margin of error is:

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

If we assume that $n_1 = n_2 = n$, we can solve for $n$ and get the following result:
The **sample size required** to obtain a \((1-\alpha)100\%\) confidence interval for \(p_1-p_2\) with a margin of error \(E\) is:

\[
n = n_1 = n_2 = \left[ \frac{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)}{\left(\frac{z_{\alpha/2}}{E}\right)^2} \right]
\]

rounded up to the next integer, if \(\hat{p}_1\) and \(\hat{p}_2\) are estimates for \(p_1\) and \(p_2\), respectively.

If no prior estimate is available, use \(\hat{p}_1 = \hat{p}_2 = 0.5\), which yields the following formula:

\[
n = n_1 = n_2 = 0.5 \left(\frac{z_{\alpha/2}}{E}\right)^2
\]

again rounded up to the next integer.

Note: As in Section 9.3, the desired margin of error should be expressed as a decimal.

Let's try one.

**Example 3**

Suppose we want to study the success rates for students in Mth098 Intermediate Algebra at ECC. We want to compare the success rates of students who place directly into Mth098 with those who first took Mth096 Beginning Algebra. From past experience, we know that a typical success rate for students in this class is about 65%. How large of a sample size is necessary to create a 95% confidence interval for the difference of the two passing rates with a maximum error of 2%?

[ reveal answer ]
Section 11.2: Inference about Two Means: Dependent Samples

11.1 Inference about Two Proportions

11.2 Inference about Two Means: Dependent Samples

11.3 Inference about Two Means: Independent Samples

11.4 Inference about Two Standard Deviations

11.5 Putting It Together: Which Method Do I Use?

Objectives

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

1. distinguish between independent and dependent samples
2. test hypotheses regarding matched-pairs data
3. construct and interpret confidence intervals about the population mean difference of matched-pairs data

In the next two sections, we’ll focus on the relationship between two means. Before we begin, we need to discuss the two possible situations.

Independent vs. Dependent Samples

In general, there are two possible situations regarding two population means. Consider the following examples:

Example 1

**Problem:** Suppose we measure the thickness of plaque (mm) in the carotid artery of 10 randomly selected patients with mild atherosclerotic disease. Two measurements are taken, thickness before treatment with Vitamin E (baseline) and after two years of taking Vitamin E daily. (Source: UCLA Dept. of Statistics)

**Discussion:** In this example, we would be comparing the mean plaque thickness before Vitamin E with the mean thickness after - so the same 10 patients would be in the sample "before" and the sample "after". When the individuals selected are paired in this manner, we call the samples dependent.

Example 2

**Problem:** Nine observations of surface soil pH were made at two different locations. Does the data suggest that the true mean soil pH values differ for the two locations? (Source: UCLA Dept. of Statistics)

**Discussion:** Unlike in Example 1, the samples in this example are completely unrelated (two different locations). In examples like this, where the individuals selected have no relation to each other, we call the samples independent.

In general, two samples are dependent if the individuals in one sample determine the individuals in the other sample. (i.e. matched-pair design) Two samples are independent when the individuals in one sample do not determine the individuals in the other sample. (i.e. completely randomized)
This will be very important as we progress, because we will need to distinguish between whether the samples are dependent or independent, because the statistical methods will be very different.

### The Difference, \( d \)

When considering dependent samples, we analyze the difference, \( d \), in each matched pair. For example, suppose we consider the thickness of plaque (mm) in the carotid artery, referenced in Example 1. If an individual had a maximal thickness of 0.92mm before the Vitamin E treatment, and 0.95mm after the treatment, the difference, \( d \), for that individual would be

\[
d = 0.95 - 0.92 = 0.03\text{mm}
\]

So in general, for this experiment, we would define the difference, \( d \), to be:

\[
d = \text{thickness after Vit. E treatment} - \text{thickness before treatment}
\]

Before we can perform any inferential statistics, we need to know the distribution of \( d \).

#### The Distribution of the Difference, \( d \)

Suppose the following are true concerning a sample:

1. the sample is obtained using simple random sampling, and
2. the sample data are matched pairs, and
3. the sample has no outliers and the population from which the sample is drawn is normally distributed, or the sample size is large (\( n \geq 30 \)).

Then the test statistic

\[
t_0 = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}
\]

follows the \( t \)-distribution with \( n-1 \) degrees of freedom.

With that in mind, we can now perform statistical inference like hypothesis tests and confidence intervals. We'll start with hypothesis tests, and follow the same steps we did when we were analyzing the population mean, when \( \sigma \) was unknown.

**A note about the difference, \( d \):** Many students find it confusing how to determine which value should go first when setting up the difference. In reality, it doesn't matter! The important thing is to note clearly in your work how \( d \) is set up, and then to construct your alternative hypothesis accordingly.

#### Performing a Hypothesis Test Regarding \( d \)

**Step 1:** State the null and alternative hypotheses.

<table>
<thead>
<tr>
<th>Two-Tailed</th>
<th>Left-Tailed</th>
<th>Right-Tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_0: \mu_d = 0 )</td>
<td>( H_0: \mu_d = 0 )</td>
<td>( H_0: \mu_d = 0 )</td>
</tr>
<tr>
<td>( H_1: \mu_d \neq 0 )</td>
<td>( H_1: \mu_d &lt; 0 )</td>
<td>( H_1: \mu_d &gt; 0 )</td>
</tr>
</tbody>
</table>

**Step 2:** Decide on a level of significance, \( \alpha \).

**Step 3:** Compute the test statistic,

\[
t_0 = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}
\]
Example 3

Problem: Suppose we want to determine if a diet drug is effective. To determine its effectiveness, we randomly select 10 volunteers, and measure their weight before the diet drug treatment, and again one month later. The results are shown below.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>190</td>
<td>211</td>
<td>198</td>
<td>203</td>
<td>262</td>
<td>224</td>
<td>251</td>
<td>238</td>
<td>219</td>
</tr>
<tr>
<td>After</td>
<td>184</td>
<td>204</td>
<td>197</td>
<td>208</td>
<td>246</td>
<td>221</td>
<td>256</td>
<td>225</td>
<td>211</td>
</tr>
</tbody>
</table>

Is there evidence to support the company's claim that the diet drug does cause weight loss at the 5% level of significance?

Solution:

First, we need to calculate the differences. It's important to always write down which direction you want to define the difference, d. In this case, we'll use:

\[ d = \text{Before} - \text{After} \]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>190</td>
<td>211</td>
<td>198</td>
<td>203</td>
<td>262</td>
<td>224</td>
<td>251</td>
<td>238</td>
<td>219</td>
</tr>
<tr>
<td>After</td>
<td>184</td>
<td>204</td>
<td>197</td>
<td>208</td>
<td>246</td>
<td>221</td>
<td>256</td>
<td>225</td>
<td>211</td>
</tr>
<tr>
<td>Difference</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>-5</td>
<td>16</td>
<td>3</td>
<td>-5</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

We need to then determine if the differences are normally distributed, since our sample size is less than 30.
We can see that the Q-Q plot is fairly linear and the boxplot shows no outliers, so it's reasonable to say that the differences are normally distributed.

**Step 1:**

- \( H_0: \mu_d = 0 \)
- \( H_1: \mu_d > 0 \)  
  (Since the company wants to show that the average weight loss is positive.)

**Step 2:** \( \alpha = 0.05 \) (given)

**Step 3:** (we'll use StatCrunch)

**Step 4:** Using StatCrunch:

**Hypothesis test results:**

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Diff.</th>
<th>Std. Err.</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before - After</td>
<td>5.6</td>
<td>2.2715633</td>
<td>9</td>
<td>2.4652627</td>
<td>0.0179</td>
</tr>
</tbody>
</table>

Differences stored in column, Differences.

**Step 5:** Since the \( P \)-value < \( \alpha \), we reject the null hypothesis.

**Step 6:** Based on these results, it would appear that there is enough evidence to support the claim that the drug causes weight loss.

**Note:** If we had first set up the difference as \( d = \text{After} - \text{Before} \), the alternative hypothesis would then be \( H_1: \mu_d < 0 \) (since the company claims the "after" is less than the "before").

---

**Confidence Intervals about the Mean Difference**

Since the distribution of \( \bar{d} \) follows the \( t \)-distribution, we can also create a confidence interval for the population mean difference.

In general, a \((1-\alpha)100\%\) confidence interval for \( \mu_d \) is

\[
\bar{d} \pm t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}}
\]
where $t_{a/2}$ is computed with $n-1$ degrees of freedom.

Note: The sample size must be large ($n \geq 30$) with no outliers or the population must be normally distributed.

### Confidence Intervals About $\mu_d$ Using StatCrunch

1. Enter the data.
2. Go to **Stat > t-Statistics > paired**.
3. Select the columns the 1st and 2nd samples are in and click **Next**.
4. Select "Confidence Interval" and select the confidence level.
5. Click on **Calculate**.

The results should appear.

Note: StatCrunch does not display the variable correctly. It will display $\mu_1 - \mu_2$ rather than $\mu_d$.

### Example 4

**Problem:** Consider the weight loss data from Example 3.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>190</td>
<td>211</td>
<td>198</td>
<td>203</td>
<td>262</td>
<td>224</td>
<td>251</td>
<td>238</td>
<td>219</td>
</tr>
<tr>
<td>After</td>
<td>184</td>
<td>204</td>
<td>197</td>
<td>208</td>
<td>246</td>
<td>221</td>
<td>256</td>
<td>225</td>
<td>211</td>
</tr>
</tbody>
</table>

Find a 90% confidence interval for the population mean difference.

**Solution:**

From Example 3, we know that the differences are normally distributed with no outliers, so we can find the confidence interval.

$$d = \text{Before} - \text{After}$$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>190</td>
<td>211</td>
<td>198</td>
<td>203</td>
<td>262</td>
<td>224</td>
<td>251</td>
<td>238</td>
<td>219</td>
</tr>
<tr>
<td>After</td>
<td>184</td>
<td>204</td>
<td>197</td>
<td>208</td>
<td>246</td>
<td>221</td>
<td>256</td>
<td>225</td>
<td>211</td>
</tr>
<tr>
<td>Difference</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>-5</td>
<td>16</td>
<td>-5</td>
<td>13</td>
<td>8</td>
<td>12</td>
</tr>
</tbody>
</table>

Using StatCrunch:

**90% confidence interval results:**

$\mu_1 - \mu_2$: mean of the paired difference between Before and After

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Diff.</th>
<th>Std. Err.</th>
<th>DF</th>
<th>L. Lim</th>
<th>U. Lim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before - After</td>
<td>5.6</td>
<td>2.2715633</td>
<td>9</td>
<td>1.4359679</td>
<td>9.764032</td>
</tr>
</tbody>
</table>

So we can say that we're 90% confident that the mean weight loss (difference) is between 1.4 and 9.8.
This work is licensed under a Creative Commons License.
Section 11.3: Inference about Two Means: Independent Samples

11.1 Inference about Two Proportions
11.2 Inference about Two Means: Dependent Samples
11.3 Inference about Two Means: Independent Samples
11.4 Inference about Two Standard Deviations
11.5 Putting It Together: Which Method Do I Use?

Objectives

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

1. test hypotheses regarding the difference of two independent means
2. construct and interpret confidence intervals regarding the difference of two independent means

In 2005, Larry Summers, then President of Harvard, gave a speech at the NBER Conference on Diversifying the Science and Engineering Workforce. In that speech, he made some very controversial remarks regarding differences in the genders. In particular,

> It does appear that on many, many different human attributes—height, weight, propensity for criminality, overall IQ, mathematical ability, scientific ability—there is relatively clear evidence that whatever the difference in means—which can be debated—there is a difference in the standard deviation, and variability of a male and a female population.

Suppose we wanted to do a comparison between the genders. In Section 11.2, we looked at comparing two means with matched-pairs data - dependent samples. What if there isn't a relationship between the two samples? We certainly can't pair them up then, and find the mean difference. What we need to analyze instead is the difference of two means. First, we need to know something about it's distribution.

The Difference of Two Independent Means

**The Distribution of the Difference of Two Means, \( d \)**

Suppose a simple random sample of size \( n_1 \) is taken from a population with unknown mean \( \mu_1 \) and unknown standard deviation \( \sigma_1 \). In addition, a simple random sample of size \( n_2 \) is taken from a second population with unknown mean \( \mu_2 \) and unknown standard deviation \( \sigma_2 \). If the two populations are normally distributed or the sample sizes are sufficiently large (\( n_1, n_2 \geq 30 \)), then

\[
 t_0 = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} 
\]

*approximately* follows the \( t \)-distribution with the smaller of \( n_1-1 \) or \( n_2-1 \) degrees of freedom.

Note: There is no exact method for comparing two means with unequal populations, but this statistic is a close approximation. It is known as Welch's approximate \( t \), in honor of English statistician Bernard Lewis Welch (1911-1989).

Now that we have the distribution of the difference between two means, we can perform statistic inference (hypothesis testing and confidence intervals).
Performing a Hypothesis Test Regarding the Difference Between Two Independent Means

**Step 1:** State the null and alternative hypotheses.

<table>
<thead>
<tr>
<th></th>
<th>Two-Tailed</th>
<th>Left-Tailed</th>
<th>Right-Tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0:$</td>
<td>$\mu_1 - \mu_2 = 0$</td>
<td>$H_0:$</td>
<td>$H_0:$</td>
</tr>
<tr>
<td>$H_1:$</td>
<td>$\mu_1 - \mu_2 \neq 0$</td>
<td>$H_1:$</td>
<td>$H_1:$</td>
</tr>
<tr>
<td>$H_1:$</td>
<td>$\mu_1 - \mu_2 &gt; 0$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 2:** Decide on a level of significance, $\alpha$.

**Step 3:** Compute the test statistic, $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$.

**Step 4:** Determine the $P$-value.

**Step 5:** Reject the null hypothesis if the $P$-value is less than the level of significance, $\alpha$.

**Step 6:** State the conclusion.

---

**Hypothesis Testing Regarding $\mu_1 - \mu_2$ Using StatCrunch**

1. Enter the data. (Note: If you're copying from another file, be careful - put the column with the most entries first. StatCrunch does not handle spaces well.)
2. Go to Stat > t-Statistics > Two Sample, then with data or with summary.
3. If you chose with data, select the columns containing the 1st and 2nd samples. Otherwise, enter all the sample statistics.
5. Set the null mean difference and the alternative hypothesis.
6. Click on Calculate.

The results should appear.

A note about the difference between the means: As with the previous section, it's often difficult for students to choose which mean to place first. Again - it doesn't matter! The important thing is to note clearly in your work what the order is, and then to construct your alternative hypothesis accordingly.

---

**Example 1**

**Problem:** Suppose we wish to test whether there is a difference in the performances of men and women in mathematics. An ECC instructor collects exam scores from 2 semesters worth of Beginning Algebra students, shown below by gender.

<table>
<thead>
<tr>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>75 90 92 88 76 61 77</td>
<td>79 97 56</td>
</tr>
<tr>
<td>43 77 63 76 67 99 93</td>
<td>43 82 87 32 71 77 77</td>
</tr>
<tr>
<td>82 55 75 47 80 92 56</td>
<td>60 64 82 91 51 60 76</td>
</tr>
<tr>
<td>94 90 96 73 71 75 52</td>
<td>77 57 79 65 68 65 55</td>
</tr>
<tr>
<td>86 93 86 50 30 46 36</td>
<td>72 66 29 69 85 82 43</td>
</tr>
</tbody>
</table>
Is there enough evidence at the 5% level of significance to support the claim that men and women perform differently in this class?

**Solution:**

First, we need to make sure that neither sample contains outliers. (We do have sample sizes of at least 30, so we don't need to check to see if they come from normally distributed populations.)

![Box plots of Exam Scores by Gender](image)

We can see that neither sample contains outliers, so we are free to continue.

**Step 1:**

In this case, since we're only testing whether the means are different, the order doesn't matter at all. It's easiest to simply take the two in the order we receive them, so $\mu_1 = \mu_W$ (women), and $\mu_2 = \mu_M$ (men).

$H_0$: $\mu_W - \mu_M = 0$

$H_1$: $\mu_W - \mu_M \neq 0$

**Step 2:** $\alpha = 0.05$ (given)

**Step 3:** (we'll use StatCrunch)

**Step 4:** Using StatCrunch:

<table>
<thead>
<tr>
<th>Hypothesis test results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$ : mean of Women</td>
</tr>
<tr>
<td>$\mu_2$ : mean of Men</td>
</tr>
<tr>
<td>$\mu_1 - \mu_2$ : mean difference</td>
</tr>
<tr>
<td>$H_0$: $\mu_1 - \mu_2 = 0$</td>
</tr>
<tr>
<td>$H_A$: $\mu_1 - \mu_2 \neq 0$</td>
</tr>
</tbody>
</table>

(without pooled variances)

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Mean</th>
<th>Std. Err.</th>
<th>DF</th>
<th>T-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1 - \mu_2$</td>
<td>4.1426797</td>
<td>4.0103817</td>
<td>70.54798</td>
<td>1.0329889</td>
<td>0.3051</td>
</tr>
</tbody>
</table>

**Step 5:** Since the $P$-value > $\alpha$, we do not reject the null hypothesis.

**Step 6:** Based on these results, it would appear that there is not enough
evidence (very little, in fact) to support the claim that men and women perform differently in Beginning Algebra.

It should be noted that larger sample sizes will most likely show a statistically significant difference, though the difference may not have any practical meaning.

One final note, you may wonder what the "pooled variances" is referring to, and why we don't use it. By not pooling the variances, we are assuming that the population variances are unequal. There is a test for equal variances (we'll cover it in Section 11.4), but like the earlier tests concerning standard deviations, the distributions must be normal. Your textbook makes this final comment:

Because the formal $F$-test for testing the equality of variances is so volatile, we are content to use Welch's $t$. This test is more conservative than the pooled $t$. The price that must be paid for the conservative approach is that the probability of a Type II error is higher in Welch's $t$ than in the pooled $t$ when the population variances are equal. However, the two tests typically provide the same conclusion, even if the assumption of equal population standard deviations seems reasonable.

Source: Statistics; Informed Decisions Using Data

Confidence Intervals about the Difference Between Two Means

Since the distribution of $\bar{x}_1 - \bar{x}_2$ follows the $t$-distribution, we can also create a confidence interval for the difference between two population means.

In general, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\alpha/2}$ is computed with $\min\{n_1-1, n_2-1\}$ degrees of freedom.

Note: The sample sizes must be large ($n_1, n_2 \geq 30$) with no outliers or the populations must be normally distributed.

Confidence Intervals About $\mu_1 - \mu_2$ Using StatCrunch

1. Enter the data. (Note: If you're copying from another file, be careful - put the column with the most entries first. StatCrunch does not handle spaces well.)
2. Go to Stat $\rightarrow$ t-Statistics $\rightarrow$ Two Sample, then with data or with summary.
3. If you chose with data, select the columns containing the 1st and 2nd samples. Otherwise, enter all the sample statistics.
5. Select "Confidence Interval" and select the confidence level.
6. Click on Calculate.

The results should appear.

Example 2

Problem: Consider the data comparing men and women from Example 1.
Find a 90% confidence interval for the population mean difference.

**Solution:** From Example 1, we know that neither sample contains outliers, so we can find the confidence interval.

Using StatCrunch:

90% confidence interval results:

- $\mu_1$: mean of Women
- $\mu_2$: mean of Men
- $\mu_1 - \mu_2$: mean difference (without pooled variances)

<table>
<thead>
<tr>
<th>Difference</th>
<th>Sample Mean</th>
<th>Std. Err.</th>
<th>DF</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1 - \mu_2$</td>
<td>4.1426797</td>
<td>4.0103817</td>
<td>70.54798</td>
<td>-2.541587</td>
<td>10.826947</td>
</tr>
</tbody>
</table>

So we can say that we're 90% confident that the difference between the two means is between -2.5 and 10.8.
Section 11.4: Inference about Two Population Standard Deviations

11.1 Inference about Two Proportions
11.2 Inference about Two Means: Dependent Samples
11.3 Inference about Two Means: Independent Samples
11.4 Inference about Two Standard Deviations
11.5 Putting It Together: Which Method Do I Use?

Objectives

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

1. find critical values of the $F$-distribution
2. test hypotheses regarding two population standard deviations

The last parameters we need to compare between two populations are the variance and standard deviation. Before we can develop a hypothesis test comparing two population parameters, we need another distribution.

Fisher's $F$-distribution

Unlike the mean, the standard deviation is extremely susceptible to extreme values, and consequently does a very poor job of measuring spread for distributions that are not symmetric. So before we do any inference regarding population standard deviations, we must first verify that the samples come from normally distributed populations.

Fisher's $F$-distribution

If $\sigma_1^2$ = $\sigma_2^2$ and $s_1^2$ and $s_2^2$ are sample variances from independent simple random samples of size $n_1$ and $n_2$, respectively, drawn from normal populations, then

$$F = \frac{s_1^2}{s_2^2}$$

follows the $F$-distribution with $n_1$-1 degrees of freedom in the numerator and $n_2$-1 degrees of freedom in the denominator.

Properties of the $F$-distribution

1. Like the $\chi^2$ distribution, it is not symmetric. It is skewed right
2. The shape depends on the degrees of freedom in the numerator and denominator.
3. $F \geq 0$

Notice here that the samples must come from normally distributed populations.

Finding Critical Values

Find critical values in the $F$-distribution using a table is done in a similar manner to the $t$ and $\chi^2$ tables, though with some differences. The values in the table still represent values with the indicated $\alpha$ area to the right, but because the $F$ distribution has two degrees of freedom rather than one, it requires a separate table for each $\alpha$.

Before we start the section, you need a copy of the table. You can download a printable copy of this table, or use
Example 1

Find the value of the F-distribution that has $\alpha=0.05$ area to the right, with 10 degrees of freedom in the numerator, and 15 degrees of freedom in the denominator.

[ reveal answer ]

Example 2

Find the value of the F-distribution that has $\alpha=0.05$ area to the left, with 20 degrees of freedom in the numerator, and 8 degrees of freedom in the denominator.

[ reveal answer ]

So the values in the table above are the critical values:

$$F_{\alpha, n_1-1, n_2-1}$$

You may wonder how we find critical values for left-tailed tests. To do that, we use the same table and the following formula:

$$F_{1-\alpha, n_1-1, n_2-1} = \frac{1}{F_{\alpha, n_2-1, n_1-1}}$$

Let's try a couple examples.
Finding Critical Values Using StatCrunch

Click on **Stat > Calculators > F**

Enter the numerator and denominator degrees of freedom, the direction of the inequality, and the probability (leave X blank). Then press **Compute**.

**Example 3**

Use the technology of your choice to find the value from the $F$-distribution with $\alpha=0.01$ area to the right if samples of size 15 and 20 are taken.

[ reveal answer ]

Performing a Hypothesis Test Regarding Two Population Standard Deviations

**Step 1**: State the null and alternative hypotheses.

<table>
<thead>
<tr>
<th>Two-Tailed</th>
<th>Left-Tailed</th>
<th>Right-Tailed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$: $\sigma_1^2 = \sigma_2^2$</td>
<td>$H_0$: $\sigma_1^2 = \sigma_2^2$</td>
<td>$H_0$: $\sigma_1^2 = \sigma_2^2$</td>
</tr>
<tr>
<td>$H_1$: $\sigma_1^2 \neq \sigma_2^2$</td>
<td>$H_1$: $\sigma_1^2 &lt; \sigma_2^2$</td>
<td>$H_1$: $\sigma_1^2 &gt; \sigma_2^2$</td>
</tr>
</tbody>
</table>

**Step 2**: Decide on a level of significance, $\alpha$.

**Step 3**: Compute the test statistic, $F = \frac{s_1^2}{s_2^2}$.

**Step 4**: Determine the $P$-value.

**Step 5**: Reject the null hypothesis if the $P$-value is less than the level of significance, $\alpha$.

**Step 6**: State the conclusion.

Hypothesis Testing Regarding Two Population Standard Deviations Using StatCrunch

1. Go to **Stat > Variance > Two Sample > data/summary**
2. Enter the sample variances or select the appropriate column
3. Select **Next**.
4. Set the null variance ratio (standard is 1) and the alternative hypothesis.
5. Click on **Calculate**.

The results should appear.

**Example 4**

**Problem**: In Example 1 in Section 11.2, we compared the average scores of men and women on a Mth096 exam. In that test, we assumed that the standard deviations of the two groups were equal. Test the assumption at the $\alpha=0.1$ level of significance.
Solution:
First, we need to check the conditions. We know from Example 1 that neither sample contains outliers, but we do not know if they come from normally distributed populations. We’ll use StatCrunch to perform Q-Q plots.

While the two plots aren’t exactly linear, it does appear that the samples could come from normally distributed populations, so our conditions are satisfied.

Step 1:
\[ H_0: \sigma_1^2 = \sigma_2^2 \]
\[ H_1: \sigma_1^2 \neq \sigma_2^2 \]

Step 2: \( \alpha = 0.1 \) (given)

Step 3: (we’ll use StatCrunch)

Step 4: Using StatCrunch:

Hypothesis test results:
- \( \sigma_1^2 \): variance of Men
- \( \sigma_2^2 \): variance of Women
- \( \sigma_1^2/\sigma_2^2 \): variance ratio
- \( H_0 : \sigma_1^2/\sigma_2^2 = 1 \)
- \( H_A : \sigma_1^2/\sigma_2^2 \neq 1 \)

<table>
<thead>
<tr>
<th>Ratio</th>
<th>n1</th>
<th>n2</th>
<th>Sample Ratio</th>
<th>F-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_1^2/\sigma_2^2 )</td>
<td>31</td>
<td>52</td>
<td>0.7472743</td>
<td>0.7472743</td>
<td>0.3959</td>
</tr>
</tbody>
</table>

Step 5: Since the P-value > \( \alpha \), we do not reject the null hypothesis.

Step 6: Based on these results, there is no evidence to support the claim that the standard deviations are not equal.
Section 11.5: Putting It Together: Which Method Do I Use?

11.1 Inference about Two Proportions
11.2 Inference about Two Means: Dependent Samples
11.3 Inference about Two Means: Independent Samples
11.4 Inference about Two Standard Deviations

11.5 Putting It Together: Which Method Do I Use?

Objectives

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

1. determine the appropriate hypothesis test to perform

Hypothesis Tests Regarding Two Populations

So we now have four new hypothesis tests to add to our arsenal. Here they are again:

Tests Regarding the Difference Between Two Population Means

In order to perform a hypothesis test regarding two population means, the following must be true concerning a sample:

1. the sample is obtained using simple random sampling, and
2. the sample data are matched pairs, and
3. the sample has no outliers and the population from which the sample is drawn is normally distributed, or the sample size is large (n ≥ 30).

Then the test statistic is

\[ t_0 = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} \]

Tests Regarding the Mean Difference

In order to perform a hypothesis test regarding the mean difference, the following must be true:

1. a simple random sample of size \( n_1 \) is taken from a population with unknown mean \( \mu_1 \) and unknown standard deviation \( \sigma_1 \)
2. a simple random sample of size \( n_2 \) is taken from a second population with unknown mean \( \mu_2 \) and unknown standard deviation \( \sigma_2 \)
3. the two populations are normally distributed or the sample sizes are sufficiently large (\( n_1, n_2 \geq 30 \))

Then the test statistic is:

\[ t_0 = \frac{\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \]

Tests Regarding the Difference Between Two Population Proportions

In order to perform a hypothesis test regarding the mean difference, the following must be true:
1. simple random samples size $n_1$ and $n_2$ are taken from two populations
2. $n_1\hat{p}_1(1-\hat{p}_1) \geq 10$
3. $n_2\hat{p}_2(1-\hat{p}_2) \geq 10$
4. both sample sizes are less than 5% of their respective populations.

Then the test statistic is:

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

**Tests Regarding Two Population Standard Deviations**

In order to perform a hypothesis test regarding the mean difference, the following must be true:

1. $\sigma_1^2 = \sigma_2^2$
2. $s_1^2$ and $s_2^2$ are sample variances from independent simple random samples of size $n_1$ and $n_2$, respectively
3. both populations are normal

Then the test statistic is:

$$F = \frac{s_1^2}{s_2^2}$$

**Choosing the Appropriate Hypothesis Test**

Now that we've done a (very) quick review of the four various tests, it's helpful to think of a flowchart when deciding which test to apply. Here's a version of the flowchart from your text:

The biggest problems usually occur between the *independent* and *dependent* samples regarding the means. They key is to determine if the samples are somehow paired. A dead giveaway is a problem with *before* and *after*. In that case, they're clearly paired, making them *dependent* samples. If you're comparing two completely different populations (the average mpg for Honda Civics vs. Toyota Camry), then you have *independent* samples.

As we mentioned in the last chapter, there's no quick and easy rule to memorize. You'll need to practice all the
problems on the following page and be sure to do all the assigned homework problems. There's also an extra review for this exam, which also helps you choose which hypothesis test to apply. It's important to practice, practice, practice!

Some Examples

It's time for examples. In each case, don't worry about actually completing the problem. Focus instead on choosing the correct hypothesis test to apply. For more practice, you should look at the Exam 4 Extra Review file, which is available in Desire2Learn.

Example 1

Janis commutes to her Statistics class at ECC. She has two possible routes and would like to determine which is optimal. To help decide, she collects travel times for 60 morning trips, 30 on each route. Her first route has an average travel time of 24.3 minutes, with a standard deviation of 3.8 minutes. The second route has an average travel time of 22.9 minutes, with a standard deviation of 4.4 minutes. Based on these data, does Janis have enough evidence to say that the second route is the optimal one?

[ reveal answer ]

Example 2

Jay and Sheila are pig farmers in south-western Minnesota. They're changing the feed they use, and they're concerned that one of the new options leads to weights in the pigs that vary too widely. To help determine which choice of feed is more consistent, they take two samples of 100 piglets each. The first sample receives feed from AgraChoice, while the second receives feed from Swine Food. After 6 months, both samples have similar average weights of nearly 200 pounds, but the standard deviations are different. The AgraChoice sample has a standard deviation of 22.1 pounds, while the Swine Food sample has a standard deviation of 24.3 pounds.

Based on these samples, are Jay and Sheila's fears founded? Does the Swine Food yield 6-month-old pigs whose weight varies more than those fed with AgraChoice?

[ reveal answer ]

Example 3

A statistics professor is interested in the success rates of his students. In particular, anecdotal evidence seems to suggest that those students who are returning to college after an absence seem to be more successful in his courses. He collects data from his and his colleagues' students over one semester, specifically focusing on whether or not students were "returning" (defined for his purposes as those with two or more years away from school) and whether or not they were "successful" (earning a C or better).

He found that out of 184 "returning" students, 132 were successful, and of 429 "traditional" students, 256 were successful. Based on these data, is there evidence to support the professor's anecdotal evidence?
Example 4

A college administrator recently learned about a new strategy for encouraging faculty participation in academic committees. She is prepared to implement it, and would like to know if it truly changes faculty participation. She chooses a random sample of 10 faculty and records their current attendance at committee meetings. She then implements the new strategy she learned and records the attendance of these same faculty. The data she collected are as follows (in meetings attended per month):

<table>
<thead>
<tr>
<th>Faculty</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>After</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Is there evidence to say the new strategy increased faculty participation in their committees?
Chapter 12: Inference on Categories of Data

12.1 Goodness-of-Fit Test
12.2 Tests for Independence and the Homogeneity of Proportions

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown? (Note: These values are different from those that used to be available on the M&M's website, but have been confirmed by ScientificAmeriken.) Do they really? Could a quality control engineer test that? How far from those expected percentages is acceptable?

These are all questions we're going to answer in Section 12.1, using something called a Goodness-of-Fit Test.

In Chapter 4, we studied relationships between two variables. We learned that we could quantify the strength of the linear relationship between two quantitative variables with the correlation.

What about qualitative (categorical) variables, though? For example, suppose we consider a survey given to 82 students in a Basic Algebra course at ECC, with the following responses to the statement "I enjoy math."

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>9</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Women</td>
<td>12</td>
<td>18</td>
<td>11</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

How do we study this relationship? Is there a way to tell if gender and whether the student enjoys math are related? In Section 4.4, we discussed construction conditional distributions and analyzing them, but can we be more precise? In fact, there is a way, and we'll study it in Sections 12.2.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::

This work is licensed under a Creative Commons License.
Section 12.1: Goodness-of-Fit Test

12.1 Goodness-of-Fit Test
12.2 Tests for Independence and the Homogeneity of Proportions

Objectives

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

1. perform a goodness-of-fit test

In Example 4 from Section 6.2, we assumed that the number of free throws made out of 10 attempts followed the binomial distribution. But does it really? And how do we know? In that problem, we assumed that the free throws were independent, but is that something we could check?

Consider a standard 6-sided die. If we assume a die is fair, each side should be equally likely. Of course, out of 100 tosses, they won't show up an equal number of times (they can't, since 1/6 of 100 is about 16.7). But how far from an equal number is acceptable?

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown? (Note: These values are different from those that used to be available on the M&M's website, but have been confirmed by ScientificAmeriken.) Do they really? Could a quality control engineer test that? Like the dice, how far from those expected percentages is acceptable?

These are all questions we're going to answer in this section, using something called a Goodness-of-Fit Test. Before we do that, we need a little background.

Observed vs. Expected Values

Consider the M&M's® example above. Suppose we purchase a standard bag of Milk Chocolate M&M's, and observe the following distribution:

<table>
<thead>
<tr>
<th>Color</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>13</td>
</tr>
<tr>
<td>Orange</td>
<td>9</td>
</tr>
<tr>
<td>Green</td>
<td>12</td>
</tr>
<tr>
<td>Yellow</td>
<td>8</td>
</tr>
<tr>
<td>Red</td>
<td>7</td>
</tr>
<tr>
<td>Brown</td>
<td>7</td>
</tr>
</tbody>
</table>

If we follow the percentages above, we would expect 24% of the M&M's® to be blue. Since we had 56 total, we would expect 24% of 56, or about 13. Similarly, we could fill out the table for the rest of the colors as follows:

<table>
<thead>
<tr>
<th>Color</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>13</td>
</tr>
<tr>
<td>Orange</td>
<td>11</td>
</tr>
<tr>
<td>Green</td>
<td>9</td>
</tr>
<tr>
<td>Yellow</td>
<td>8</td>
</tr>
<tr>
<td>Red</td>
<td>7</td>
</tr>
<tr>
<td>Brown</td>
<td>7</td>
</tr>
</tbody>
</table>

What we're not able to answer now is the severity of these differences - is it significant enough for us to say that the distribution is different from...
**Expected Counts**

In general, the expected count for each category is the number of trials of the experiment, multiplied by the probability of that particular outcome.

\[ E_i = n \cdot p_i \]

To test whether the observed values fit the stated distribution, we compare them with the expected, using the Goodness-of-Fit Test. Go to the next page to see the details and some examples.

You can also go to the video page for links to see videos in either Quicktime or iPod format.

---

### The Goodness-of-Fit Test

The Goodness-of-Fit Test is used to test the distribution of a single variable. In essence it compares the observed values with what we would expect. Before we can begin, we need a new test statistic.

### The Test Statistic for Goodness-of-Fit Tests

If we let \( O_i \) represent the observed counts for category \( i \), and \( E_i \) represent the expected counts, with \( n \) independent trials and \( k \) categories, then the formula

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]

approximately follows the chi-square distribution with \( k-1 \) degrees of freedom, provided that

1. all expected frequencies are greater than or equal to 1, and
2. no more than 20% of the expected frequencies are less than 5.

Note: If 1 or 2 fail, we can combine categories so they are satisfied.

### Performing a Goodness-of-Fit Test

**Step 1:** State the null and alternative hypotheses.

- \( H_0 \): The random variable follows the claimed distribution.
- \( H_1 \): The random variable does not follow the claimed distribution.

Note: The test is always a right-tailed test, since larger deviations from the expected values will result in larger \( \chi^2 \) values.

**Step 2:** Decide on a level of significance, \( \alpha \).

**Step 3:** Compute the test statistic,

\[ \chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \]

**Step 4:** Determine the \( P \)-value.

**Step 5:** Reject the null hypothesis if the \( P \)-value is less than the level of significance, \( \alpha \).
Step 6: State the conclusion.

Example 2

Consider the M&M's®, from Example 1. Based on the observed counts in that bag, is there enough evidence at the 5% level of significance to say that the distribution is different from what the company claims?

Chi-Square Goodness-of-Fit Test Using StatCrunch

1. You'll need to calculate the expected counts based on the assumed distribution.
2. Enter the observed counts in the first column, and the expected counts in the second column.
3. Choose Stat > Goodness-of-fit > Chi-Square test
4. Select the columns for the observed counts.
5. Select the columns for the expected counts.
6. Select Calculate.

The results should appear.

More information is available in the help file through StatCrunch.

Example 3

Consider the standard 6-sided die we mentioned earlier this section. If we assume a die is fair, each side should be equally likely. Suppose we roll a die 100 times and observe the results shown below. Is there enough evidence at the 5% level of significance to say that the die is not fair?

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Frequency</td>
<td>22</td>
<td>12</td>
<td>17</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>

[ reveal answer ]
Section 12.2: Tests for Independence and the Homogeneity of Proportions

1.2.1 Goodness-of-Fit Test

12.2 Tests for Independence and the Homogeneity of Proportions

Objectives

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

1. perform a test for independence
2. perform a test for homogeneity of proportions

In the previous section, we considered the relationship between a student's gender and whether he or she enjoys math. One question we might have as a result of this is whether we can determine whether there is a statistical test to determine if there is a relationship between the two variables.

Of course, we wouldn't be mentioning it if there wasn't! Before we discuss that test, we need a little background.

Determining Expected Counts

Let's assume that a student's gender and whether he or she enjoys math are independent. What frequencies would we expect in that case? Let's consider again the survey data from Example 2 in Section 4.4:

In that example, a survey was given to 82 students in a Basic Algebra course at ECC, with the following responses to the statement "I enjoy math."

<table>
<thead>
<tr>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Neutral</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>9</td>
<td>13</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Women</td>
<td>12</td>
<td>18</td>
<td>11</td>
<td>6</td>
</tr>
</tbody>
</table>

We then created a **relative frequency marginal distribution**, which was calculated by taking the row/column totals and dividing by the sample size of 82.

<table>
<thead>
<tr>
<th>SA</th>
<th>A</th>
<th>N</th>
<th>D</th>
<th>SD</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>9</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>30/82 ≈ 0.37</td>
</tr>
<tr>
<td>Women</td>
<td>12</td>
<td>18</td>
<td>11</td>
<td>6</td>
<td>52/82 ≈ 0.63</td>
</tr>
<tr>
<td>Total</td>
<td>21/82</td>
<td>31/82</td>
<td>16/82</td>
<td>8/82</td>
<td>6/82</td>
</tr>
</tbody>
</table>

Let's focus on the first cell - "Men" and "Strongly Agree". From the table, we can see that 30/82 or about 37% of the students were men, and 21/82 or about 26% of the students strongly agreed with the statement "I enjoy math." If they really are independent, we can use the **Multiplication Rule for independent events**, where \( P(E \text{ and } F) = P(E) \cdot P(F) \).

So if they are independent, the probability that a student is is both male and strongly agrees would be:

\[
P(\text{male and strongly agrees}) = \frac{30}{82} \cdot \frac{21}{82} \approx 0.094
\]
We can then use this probability to determine how many we would expect in that cell, if the two variables are actually independent. We just multiply the total number of individuals by the probability of being both male and strongly agreeing:

\[
\text{Expected number of students who are male and strongly agree} = \frac{30 \cdot 21}{82} = \frac{30 \cdot 21}{82} \approx 7.68
\]

In general, we can find the expected values using this formula:

\[
\text{Expected Frequency} = \frac{(\text{row total}) \cdot (\text{column total})}{\text{table total}}
\]

**Example 1**

Use the table provided and find the expected frequency for each outcome.

[ reveal answer ]

Now that we have the expected frequencies for each outcome, we need a new hypothesis test to see if these expected counts are far enough from what we actually observed to say that the variables aren’t independent.

**The Test for Independence**

The test we use to determine if there is an association between two qualitative variables is called the **chi-square test for independence**. In this test, the null hypothesis is always that the variables are not associated (independent), and the alternative is that they are associated (dependent).

The test works by comparing the observed counts with the expected counts if we assume the two variables are related. If those are far enough apart, we can say that we think there is a relationship. Here are the details:

**The Test Statistic for the Test of Independence**

If we let \( O_i \) represent the observed counts for the \( i^{th} \) cell, and \( E_i \) represent the expected counts, then

\[
\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}
\]

approximately follows the chi-square distribution with \((r-1)(c-1)\) degrees of freedom, where \( r \) is the number of rows and \( c \) is the number of columns, provided that:

1. all expected frequencies are greater than or equal to 1, and
2. no more than 20% of the expected frequencies are less than 5.

Note: If 1 or 2 fail, we can combine categories so they are satisfied.

**Performing a Chi-Square Test for Independence**
Step 1: State the null and alternative hypotheses.

H₀: The row and column variables are independent.
H₁: The row and column variables are dependent.

Note: Like the Goodness-of-Fit Test, this test is always right-tailed, since larger deviations from the expected values will result in larger $\chi^2$ values.

Step 2: Decide on a level of significance, $\alpha$.

Step 3: Compute the test statistic, $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$.

Step 4: Determine the $P$-value.

Step 5: Reject the null hypothesis if the $P$-value is less than the level of significance, $\alpha$.

Step 6: State the conclusion.

Example 2

Use the data from earlier examples to determine if gender and whether a student enjoys math are related. Perform the test at the 5% level of significance.

From earlier, we have the observed counts:

<table>
<thead>
<tr>
<th>SA A N D SD</th>
<th>Men</th>
<th>9 13 5 2 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Women</td>
<td>12 18 11 6 5</td>
</tr>
</tbody>
</table>

And from Example 1, we know the expected counts are:

<table>
<thead>
<tr>
<th>SA A N D SD</th>
<th>Men 7.68 11.34 5.85 2.93 2.20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Women 13.32 19.66 10.15 5.07 3.80</td>
</tr>
</tbody>
</table>

Chi-Square Test for Independence Using StatCrunch

1. You'll need to first enter the data, with row and column labels. (Leave the first column for row labels.)
2. Choose Stat > Tables > Contingency > with summary
3. Select the columns for the observed counts.
4. Select the column for the row variable.
5. Click Next.
6. Check "Expected Count" and select Calculate.

The results should appear.

More information is available in the help file through StatCrunch.

You can also go to the video page for links to see videos in either Quicktime or iPod format.
The Test for Homogeneity of Proportions

Suppose the Math Department at ECC would like to compare success rates in its College Algebra course based on how students placed into the class. There are currently three ways of placing into the course:

1. earning a C or better in the Mth098 - Intermediate Algebra; or
2. an appropriate placement test score; or
3. a Math ACT of 23 or better.

In this case, the department might want to analyze the proportion who are successful in College Algebra (i.e. earning a C or better). They wonder if the proportions are all the same, or if one is different. One way to answer this would be to do proportion tests with all of the possible pairs, but that would entail three separate tests like those we studies in Section 11.3.

Another option is a new test - the **chi-square test for homogeneity of proportions**. In a chi-square test for homogeneity of proportions, we test the claim that different populations have the same proportion of individuals with a certain characteristic.

Interestingly, the procedures for performing a chi-square test for homogeneity of proportions is identical to that for the test of independence.

Example 4

In the Fall of 2005, the ECC math department asked the Institutional Research department to collect data from previous semesters to analyze. The table below shows the results for Fall 2004 and Spring 2005.

<table>
<thead>
<tr>
<th>Mth098 placement ACT 23+</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>successful</td>
<td>132</td>
<td>94</td>
</tr>
<tr>
<td>not successful</td>
<td>140</td>
<td>52</td>
</tr>
</tbody>
</table>

Is there evidence to indicate that the proportion of students in each group who are successful is different at the $\alpha = 0.01$ level of significance?
Chapter 13: Comparing Three or More Means

13.1 Comparing Three or More Means (One-Way Analysis of Variance)

In Section 11.3, we compared two means from independent populations. What if we have more than two means to compare?

Suppose you own a chain of four boutique resale clothing shops. All four have been open for at least three years, and you want to do some analysis regarding their performance. You suspect that the managers at the four shops are not hiring staff of equal quality, so you take a sample of the weekly sales amounts for the employees at each location. You want to determine if any location is performing statistically lower than any of the others.

Using the strategies from Section 11.3, this would mean **six hypothesis tests** - one for each comparison (shop 1 vs. shop 2, shop 1 vs. shop 3, etc.). There has to be a better way! This is what One-Way Analysis of Variance is.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

:: start ::

This work is licensed under a Creative Commons License.
Section 13.1: Comparing Three or More Means (One-Way ANOVA)

13.1 Comparing Three or More Means (One-Way Analysis of Variance)

**Objectives**

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

1. verify the requirements to perform a one-way ANOVA
2. test a hypothesis regarding three or more means using one-way ANOVA

**What is One-Way ANOVA?**

Suppose you're an instructor teaching three sections of the same course. You suspect that the three sections are not equivalent, but you can't be sure. Since the students won't always earn the same score on every exam, we can treat an individual exam given to all the students in each section as a "sample" from each section.

Suppose the results from the first turn out as follows:

9am: $\bar{x} = 72.1, s = 2.5$

10am: $\bar{x} = 78.8, s = 2.3$

1pm: $\bar{x} = 86.5, s = 2.4$

We can clearly see that these appear to be from very different populations - the means are very different, and there's very little variation within each group. With that description, we can be pretty sure they're different.

What about a different situation?

9am: $\bar{x} = 78.8, s = 9.6$

10am: $\bar{x} = 80.4, s = 10.1$

1pm: $\bar{x} = 81.1, s = 8.7$

In this case, we can see that the means do appear to be different, but not by much. And within each sample, there is a lot of variation, so the difference in the means could just be due to the wide variation within each group.

The point here is that we can't just consider the differences in the means - whether those differences are significant or not depends on the standard deviations (and sample sizes, of course).

This is the basic idea behind One-Way ANOVA. It's one-way, because we're focusing on a single characteristic (time of class period, in our example above). And the ANOVA stands for analysis of variance. While it may seem odd that the title refers to variance when we're actually comparing means, it's actually because of a significant assumption - to perform the statistical analysis.
To perform the test, we focus on the *between-sample variation* (between the means) and the *within-sample variation* (i.e. the standard deviation). If the former is large in comparison to the latter, we can say that one of the means must be different. The test statistic that we use is another F-statistic, and it's the ratio of these two variations:

\[
F_0 = \frac{\text{between-sample variability}}{\text{within-sample variability}}
\]

In order to analyze this statistic, though, there are several requirements that need to be met.

**Verifying the Requirements**

In order to perform one-way ANOVA, the following requirements must be met:

<table>
<thead>
<tr>
<th>Requirements to Perform a One-Way ANOVA Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. There must be k simple random samples, one from each of k populations or a randomized experiment with k treatments.</td>
</tr>
<tr>
<td>2. The k samples must be independent of each other; that is, the subjects in one group cannot be related in any way to subjects in a second group.</td>
</tr>
<tr>
<td>3. The populations must be normally distributed.</td>
</tr>
<tr>
<td>4. The populations must have the same variance; that is, each treatment group has population variance $\sigma^2$.</td>
</tr>
</tbody>
</table>

Luckily, the procedure is *robust*, so slight variations from these criteria are OK. In particular, a good rule of thumb is that as long as the largest variance is no more than double the smallest, we can assume point #4 above has been satisfied. If not, there are other tests that can be performed, but they're beyond the scope of this course.

**Example 1**

**Problem:** Referring back to the example above, does it appear that the conditions are met to perform one-way ANOVA?

**Solution:** To answer this, let's consider the box plots we were provided:

9am: $\bar{x} = 78.8, s = 9.6$

10am: $\bar{x} = 80.4, s = 10.1$

1pm: $\bar{x} = 81.1, s = 8.7$

Granted, the information is a little imprecise here, but we do have 3 samples that are independent (3 different sections). Based on the boxplots, the scores do appear to be normally distributed, though without a histogram that's difficult to tell precisely. And the largest standard deviation of 10.1 is not more than double the smallest of 8.7.

So yes, it does appear that the conditions have been met and one-way ANOVA could be performed.
Performing a Hypothesis Test Regarding Three or More Means Using One-Way ANOVA

**Step 1:** State the null and alternative hypotheses.

- \( H_0: \mu_1 = \mu_2 = \mu_3 = \ldots \)
- \( H_1: \) At least one of the means is different

**Step 2:** Decide on a level of significance, \( \alpha \).

**Step 3:** Compute the test statistic, \( F_0 = \frac{\text{between-sample variability}}{\text{within-sample variability}} \) (using StatCrunch).

**Step 4:** Determine the \( P \)-value.

**Step 5:** Reject the null hypothesis if the \( P \)-value is less than the level of significance, \( \alpha \).

**Step 6:** State the conclusion.

---

**Hypothesis Testing Regarding Three or More Means Using One-Way ANOVA with StatCrunch**

1. Either enter the raw data in separate columns for each sample or treatment, or enter the value of the variable in a single column with indicator variables for each sample or treatment in a second column.
2. Go to Stat > ANOVA > One Way.
3. If the raw data are in separate columns, select “Compare selected columns” and then click the columns you wish to compare. If the raw data are in a single column, select “Compare values in a single column” and then choose the column that contains the value of the variables and the column that indicates the treatment or sample.
4. Click Calculate.

The results should appear.

---

**Example 2**

**Problem:** Referring back to the example above, use the data provided in the link below.

[test scores by section (CSV)](link)

Is there enough evidence at the 5% level of significance to support the claim that one of the classes is performing differently from the others?

**Solution:**

Assuming again the conditions have been met, we have the following results:

**Step 1:**

- \( H_0: \mu_1 = \mu_2 = \mu_3 = \ldots \)
- \( H_1: \) At least one of the means is different

**Step 2:** \( \alpha = 0.05 \) (given)

**Step 3:** (we'll use StatCrunch)

**Step 4:** Using StatCrunch:
Analysis of Variance results:
Responses stored in var2.
Factors stored in var1.

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>10am</td>
<td>24</td>
<td>80.375</td>
<td>10.124496</td>
<td>2.0666542</td>
</tr>
<tr>
<td>1pm</td>
<td>24</td>
<td>81.083336</td>
<td>8.722388</td>
<td>1.7804459</td>
</tr>
<tr>
<td>9am</td>
<td>26</td>
<td>78.84615</td>
<td>9.611212</td>
<td>1.8849137</td>
</tr>
</tbody>
</table>

ANOVA table

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>2</td>
<td>65.819214</td>
<td>32.909607</td>
<td>0.36413267</td>
<td>0.6961</td>
</tr>
<tr>
<td>Error</td>
<td>71</td>
<td>6416.843</td>
<td>90.37807</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>73</td>
<td>6482.662</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Step 5:** Since the $P$-value > $\alpha$, we do not reject the null hypothesis.

**Step 6:** There is not enough evidence at the 5% level of significance to support the claim that the mean score from one of these sections is different from the others.