Elementary Statistics

Lectures, Exercises, Activities, and Sample Tests

by

Deborah H. White

Table of Contents

Lecture #1: Course Introduction

Almost all of the statistical analysis we'll be doing in this course will be based on the Class Data Base, which is a spreadsheet in which all of your answers to the Student Survey you're going to fill out will appear. Each of you will have a line in the spreadsheet for your answers.

In addition to getting information for our data base, your job today is to get acquainted with certain words which are used to describe the **variables** that I'll be asking about. Variables in this sense mean things that can vary from person to person. For instance, some of you are male, and some of you are female. Sex is a variable. So are ZIP code, attitude toward taking math classes, and height. And these variables have characteristics which we describe with special words that I'll be using all semester.

So even though you're pretty sure you know the answers to these questions, I'm going to give a little spiel about how each should be answered, using the special words. Feel free to make notes on your Student Survey sheet; you'll get it back after I compile the data.

Here's the first question:

What sex are you (M or F)?

This seems simple enough. Your response should be either an M or an F. We call this kind of variable **binomial** – two names, like the $x + y$ of algebra. It's true that some people prefer to be thought of as being neither male nor female, or being both male and female, but for the purposes of this survey I'm asking you to make a choice. And the thing about binomial variables is that the only way we can analyze them is to say what fraction of the group is in each category. We can't find the "average" of the genders – the very idea sounds silly. We can only say that ~~% of the group is male, and ~~% of the group is female.

In addition to being binomial, sex is also a **categorical** or **qualitative**, as opposed to a **quantitative**, variable. In other words male and female are categories and qualities, not numbers. It's true that we could artificially assign numbers to them – call being male #1 and being female #2 or *vice versa*, but these numbers would be meaningless.

You probably didn't think there was so much to say about such a straightforward question. Statistics is a very wordy subject!

Here's the next question:

Are you a native of Mendocino or Lake County?

The answer I'm looking for is a Yes or a No. So this is also a categorical, binomial variable, though the categories are different than those for the question about

what sex you are. I'm considering the two-county area to be a single entity. Don't put down which county you were born in of the two. (Don't worry if you already did, though; I'll take it as a Yes.)

How about this situation: Your parents live in Ukiah. One morning your mother wakes up and notices that she's gone into labor. Her obstetrician or midwife is in Santa Rosa. She goes to Santa Rosa, and while she's there you are born. The next day or so she comes back to Ukiah and you begin your childhood. Are you a native of Mendocino or Lake County? I'm not sure. This is one of many cases that illustrate the phrase I once saw on a T-shirt: Statistics is never having to say you're certain. If this or a similar story describes your birth, you'll have to make your own decision. There's not a whole lot at stake here, so do what you wish.

Are you a graduate of a Mendocino or Lake County high school?

Another qualitative, binomial variable, with possible answers Yes and No. Nothing new here. But there are a couple of dicey cases. What if you're still in high school in Mendocino or Lake County? It's not like you went to some out-of-area high school and then came to Mendocino College. You're probably going to graduate, given that you're already taking college-level classes. Or how about if you got a GED? Or you were home-schooled? In all these cases you should answer the way that feels right to you. If you despise the very idea of high school, say No. Otherwise, you might as well say Yes, because to me you are a local student, and that's what this question is getting at.

Now here's something new:

What is your favorite color?

Obviously there are more than two answers to this question, so instead of being binomial, favorite color is what is called a **multinomial** variable – **many** names. But it is still a categorical variable, not a numerical or quantitative one. I suppose you could give it as a wavelength, but that isn't practical. You're only allowed one favorite color, so don't put down "blue or green" or something like that. You have to choose. But you can give it as many descriptive adjectives as you like. My favorite all time was Blood-and-Guts Red. Very vivid. Some people don't have a favorite color; they don't really care. If you're one of those, leave this question blank. We're not forcing you to pick. Or what if you like two or more colors equally and can't bring yourself to insult one or more by picking a favorite? Leave it blank. These are two very different ways of not having a favorite color, but they will look the same in our survey.

Just like with the first question, we'll be able to analyze the data only by giving the percent of the group which picked each color. There will be more than two percents needed, but we won't be able to give an "average" favorite color.

And here at last is a number:

What is your ZIP code?

Your answer should be a five-digit number. If you've just come here from somewhere else, give your ZIP code here. Most likely it will be 954--.

We didn't use to have ZIP codes. You'd just write the address, the name of the city or town, and the state. Large cities did have something called zones, so you'd write, for instance, New York 9, New York. Then in 1963 the **Z**oning **I**mprovement **P**lan went into effect, and every location in the United States had a ZIP code.

But even though the ZIP code is a number, it doesn't act like a number. You can't give an average ZIP code for a group; that makes no sense. You can't compare ZIP codes for size. People from Willits cannot claim they're better than people from Ukiah because their ZIP code (95490) is 8 bigger than Ukiah's (95482). Then how are they assigned? Well, each area is given three digits (954 in our case), and these are arranged geographically. But within each such group the final two digits are assigned alphabetically: **B**oonville, 95415; **L**akeport, 95456; **P**otter Valley, 95469; **R**edwood Valley, 95470, etc., with little gaps allowed for future towns perhaps.

So again the only way to analyze a set of ZIP codes is by saying what percent of a group has each ZIP code. You could say then that ZIP code is a multinomial, categorical variable. But there **is** something numerical about ZIP codes; they **are**, after all, numbers, so you could also say that ZIP codes are a quantitative variable, but at the very lowest rung on a ladder which we call **levels of measurement**. This first level is the **nominal** level, meaning that the number is really only being used as a name. We'll be going up through these levels for the rest of the survey.

The next level is represented by these questions:

Which statement best characterizes your attitude towards taking math classes?

1) **If I never have to take another math class it will be too soon.**
2) **I would prefer not to take a math class.**
3) **I can take them or leave them.**
4) **I enjoy math classes.**
5) **I adore taking math; it's my favorite thing to study.**

Which statement best characterizes your attitude towards social media? (Facebook, Snapchat, Instagram, Twitter, etc.)

1) **I never use social media and can't stand them.**
2) **I seldom use social media.**
3) **I can take them or leave them.**
4) **I enjoy social media.**
5) **I love using social media; I can't get through the day without them.**

Please pick one number to answer each question; don't, say, circle the 3 and the 4. Pity the poor person compiling the surveys; don't make me choose.

Now, these answers are unquestionably numbers, and they function as numbers in many ways. Yes, you could say that the name of your attitude toward math classes is 5, or 1, but that loses the richness of the response. Answering with a 5 means you have a very positive attitude about the subject; 1 a very negative one. There is an order to the responses; the larger the number the more positive the response. We call this the **ordinal** level of measurement for this reason. Analyzing a set of responses we might give the percents for each answer, but we might also pool the 5's and the 4's as those with positive attitudes, or the 1's and 2's as those with negative attitudes, or even the 3's, 4's and 5's as those who **don't** have negative attitudes, and so on. We can rank the attitudes in a way that makes sense, hence the name ordinal.

But surely some characteristics of numbers are lacking here. For instance, we can't say that a person who answers with a 5 is more positive than a person who answers with a 4 **to the same extent that** a person who answers with a 4 is more positive than a person who answers with a 3. In other words, we can't make inferences about the differences between the responses, only their rank. And it would be ridiculous to assert that a person answering 4 is twice as positive as a person answering 2. These sorts of numerical characteristics belong to the remaining, higher levels of measurement.

Next question:

If you own a car, what year is it?

You should either leave this question blank, if you don't own a car, or put in a four-digit number starting with either 19 or 20. No fractions (though I did once own a 1988½ Ford!). What if you own two cars, or a car and a truck? Pick one and give its year. The other won't know it's being slighted.

The year your car was made is definitely a number, and the newer the car the bigger the number, so this variable is ordinal, but it's also something more. A car made in 2003 is 3 years newer than a car made in 2000, and a car made in 2006 is also 3 years newer than a car made in 2003, so the **differences** between years have significance. We call these differences **intervals**, and so year of car is at the **interval** level of measurement. That means we can reasonably say how much newer or older one car is than another. But, although we can say that a certain car is half as old as another, or twice as old, we can't do so using the years the cars were made without subtracting them from the current year. If cars had been made in the year 1000, we couldn't say that such a car is twice as old as a car made in the year 2000.

How about this question:

What size shoes do you wear?

I know that sizes vary and that men's sizes are different from women's sizes (a man's 6, I believe, is the same actual size as a woman's 7½, on the whole) and also that there are different size systems in different parts of the world (a man's 6 in the United States would be a 38 in Europe, for instance). Then there's the fact that the sizes go up to 13 ½ for children and then start all over again at 1 for larger feet. However, for the

purposes of this survey, just put down one whole number, or a whole number followed by ½, and leave it at that. No need to say if it's a man's or a woman's size; you've already stated your sex. Plus, I've found analyzing these numbers later in the semester goes just as smoothly if we ignore the gender of the shoe size.

Shoe size is definitely a number, definitely ordinal (10 is bigger than 8, 8 is bigger than 6), and definitely at the interval level of measurement. Size 10 is the same amount bigger than 8 as 8 is bigger than 6. In fact, each size bigger is supposed to be ¼ inch longer. A woman's 6 is 9¼ inches in length, a 7, 9½. Why don't they just use actual lengths instead of these crazy sizes?? Well, you'd still have to decide between inches and centimeters, in which a 6 is 23.5 cm and a 7 is 24.1 cm. Korea seems to have the idea – the size is given as the actual length of the shoe in millimeters! (See http://www.i18nguy.com/l10n/shoes.html#adult for the full story.)

So shoe sizes in our system still lack that last, most sophisticated property of numbers – the ability to say that one size is twice as big as another (the **size** might be, but the shoes aren't). But we get that property with the next question:

How old did you turn on your last birthday?

You'll want to put down a whole number here. Notice I didn't ask how old you are. Some people might answer, 20 and 4 months, or 22½ , or, in the manner of children eager to be accorded their full measure of maturity, 19 7/12.

Age on last birthday is certainly a variable at the ordinal level of measurement, showing that one person is older than another, and certainly a variable at the interval level of measurement, enabling us to talk meaningfully about the difference in age of two people, but it's something else also. We **can** say that one person is **twice** as old as another, say if one is 36 and one is 18. We can talk about the **ratio** of their ages as 2:1. So this variable is at the **ratio level of measurement**, and that's the highest you can get. You can also talk meaningfully about an age of 0, though few people do, although my children used to refer to babies before their first birthday as being 0.

But something else is going on here too. You were asked to **round** your age to the nearest whole number. In fact you were asked to round your age **down** to the nearest number (though if you have a birthday in the next week it would be okay to put your new age). And that's because you don't **count** your age, you **measure** it. This is a fundamental distinction between different uses of numbers for different variables. Some variables are measured, like age, or height, as you'll see below, and some are counted. When you measure a variable, you always have to determine how fine a measurement to use. You could give you age to the nearest year, or month, or week, or day, or hour, or minute, or second or even nanosecond, but you'd have to keep changing your answer as you got more precise. We say that age is a **continuous** variable, meaning that it could be measured to any degree of fineness, so to avoid people measuring to different degrees I asked you to round down to the nearest whole number. Only variables which are measured are continuous.

Here's another measured, continuous variable at the ratio level:

What is your height to the nearest inch?

If you wrote 5'6", kindly convert that to inches ($12 \times 5 + 6 = 66$ inches). How annoying that a foot has 12, not 10 inches, which would certainly simplify the calculation! And if you wrote down a height ending in ½, round it up or down as you please.

Even though your height wouldn't continually have to be changed if a finer measurement were asked for, as in the case of age discussed above, all the other considerations would apply, so I specified where to round to keep the responses all having the same format.

And the final question:

How many pets do you have?

What's a pet? How about that goat? Is it a pet, or a farm animal? It's up to you. How about pets that belong to your whole family, not just you? It's also up to you whether to include them or not (but no siblings, please – they are **not** pets). And those guppies? Hard to tell how many of them there are at any time, but, again, your decision.

So clearly I'm expecting a whole number here, and 0 is a perfectly fine one. Don't put down "None." In that case you have a number of pets, and that number is 0.

You can see that number of pets is at the ratio level of measurement – you can have twice as many pets as someone else, and you can have 0 pets.

But what you can't have is 4 ½ pets, or anything ending in a fraction. And that's because you don't measure your pets (at least not to find out how many you have). You **count** them. Number of pets is a variable which is counted, and we call that kind **discrete**. This isn't the kind of discreet that describes people who are good at keeping secrets – note the different spelling. This kind of discrete has the same ending as "concrete," which is derived from Latin words meaning "growing together," and the "crete" is the growing part, so discrete meaning growing apart. In other words, discrete numbers have gaps between them. You can have 2 pets, or 3 pets, but no number of pets between those two numbers.

Okay, you've done the survey. And you've encountered a lot of the basic words we used in **descriptive** statistics, which is the part of statistics, obviously, in which we **describe** things. I wish that language were neat enough to make a nice flow chart of these words in a well-defined order, but language isn't like that. Concepts overlap and intertwine.

Next class we'll go over all these words again, and some others too.

Lecture #2: What Is Statistics Anyway?

Perhaps somebody will ask you this semester, "What **is** statistics, anyway?" It's good to have an answer, if only to shut them up. Here's one from the PBS Parents website (http://www.pbs.org/parents/earlymath/resources4.html): Statistics is the mathematics of collecting and analyzing data to draw conclusions and make predictions. Good enough.

Be sure to distinguish this meaning from the common usage of the word statistics in our society, typified by the sports pages with their small-type "stats." They **are** a form of statistics, but only a very small part of what we're doing in this course. They're also a plural noun – we might say that someone's stats are good – but when we use the word, oddly enough, it's singular. See the definition above. It's no coincidence that the word shares its first four letters with the word "state," because it was first used in connection with data gathered by the state, a general term for an independently-governed unit. I can imagine that the chief use was information gathered for purposes of taxation, representation, and the like.

As you saw in the last lecture, the way we practice statistics is by gathering information about a **variable**, which is something that can take on different **values** – let's just say something that can **vary**. (When they're numbers they're often referred to as **random variables** in the context of statistics.) And a value that a variable **does** take on in a certain person or whatever we're studying is called a **datum**, Latin for "given." Datum is a singular word, and the plural is **data**, which is a little confusing since we're so used to the Spanish ending -a being singular. It even sounds a little funny when we say, "The data show…" or, "The data do…," but it's correct.

If we collect data on a variable, we call that collection a **data set**.

Descriptive and Inferential Statistics

There are two parts to this course, with a small bridge-like part in between. The first part is about **descriptive** statistics, in which we develop ways to describe data sets.

But we're interested in more than just what **is**, and what data we've been able to gather. We want to draw conclusions and make predictions in general, not just talk about what we've found out. Yes, 460 of the 1000 people we asked said they're going to vote for Candidate X, but can we say that she has 46% of the entire vote committed to her? The data set consisting of the 1000 yes's and no's that we collected is called a **sample**, and any number that describes this sample (for instance that 46% are yes's) is called a **statistic** (singular).

Now **probability** comes in. It's the "bridge" I described above. We calculate how likely it is that, assuming we didn't ask our question at the headquarters of Candidate X's campaign, or that of her opponent, the entire electorate that turns out to vote will give 46% to her, or close to that, and if so, how close can we expect?

Probability, which is the study of likelihood of events, is the tool used to make the transition from descriptive to **inferential** statistics. An **inference** is a conclusion that you draw from information you receive. Inferential statistics takes information from the sample, in particular its statistics, in the above sense of the word as the plural of statistic, and applies it to the **population**, which is the entire potential collection of yes's and no's, to make guesses about the numbers which describe the population, which are called **parameters**. Again, this isn't exactly how we use that word in everyday conversation, if it comes up, but this is the specialized vocabulary of statistics.

It's easy: **S**amples have **S**tatistics; **P**opulations have **P**arameters. And though you may well want to say that they're equal, as in the case of our 46%, they have totally different significances: the former is reality, the latter a guess.

Variables

Now we'll discuss briefly the ways that variables are classified, and many of these ways should sound familiar from the last class.

First, there's the **qualitative/quantitative** distinction. Some variables are characteristics – not numbers anyway – and some are numbers. Qualitative variables are also called **categorical** because the data fall into categories.

If a variable is quantitative, it can be at one of four possible **levels of measurement**. The lowest level is the **nominal**, where the number is simply a name. Last time we had ZIP codes. There's also bus route numbers, Social Security numbers, phone numbers, etc. Next comes the **ordinal** level, where numbers can be ranked. Attitudes toward math and social media were used last class; further examples are numbers of stars in restaurant or movie reviews and ranking of tennis players or golfers, to name a few.

The **interval** level may be the hardest to understand. Variables at this level may be ranked, with the differences between data values having a definite meaning and significance, but there is no meaning in phrases like "twice as big," and there is no meaningful zero. So, as with shoe size, temperatures in Fahrenheit and Celsius are examples of this level. 0° F has absolutely nothing special about it, and 0° C is special only because it's the freezing point of water. Now, with temperature measured on the Kelvin scale, where 0 means no thermal activity whatsoever, you would be able to talk about "twice as hot," because that would mean twice as much thermal activity, and, of course, there **is** a meaningful zero. But I don't see us using that scale in ordinary life any time soon (0° K is -459.67° F and -273.15° C).

The highest level, the **ratio** level, is achieved by variables like age, height and number of pets from last class, the Kelvin scale, income, weight, speed, and many, many more.

When variables are at the interval or the ratio level of measurement, they can be divided into two kinds, those that are **counted**, and those that are **measured**. The

counted ones are called **discrete**, because they have gaps between them, since they are whole numbers. The ones that are measured are called **continuous**, which means that they can be given to any degree of accuracy, depending on the measuring device and on how you decide to round the numbers. This leads to the concept of **boundaries**, which we'll talk about now and then hardly ever again.

Boundaries

What's the difference between saying that something is 28.5 cm long and that it's 28.50 cm or 28.500 cm, for instance? This is the subject of boundaries. Let's say that I weigh something on a scale that measures whole grams, and that it comes out to be 27 g.

(I know that grams aren't a measure of **weight**, really, but rather of **mass**. In other words, something that has a mass of 27 g has that mass no matter where you weigh it – on earth at sea level, on earth on a high mountain, on the moon, on Jupiter, wherever. But its weight varies depending on the gravitational force acting upon the mass. The metric unit of weight is actually called the **newton**, and it's the force required to cause one kilogram to accelerate one meter per second every second. A kilogram at the earth's surface has a weight of about 9.80 newtons. The unit of weight in our system (the U.S. or imperial system) is the pound, and the unit of mass has the funny name of "slug." Nevertheless, I'll continue to talk about the weight of 27 grams, even though it's not correct.)

Anyway, nothing really weighs exactly 27 g. It's a little more or a little less. So when you've weighed something to the nearest gram and you get 27 g, that actually means that the object weighed between 26.5 g and 27.5 g. These numbers are called its **boundaries**, and the difference between them (1 g) shows the accuracy to which you're weighing the object.

Now suppose you get a more accurate scale, one that weighs to the nearest tenth of a gram, and you weigh the object on it, and it reads 27.3 g. (So it would have read 27 g on the first scale.) Again, nothing weighs exactly 27.3 g. An object which registers that weight actually weighs between 27.25 g and 27.35 g – those are its boundaries, and they are 0.1 gram apart. (Don't worry about something weighing exactly 27.35 g, and wouldn't it be rounded to 27.4 g, because nothing weighs exactly 27.35 g.)

Here comes another scale, one that weighs to the nearest hundredth of a gram. (A small paper clip weighs about a gram, so you can see how accurate this is getting!) Now the object registers 27.32 g. That checks with the previous measurements. You know the drill: its boundaries are 27.315 g to 27.325 g, a difference of 0.01 grams. One further scale, this one weighing to the nearest thousandth of a gram, tells us the object weighs 27.318 g. Its boundaries are 27.3175 g to 27.3185 g, a difference of 0.001 g.

These are boundaries. They split the difference. You can see how to form them. The upper one is easier: just tack on a 5. The lower one involves decreasing the last digit by 1 and **then** tacking on a 5. This table summarizes what we just did:

Accuracy of Scale	Weight	Lower Boundary	Upper Boundary
Whole gram	27	26.5	27.5
Tenth of a gram	27.3	27.25	27.35
Hundredth of a gram	27.32	27.315	27.325
Thousandth of a gram	27.318	27.3175	27.3185

It's a little harder when the weight ends in a 0. Suppose you weigh a different object, and when it comes to the scale that weighs to the nearest tenth of a gram, it registers 27.0 g. (This is a slightly lighter object.) You form the boundaries the same way, but when you decrease the last digit by 1, you have to borrow from the 7, giving a lower boundary of 26.95 g. The upper is 27.05 g.

If you then use a scale that weighs to the nearest hundredth of a gram, and it comes out 27.00 g, the boundaries are 26.995 g to 27.005 g.

So the significance of those zeros at the end is to tell you how accurately the object was weighed. An object described as weighing 27 g might be the same as one weighing 27.0 g, but you can't tell for sure.

These are boundaries. They come into play only for continuous, measured variables. If you say you have two pets, you are most definitely **not** saying you have between 1.5 and 2.5 pets. Well, you are, but it's kind of silly.

Sampling

This is a course in statistics, not research methods, also a very interesting field of study, so for the most part we just say we have a data set from a sample to analyze, and we don't go into how we chose the sample, but just for today we're going to talk about methods of collecting data, often referred to as **sampling techniques**.

First, there is the **random** sample, in which every member of a population has an equal chance of being selected for the sample. We'd like all samples to be random, but how do you actually accomplish that? Say you're doing a poll. You assign a random number to each registered voter, select a certain number of these numbers, and contact those voters. Oops. Some of them aren't home, some have had their phones disconnected, some won't talk to you, and sometimes the phone is answered by a three-year-old.

You could try a **systematic** sample. You pick every tenth number in the phone book, for instance. (Let's ignore the problem of cell phones.) Even if every person answered the phone and your question, you wouldn't have a truly random sample, because people with the same name would have their chances of being selected greatly decreased, because of not being ten or more apart in the phone book. Still, it's less work

than generating and assigning random numbers, and not at all a bad way to get close to a random sample.

What you'd like to avoid is a **convenience** sample. You're doing product research on paper towels in a mall, and you flag down shoppers and ask them if they'd mind answering a few questions about paper towels. Who's going to say yes? People with time on their hands, not dragging little children or having any number of other impediments. ESPN scrolls: "Should the manager remove the pitcher after this inning? To vote, log on to www.espn.com, etc., etc." Who is going to do that? People who care one way or the other. We won't know which group, the yes's or the no's, will respond more, but we know it's not in any way a random response. Of course, it's not a matter of life and death anyway. The advice columnist asks, "Parents, if you had it to do over again, would you have children?" You get the idea.

Here are two kinds of sampling techniques, each a valid approach, which sound a bit alike but aren't. Try to distinguish between them.

First, there's **stratified** sampling. Think of strata, or layers, of rock. Say you want to know how many units students are taking, but you want to be sure that you get a fairly equal number of men and women in your sample. So instead of getting a random sample from the whole group, first split the groups into the two sexes and then get a random sample from each group.

Then there's **cluster** sampling. This is actually how unemployment figures are estimated, and it's often used in evaluating medical techniques. You select at random a bunch of neighborhoods, or hospitals, and then you get the information about every person or procedure in the selected parts. It's just the reverse of stratifying, where you split first and then randomly select; here you randomly select the groups and then try to do a **census** of the groups chosen.

The mention of a census brings up a final point. Sometimes you **don't** want a sample – you want information about the entire population. That's what a census is, and of course it appears in the U.S. Constitution as something we do as a nation every ten years. Of course the problem with a census is not being able to track down every last person, no matter how much effort is made. Recently there has been a movement to apply concepts of statistics to the census results, making estimates of how many people have been overlooked, and what their characteristics are, which is rather a controversial approach, because some politicians would rather just overlook the missed people, and some wouldn't.

Several years ago it occurred to me to wonder why the word "census" was so similar to the word "censor," since it seemed that their meanings had nothing in common. But it turns out that in ancient Rome the government official responsible for counting everything was also in charge of deciding which writings and performances were not suitable for public consumption. Hence the census official, or censor, became a censor in our sense of the word.

Observational and Experimental Studies

Here's another topic best left to a course in research methods but which we'll touch on today. There are two ways to get data – to notice it and to make it happen.

If we just notice it, or observe it, we're doing an **observational study**. We call it this even if we are asking someone a question or measuring them or having them fill out a survey. The point is, we're not altering the information, just collecting it.

But sometimes you want to manipulate things a little to see what effects are produced. When you do that, you're conducting an **experimental study**, and there are all sorts of terminology and protocols and ethics associated with doing so. Let's use as an example trials of a new drug meant to reduce blood pressure. The drug (and the dosage at which it is given) is the **independent,** or **explanatory variable** – you want to see if it has an effect on blood pressure – and the blood pressure readings are the **dependent,** or **outcome variable**.

But of course you don't just give the drug to all patients with high blood pressure and then proclaim that your drug works when average blood pressure drops, because of course there might be other reasons for the drop. So you give some patients the drug, and they make up the **treatment group**, and you withhold the drug from others, and they comprise the **control group**. And, since people's blood pressure might well drop just because they believe that the new medicine will be effective, or just because they believe that someone is taking their problem seriously, you don't just ignore the control group. You give them a fake pill, a **placebo** (Latin for "I will please") and make sure that the patients have no idea which group they're in and which pill they're taking. And when they have their blood pressure checked, you have it done by people who are similarly ignorant; otherwise they might inadvertently affect the reading. An experiment done this way is called **double blind**. You look for a difference in the average blood pressure of the treatment group compared to the control group after the drug has been taken for a suitable length of time, and if the difference is big enough (and we spend a lot of time in this course talking about how to find out if it **is** big enough), you declare that your drug works.

Even with all of these precautions, it might turn out that your drug "worked," not because of the specially-developed chemical you put it in, but because of some supposedly unimportant substance you used to fill up space in the pill. This might be discovered some time later when the drug stops working because you started using some other filler. When something else beside the independent variable is responsible for a difference in the dependent variable between the control group and the treatment group, we call that something else a **confounding variable**, because it confounds, or confuses, the analysis of the effect.

Lecture #3: Making Data Sets into Tables and Graphs

Looking at something like the Class Data Base can be a little overwhelming. It contains **so** much information, so much raw (unsorted) data. I've helped a little by putting all the males first and then all the females, both because this gives us two smaller sub-groups to look at and also because later in the course we might want to compare the statistics for males and females.

But you'd be hard put to summarize any aspect of the class using this sheet alone. So today we'll consider ways you can organize the data sets into tables and graphs to help see their essential natures. And to do this you have to consider whether the variable is qualitative or quantitative, and, if quantitative, what level of measurement it possesses, and if it's discrete or continuous, because this determines the kinds of tables and graphs you can use.

Frequency Distribution Tables

First of all, let's do tables. Say, for instance, that you wanted to make sense of the men's ZIP code data. ZIP code is a quantitative variable at the nominal level of measurement, almost like a categorical variable, in that there's no real order to the ZIP codes. So we make what's called a **categorical frequency distribution table** (easier than it sounds). We start reading down the list of ZIP codes, and every time we find a new ZIP code we write it down and place a tally mark next to it. When we run across it again, we just make another tally mark. Tally marks are a handy way of keeping track, because after four vertical lines we cross them diagonally for the fifth line, like a little bundle, and then when we're done we can just count by fives and add on any extras. Other cultures use other systems to make the bundles. Speakers of Mandarin do it like this:

The completed pattern is the character "jen," and it means straight, or upright.

After you've tallied all the men's ZIP codes, count the tally marks and write that number in a new column. These numbers have a special name that we'll be using all semester. They're called **frequencies**, and that just means how many there are in a category, or class. The symbol for frequency is ***f***. When we add up all the frequencies, which we symbolize by writing $\sum f$, where $\sum$ is the capital Greek 'S,' (pronounced "**sigma**") and obviously stands for sum, we get the size of the sample which we are tabulating. We call this sample size ***n***, and we will all semester. Make sure you use a

lower-case *n* for the sample size; we need the upper-case for something else. So here's an equation: $n = \sum f$. You're learning the notation of descriptive statistics!

Finally, maybe you're interested not so much in how many men have, say, ZIP code 95490, but in what fraction of the group does. We call this fraction the **relative frequency**, and it has the formula $\frac{f}{n}$. We can give it as a fraction (not so helpful) or a decimal or a percent, and if it's one of the last two, we might have to round, so we specify how that should be done. Let's give it as a percent to the nearest whole percent.

Here's how the finished product looks:

Class	**Tally**	**Frequency (*f*)**	**Relative Frequency** ($\frac{f}{n}$)
95490	~~1111~~	5	9%
95482	~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ 111	38	72%
95458	1	1	2%
95453	11	2	4%
95422	1	1	2%
95481	1	1	2%
95469	1	1	2%
95470	11	2	4%
95449	1	1	2%
95444	1	1	2%
		$n = \sum f$ =53	$\sum \frac{f}{n}$ = 101%

The percents don't always add up to 100, as you can see, what with rounding up and down. As long the sum is close to 100, you probably haven't made a mistake.

So much for quantitative data at the nominal level of measurement, and categorical data. What about a variable at the interval or ratio level? (We'll skip the ordinal in this discussion, but you might want to think about it.)

Let's look at the women's ages, as an example. Instead of just listing them as they come, we want to look at them in order, since the order of the numbers means something. Furthermore, we don't want to look at each age separately. That would be

too many separate numbers. What we want to do is **group** the ages into a small enough number of groups that we can see any patterns that exist. So what we're going to make is a **grouped frequency distribution table**.

But how do we group them? What we do is pick two numbers, one called the **lower limit of the first class**, and one called the **class width**, such that if we all use these numbers, and use them correctly, we'll get identical tables.

The lower limit of the first class is the smallest number we're going to tally. It must of course be either the age of the youngest person or an even younger age. The class width is a little more complicated. It is how many separate ages are in each class. It is also the difference between the lower limit of successive classes (or for that matter of the upper limit of successive classes).

Let's use 15 as the lower limit of the first class and 5 as the class width. This gives us class limits of 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59 and…oh, we're done now, because nobody older than 59 was in the group.

We're going to tally the ages just as we did the ZIP codes, but first we need to address a concept that's important when the variable we're looking at is continuous: What if we were rounding to something finer than whole years? What if a person were 19 and 4 months, or something like that? We don't really want gaps between our classes. So what we do is create something called **class boundaries**, as distinct from class limits, which split the difference between the upper limit of one class and the lower limit of the next higher class. So the boundary between the first two classes is 19.5, halfway from 19 to 20, and it is both the upper class boundary of the first class and the lower class boundary of the second class. (This should remind you of the concept of boundaries as covered in the last lecture.) We continue in this pattern creating class boundaries, and we follow the pattern for the lower class boundary of the first class and the upper class boundary of the last class, even though there is no splitting the difference going on. In this way the classes butt right up against each other, which will be useful later in the course.

This time, let's write the relative frequencies as decimals to the nearest hundredth. Here's the table:

Class Limits	**Class Boundaries**	**Tally**	**Frequency (*f*)**	**Relative Frequency** ($\frac{f}{n}$)
15-19	14.5-19.5	~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ ~~1111~~ 1	26	0.39
20-24	19.5-24.5	~~1111~~ ~~1111~~ ~~1111~~ 11	17	0.25
25-29	24.5-29.5	~~1111~~ 11	7	0.10
30-34	29.5-34.5	1111	4	0.06

35-39	34.5-39.5	111	3	0.04
40-44	39.5-44.5	11	2	0.03
45-49	44.5-49.5	111	3	0.04
50-54	49.5-54.5	111	3	0.04
55-59	54.5-59.5	11	2	0.03

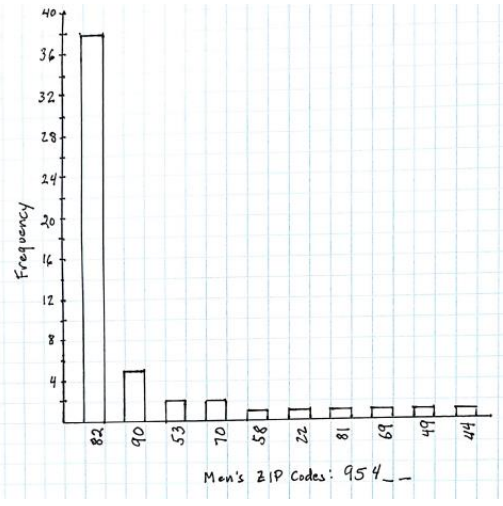
			$n = \sum f = 67$	$\sum \left(\frac{f}{n} \right) = 0.98$

The thing about both of these tables is that they show the nature of the data sets so clearly. ZIP code 95482 is by far the most frequent ZIP code for the men. Women are most likely to be between 15 and 19 years old.

Graphs

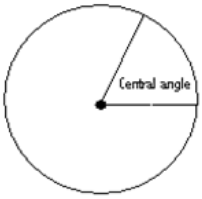
So now we'll make pictures, or graphs, of these tables, as another way of comprehending the patterns. Again, there will be distinctions among the ways of doing so depending upon the nature of the data set being portrayed.

In dealing with categorical data or quantitative data at the nominal level of measurement, a very good picture is called a **Pareto chart** (named after Vilfredo Pareto, 1848-1923, who was known for much more important achievements in the field of economics). His idea was to make a bar graph, with the bars usually not touching, arranged from the class of the largest frequency to that of the smallest. The bars could be horizontal or vertical. I've seen it most often used in the newspaper, when readers are asked, say, to name the most important problem facing their community, and a categorical frequency distribution is made to determine the length of the bars. Usually relative frequencies are used, but in this example of a Pareto chart of men's ZIP codes, I'll stick with plain old frequencies.

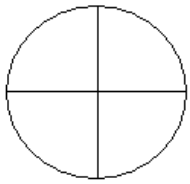


Do you see how powerful this image is? It's obvious which the most common ZIP code is, from its position on the far left and by the length of its bar. When ZIP codes are tied in frequency, they can go in any order.

Another graphical technique to use with qualitative variables and quantitative variables at the nominal level of measurement, and one you're surely acquainted with, is the **pie chart**, much beloved of comic-strip writers and advertisers. What you do is slice up a "pie" so that the size of its pieces indicates the size of the frequencies of the different categories. And it happens that the size of the piece is determined by what's called the **central angle**, which is the angle made by the two edges of the slice assuming you've started from the exact center of the pie:



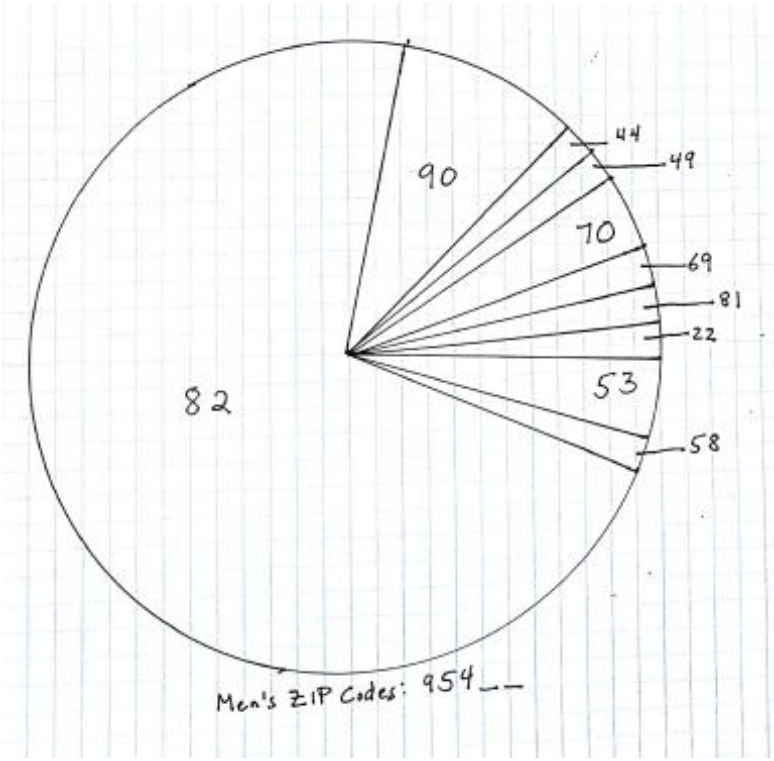
We measure angles in **degrees,** and there are 360 of them around the center of the pie. (That's because you make four right, or 90°, angles if you cut the pie into four equal pieces):



So we have to determine how many degrees to make the central angle for the slice representing each class, and for that we use the formula $\frac{f}{n} \cdot 360^\circ$, which splits up the central angles precisely proportionately to the frequency of the classes.

The angle for the 95482 slice would be $\frac{38}{53} \cdot 360° \approx 258°$, for 95490, $\frac{5}{53} \cdot 360° \approx 34°$, for 95470 and 95453, $\frac{2}{53} \cdot 360° \approx 14°$, and for each of the other six, $\frac{1}{53} \cdot 360° \approx 7°$. These angles total 362°, very close to 360°.

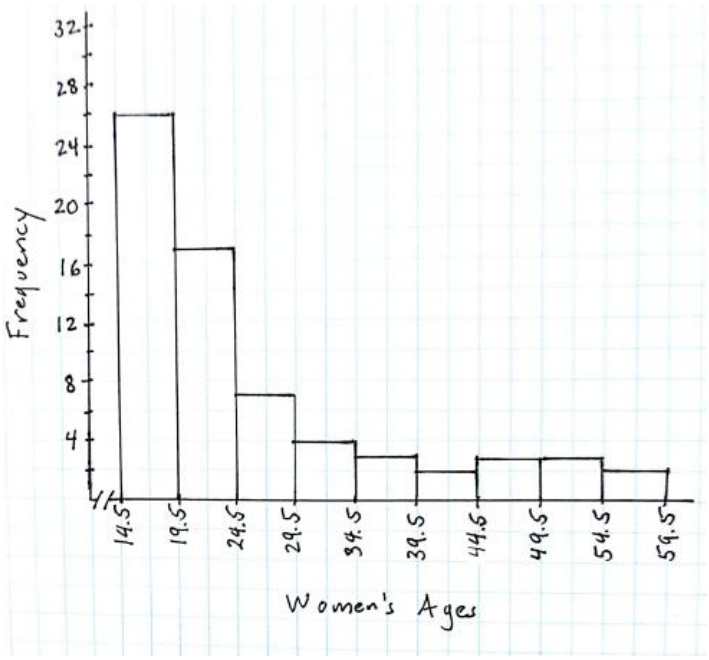
If you're not sure how to use a protractor, consult a website like this one: http://www.mathsisfun.com/geometry/protractor-using.html. Then start cutting your pie. You can start with any ZIP code you want. Make a first cut, and then use that to make a second cut that makes the slice have the central angle of that ZIP code. Use that second cut as the basis of the central angle of another ZIP code, and so on. The final piece won't have to be measured; it will be what's left from the one before it and the first slice. Don't forget to label each slice with its ZIP code, or the graph will be meaningless. If a piece is too small to label, do what maps do with states like Rhode Island: write the name outside the circle with a line to the appropriate slice. Here is the **pie chart:**



Histograms

Neither the Pareto chart nor the pie chart is suitable for variables at the interval and ratio levels of measurement, because you can't go putting them in any order in the chart. The best way to convey them graphically is called a **histogram**. It's a bar graph in which the bars touch, and so we use the class boundaries when we mark off the scale on the horizontal axis. As with the Pareto chart, the vertical axis shows the frequencies. Be sure the frequency scale goes high enough to accommodate the class with the greatest frequency but not too much higher, so that you don't have a lot of wasted space.

Here's the histogram for the women's ages:



Once you've decided where to place the 14.5 and the 19.5 on the age axis, all other numbers have their places determined, because you're saying that the distance between those two positions represents five years. So you can't just put the other boundaries any old place. Also, this decision means that the distance from the corner to the 14.5 can't really **be** 14.5 years, so you interrupt the axis and put the two diagonal hash marks to indicate this disruption of the scale. Lightning bolts are also used for this.

Representations of Bivariate Data

Sometimes we want to look at two variables at once, and for these situations we use the word **bivariate**, meaning two variables. Say we're studying the connection between people's ages and the number of pets they have. To do this visually, we make a **scatter plot**, which I'm sure you've done before. Each point in the scatter plot gives us two pieces of data about a single member of the sample, one datum for each variable. We have to specify which variable is represented by the horizontal position of the dot and which by the vertical. The horizontal variable is of course the *x*-variable, and it's sometimes called the **independent** variable (like the terminology we used in describing experimental studies) and the vertical variable, the *y*-variable, is also called the **dependent** variable, though in the case of scatter plots we are not trying to imply by this terminology that the one causes the other.

You have to make two scales, one for each variable, and to use diagonal hash marks or lightning bolts as breaks if appropriate. One problem is what to do if two people have the same values for both variables. You can indicate this simply by making another point close to the first, and so on if there are more than two, or you can count up how many identical ones there are altogether and put the figure in parentheses next to a single dot in the correct position, like this:

• (3)

Activity #3: Making a grouped frequency distribution table

Construct a grouped frequency distribution table for the heights of the men in the Class Data Base, using 61 as the lower limit of the first class and 3 inches as the class width. Have columns for the Class Limits, the Class Boundaries, the Tally, the Frequency, and the Relative Frequency to the nearest hundredth.

For use in Assignment #1:

Assignment #1

1) Construct a categorical frequency distribution for the men's favorite color data in the Class Data Base. Consider anything that contains the word 'orange' to be orange, 'red' to be red, 'blue' to be blue, and 'purple' to be purple. All other colors are separate. Don't make a category for men with no favorite color. Include the name of the class, the tallies, the frequencies, and the relative frequencies to the nearest percent.

2) Construct a grouped frequency distribution table for the men's age data in the Class Data Base. Let the lower limit of the first class be 15 and the class width be 3 years. Include in your table the class limits, class boundaries, tallies, frequencies, and relative frequencies as decimals rounded to the nearest hundredth.

3) Make a Pareto chart for the men's favorite color data. Label each axis. **Must be on graph paper.**

4) Make a pie chart for the men's favorite color data.

5) Make a histogram for the men's age data. Include a label for each axis. **Must be on graph paper.**

6) Using only the men in the Class Data Base, make a scatter plot with shoe size as the *x*-variable and height as the *y*-variable. Label each axis. **Must be on graph paper.**

7) Looking at questions titled Native and Grad (local native, local high school graduate) on the survey, there are four possible outcomes for each person: YY, YN, NY, NN. Figure out a way to represent these categories, and have your representation include how many students are in each of the four categories. **(Many different ways to do this.)**

Lecture #4: Measures of Central Tendency

When you look at a data set, there are three different facts about it that you want to know. First, what is its most typical member? This is called a **measure of central tendency**, and there are several different kinds. This lecture is about these measures. Then there's the question of how spread out or varied the members are. These are called **measures of variation**, and they are the topic of the next lecture. Finally, what positions do the various members of the data set have relative to the most typical member? These are called **measures of position** and will be covered in the lecture after the one about measures of variation.

The kind of measure of central tendency which is appropriate for a certain data set depends on whether the variable is qualitative or quantitative, and, if it's quantitative, what level of measurement the data set has. If a data set is at the nominal level, the only fitting measure is the **mode**, which is the most common value the data set takes on. When I was a child I thought that the word mode meant ice cream, because people would talk about *pie à la mode*, which was pie with ice cream on top, but actually the phrase refers to the popular or stylish way of having pie, so mode means popular.

So in the categorical frequency distribution of men's ZIP codes in the last lecture, the mode was 95482, which was given by 38 of the men in the group. In the grouped frequency distribution of women's ages, the mode was the age class 15-19, which had a frequency of 26. This class is also called the **modal class**, because it is the class with the largest frequency in the distribution. (If two classes are tied for the highest frequency, the distribution has two modes and is called **bimodal**.)

With a quantitative variable at the nominal level of measurement, the mode is as far as you can go in measures of central tendency. But at higher levels, other measures of central tendency can be named. For instance, at the ordinal, interval, and ratio levels, you can talk about the **median**. This is the value of the variable that has half the data set less than or equal to it and half the data set greater than or equal to it. It divides the data set into two parts of equal frequency.

Let's use the first five men's ages as an example. We'll call the age variable *x*. Here's a table with these numbers:

x
22
22
16
26
20

A table like this contains **raw data**, which means it hasn't been sorted in any way. To determine the median, we sort the data, either **ascending**, from lowest to highest, or **descending**, from highest to lowest. When we do this, we're creating what's called an **array**, which is a data set sorted according to order. Let's do ascending:

x
16
20
22
22
26

Now we can see that the third of these numbers, 22, is the middle one, and so it's the median of this very small data set. If the sample size is *n*, then if *n* is an odd number the median occupies the $\frac{n+1}{2}$ position, in this case $\frac{5+1}{2} = 3$, or the 3rd position.

I know what you're thinking: what if *n* is an even number, so the position formula doesn't yield a whole number? Let's try putting in the sixth age, 18, in the list, and then arranging the new set in ascending order:

x
16
18
20
22
22
26

Now there is no exact middle number. The middle two numbers are in the $\frac{n}{2}$ position (3rd in this case, *n* being 6) and the $\frac{n}{2}+1$ position (4th in this case). So here we define the median as being the number halfway between these two values, in other words $\frac{20+22}{2} = 21$. This median has half the data set less than it and half greater than it. So you can see that the median doesn't have to be a member of the data set. If the two middle numbers are not the same, it will be their average.

Different symbols are used for the median. The calculator calls it "med," which we will too.

Back to modes for a minute. The mode of both sets is 22. If the sixth age had been, say, 16 instead of 18, then the second set would have had two modes, 16 and 22. In the first set, if one of the 22-year-olds turned out to be, for instance, 24, the set would have had no mode. Data sets can have no mode, one mode, two modes, etc. Modes are very flexible that way, unlike other measures of central tendency. A data set has one and only one median.

The most commonly used measure of central tendency, and the one you've always heard called the **average**, is what we call in statistics the **mean**. It is appropriate to use when the variable is at the interval or ratio level of measurement. You know how it works: you add up all the numbers and then divide by the number of numbers. But

now we have some fancy notation for it. Using our first data set, we add up the values of *x*:

x
22
22
16
26
20
106

Using our sigma notation, we say that $\sum x = 106$, and since we use *n* for the sample size, dividing by it gives us $\frac{\sum x}{n} = \frac{106}{5} = 21.2$. So the mean of our sample is 21.2, and we have a symbol for the **sample mean** which we'll be using throughout the course: $\bar{x}$, pronounced "*x* bar." To summarize, the formula goes $\bar{x} = \frac{\sum x}{n}$.

When $\bar{x}$ isn't a whole number, like our example, the convention for rounding it is to go one place beyond the data. Since the ages were whole numbers, that means that if the mean hadn't terminated by the tenths place, we would round it to that place.

The thing about the mean is that it's the number that, if all data values were equal to it, the sum of the values, the $\sum x$, would be the same as it actually is. Say you go bowling, and you bowl a 121, a 140, and a 168. Your total for the three games is 429, and your average score is $\frac{429}{3} = 143$. If you'd scored 143 on each of the three games, your total would still be 429.

Likewise, if every age in the data set were 21.2 years, the total would still have been 106.

So the mean is a kind of balancing point. If you take each member of the data set and subtract the mean from it and then add up the differences, you always get 0. Take the bowling example: $121 - 143 = -22$, $140 - 143 = -3$, and $168 - 143 = 25$. Notice that $-22 + (-3) + 25 = 0$. This always works.

The final measure of central tendency, appropriate when the variable is at the interval or ratio level of measurement, is the **midrange**. It's the number halfway between the smallest and the largest value of the variable, hence the name: it's the middle of the range of values. We have names and notations for these concepts. The smallest value is called the **minimum** (min) and the largest the **maximum** (max) so the midrange can be found by this formula:

$$\frac{\min + \max}{2}$$

In other words, the midrange is the mean of the min and the max. Lots of m's here. For our original data set, min = 16 and max = 26, so the midrange is $\frac{16+26}{2} = \frac{42}{2} = 21$. Like the median, the midrange can be a member of the data set, but it doesn't have to be. It might not even be a whole number.

(Here we must make a little detour for an annoying little topic that becomes important later in the course, when we get to inferential statistics. The measures we've talked about so far are **statistics**, numbers which describe a sample, if in fact our little data set **is** a sample. Is it? It is if we say it is. So the mean $\bar{x}$ is actually the **sample mean**. But what if our data set is actually a population? How could it be? Well, it can be if we say it is. And if it is, the 21.2 that we found has a different symbol and a slightly different formula, even though the result is the same. This parameter, the **population mean**, is given the symbol μ, spelled "mu" in English. It's the Greek "m," and a very lovely one it is! Often parameters are symbolized by Greek letters while statistics have English ones. Anyway, the formula for μ, instead of being $\frac{\sum x}{n}$, is actually $\frac{\sum x}{N}$, because the symbol for the **population size** is *N*, as opposed to the **sample size**, *n*. *N* is a parameter, and *n* is a statistic. You probably think I've gone completely mad, but this is how it's done, and just wait till the next lecture where we encounter a number which is actually a different number depending on whether it's a statistic or a parameter! At least $\bar{x}$ and μ were both equal to 21.2.)

So we have found four measures of central tendency for our data set, and there are three different results:

Mode	22
Median	22
Mean	21.2
Midrange	21

Is this a problem? Not really, but you might be asking how you'd know which measure would be the best to use in a certain situation.

There's no one simple answer, but there are guidelines. They have to do with the concept of **resistant** statistics. A resistant statistic is one that is not affected very much by small changes in the data set, particularly adding some very high or very low values.

Let's begin with an example. You're applying for a job at a small company. The boss makes $120K per year, and each of the four employees (one of whom is leaving the firm) makes $20K. The total payroll is $200K, and when the boss tells you that the average (mean, in this case) pay is $40K, you're pretty happy about it. But then you get the job and find you're making only half that much. Wouldn't it have been more informative to be told the median salary, which is $20K, rather than the mean, which is inflated by having that comparatively huge salary thrown in?

So the median is a much more resistant statistic than the mean and should be used whenever there are a few very large or unusually small values of the variable in the data set. That's why you see data on median, not mean, home prices, because a few mansions make the mean frighteningly large. And the mean in this case would not be as informative as the median, because you're probably not looking to buy a mansion anyway. The median is a much better indicator of what's out there.

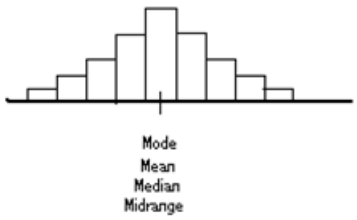
The midrange suffers even more than the mean from the defect of being unduly affected by a high or low value. In its case altering a single datum, if it were the min or the max, would greatly alter the midrange. And the mode – well, nobody wants to use it unless they have to, as in the case of the nominal level of measurement. It just doesn't do much for us mathematically.

The mean is definitely the most useful for us mathematically, as you'll see, so even when it isn't the most appropriate measure of central tendency to use, people tend to use it. This happens even in situations where it's not only inappropriate – it's downright unethical. Take GPA, or grade-point average. An A is 4, a B is 3, etc. You multiply the number of units the course was by the number of the grade you got, add these products up, and divide by the total number of units, rounding to the nearest thousandth. Sounds kosher, but it assumes that there's some meaning to the differences between grades, whereas grades are actually at the ordinal level of measurement, no matter what anybody tries to maintain to the contrary. A's are better than B's, B's than C's, but are they the **same** better? I just don't think you can say that. Finding a GPA (or an average rating when students rate professors on a scale of 1 to 5) to the nearest thousandth makes it seem that the grading process is a lot more precise than it really is. At most you could give the median grade – the number that half the grades are at or above, and half at or below. But you can see the problem with this: it would be impossible to rank students except in very large groups, because the medians would either end at the decimal point or at .5. (I don't expect anybody to wage a campaign to stop the unethical use of the mean, though. People will continue to employ it because it gives the illusion of precision.)

Shapes of Distributions

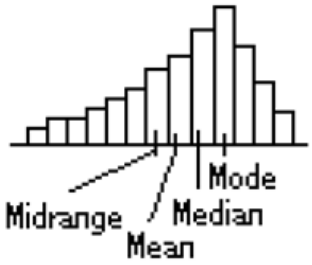
Sometimes it doesn't matter which measure of central tendency you use, because they're all the same. This leads to the idea of the **shape** of a distribution. We're thinking here of the histogram made by a data set of a variable at the interval or ratio level.

Maybe it's **symmetrical** – you could fold it along a vertical line and the two sides would match:



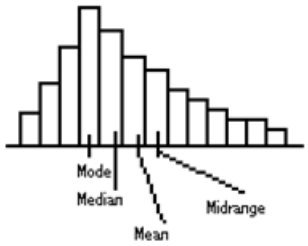
In this case the four measures of central tendency are the same, right in the middle.

But what if this isn't the case? A distribution could be what we call **skewed**: This one is **skewed to the left** (we give the direction of skewing as where the tail goes):



The low values of this distribution pull the mean and the midrange to the left of the median and mode.

Here's a distribution that is **skewed to the right**:



This is what the home prices would look like. The high-priced houses have pulled the mean and the midrange to the right of the median and the mode.

Weighted Means

Here's a final concept for this lecture, and it's one which is very useful in various parts of real life. You already have a feel for it, I'm sure. Say you're taking a course, and you're just about to have the final, which is worth 20% of the grade. So far you've got an A, but you have a terrible feeling about the final. Well, not to worry! Even in the unlikely event you get a 0 on the final, you couldn't possibly get below a C in the course (not what you want, but still, you passed, and it's a respectable grade). And you're not **going** to get a 0. How about a 50? Well, in that case, you'd get a B, not so bad. The thing is that the majority of the grade has already been determined; the instructor isn't going to average your A and whatever you get on the final, because they don't have equal **weight**.

A baseball example: At the beginning of the season, every at-bat affects your batting average tremendously. But as time goes on, it will take a big slump or a long hot spell to have much of an effect on your average. Your average is based on an increasingly large share of the season.

Let's do this problem: The mean height of the men in the Class Data Base (call it $\bar{x}_M$) happens to be 70.189 inches to the nearest thousandth (I know we're supposed to round to the tenths, but I'm using more places to show the precision of our method). The women's is $\bar{x}_F = 64.910$ inches. Is it possible to determine from the figures the mean of the entire sample? Yes, but first here's how **not** to do it:

$\frac{\bar{x}_M+\bar{x}_F}{2} = \frac{70.189+64.910}{2} \approx 67.550$ inches to the nearest thousandth of an inch.
The mean of the entire sample is actually 67.242 inches, so averaging the men's and women's means doesn't work, and that's because there are different numbers of men and women making up the separate means, and we have to **weight** them accordingly.

There are 53 men ($n_M = 53$) and 67 women ($n_F = 67$). First of all, since there are more women, the sample mean will be closer to the women's mean than the men's, and you can see that it is, since 67.242 < 67.550, which was the half-way point.

Remember that the mean of a data set is the value of the variable that, if every member of the set were that value the sum of the values would be the same. In other words, if the 53 men were each 70.189 inches tall, their total height would be the same as the actual total height of the men. Imagine them all standing on top of each other's heads in a tall column. That total height would be $n_M \cdot \bar{x}_M$. Likewise, the total height of the women is $n_F \cdot \bar{x}_F$. Now, put the two columns on top of each other, and you have the total height of the sample: $n_M \cdot \bar{x}_M + n_F \cdot \bar{x}_F$. And the size of the sample? It's $n_M + n_F$.

Putting it all together, the mean of the whole sample (we call it the **weighted** mean because we calculated it from separate means, using their sample sizes, or weights), looks like this:

$$\bar{x} = \frac{n_M \cdot \bar{x}_M + n_F \cdot \bar{x}_F}{n_M + n_F} = \frac{53 \times 70.189 + 67 \times 64.910}{53+67} \approx 67.242$$

And that's just what I know it is from using the Excel software on the Class Data Base, within a thousandth of an inch, which is certainly more than close enough.

Lecture #5: Measures of Variation

Measures of variation are not as familiar as measures of central tendency. We're not looking for the most typical number here; we're looking for a way to describe numerically how **spread out** the members of the data set are.

Of course, if a variable is at the nominal level of measurement, we can't talk about spread at all, so these variables have no measure of variation, but for the other three levels there's the very simplest of the measures, the **range**, which is the spread itself, the difference of the highest and the lowest member of the set, the max minus the min. Here's the data set we used in the last lecture:

x
22
22
16
26
20

The largest age (the max) is 26, and the smallest (the min) is 16. I know that in everyday speech you might say that the range is 16 to 26, but in statistics we use the word "range" to refer to the **difference** itself, 10 years. We can say that the people whose ages are in this data set all have ages within 10 years of each other. The formula for the range is $\max - \min$.

Couldn't be simpler, but the problem is that it's not a resistant statistic as defined in the last lecture. Add one really old person and the range is altered tremendously! So as long as we're dealing with a variable at the interval or ratio level and can calculate the mean of the data set, our measure of variation will be the **standard deviation**. It's called "standard" because it's what people normally use to quantify (make a number to express) variation. Everyone uses it when doing statistics; we'll use it all semester, and pretty soon you'll feel that you've always known about it. But in the meantime it will probably seem very odd to you, if not downright bizarre, and you'll want to know **why** it's done the way it is. Try not to ask. Just accept that it is the **standard** deviation.

Your calculator can supply you with the standard deviation of a data set with no problem, but we're going to build up its definition and calculation in what is called the **table method**. We've already started the table, which is the list of *x* values. Then we must find the mean of the data set, which we did in the last lecture. It's 21.2. We now make another column in the table, headed $x - \bar{x}$. We take each datum and subtract the mean from it:

x	$x-\bar{x}$
22	0.8
22	0.8
16	−5.2
26	4.8
20	−1.2

The $x-\bar{x}$'s are called **deviations from the mean**, because they express how far the datum is from the mean. If the *x* is less than the mean, its deviation from the mean is negative, and if it's greater than the mean, positive. Notice that the $x-\bar{x}$'s add up to 0, which as mentioned in the last lecture is the way it should be. Seeing that they do add to 0 is a good check both on the correctness of $\bar{x}$ and of the $x-\bar{x}$'s.

The next, and final, column of the table comes from taking the $x-\bar{x}$'s and squaring them. In other words, the heading is $(x-\bar{x})^2$. These are called **squared deviations from the mean** for obvious reasons. Remember that squaring a number can never result in a negative, even if your calculator seems to be telling you so. Multiplying two identical positive numbers (which is what squaring is) gives a positive product, and multiplying two identical negative numbers also gives a positive product, and $0\cdot 0=0$. Here's the table now:

x	$x-\bar{x}$	$(x-\bar{x})^2$
22	0.8	0.64
22	0.8	0.64
16	−5.2	27.04
26	4.8	23.04
20	−1.2	1.44

And that's the end of the table, which consists of the three columns. Now we do various things with the third column, $(x-\bar{x})^2$. First we add it up:

$$\sum(x-\bar{x})^2 = 52.8$$

Then we divide this sum by one less than the sample size, $n-1$. Why $n-1$? Well, it turns out that this number is a better fit for the population of which this data set is a sample. (Try not to worry about this.) We label this quotient s^2, and it's called the **sample variance:**

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{52.8}{4} = 13.2$$

So 13.2 is the sample variance. Finally, to find the **sample standard deviation**, labelled *s*, because $s = \sqrt{s^2}$, we take the square root of the sample variance:

$$s = \sqrt{13.2} \approx 3.633,$$

and we round this to the nearest tenth, following the same rule as for the sample mean. So $s \approx 3.6$.

To summarize, the formula for the sample standard deviation is

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

and you should memorize it. Actually you will have memorized it, or at least learned how to do it, by the time you've practiced a few times.

The final measure of variation concerns seeing how the sample standard deviation measures up against the sample mean. Is the standard deviation large or small compared to the mean? We find the ratio of the standard deviation to the mean, and this is called the **coefficient of variation**, or C_{VAR}:

$$C_{VAR} = \frac{s}{\bar{x}} = \frac{3.633}{21.2} \approx 0.171.$$

To the nearest whole percent, this is 17%. Notice that I used more of *s* than just to the nearest tenth in doing the further calculation. That's because once you start rounding and use the rounded numbers to get other numbers you get further and further away from an accurate figure.

The sample variance, the sample standard deviation, and the sample coefficient of variation are all statistics, because they are numbers that describe a sample. The corresponding parameters, describing a population, not only have different symbols but also a slightly different formula. It is annoying but necessary to look at this.

You use the same table, except that the headings are now *x*, $x - \mu$, and $(x - \mu)^2$, because you're talking about a **population** mean, μ, instead of a sample mean, $\bar{x}$, though the numbers are the same. Then you add up the $(x - \mu)^2$'s and get $\sum (x - \mu)^2$ which is also the same as when the data set was considered a sample. Here is the change: instead of dividing by $n-1$, one less than the sample size, you divide by *N*, the **entire** population size. Try not to wonder why. And this new number, the **population variance**, is given the symbol σ^2, that funny letter being the lower-case Greek sigma, like our "s," whose capital form, Σ, we've been using for quite some time now. This gives us

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} = \frac{52.8}{5} = 10.56,$$

and then σ, the **population standard deviation**, is the square root of σ^2:

$$\sigma = \sqrt{10.56} \approx 3.2496 \approx 3.2.$$

To summarize the formula for the population standard deviation, it's

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

As you can see, a data set considered as a sample has a larger standard deviation than when it's considered as a population and we look at its population standard deviation: in this case, 3.6 > 3.2. This happens because the quotient in the sample variance has a denominator one smaller than the denominator in the population variance, and dividing by a smaller number gives a larger quotient.

In practice, a data set will not be looked at as both a sample and a population, and it usually will be seen as being a sample, but we have to introduce these distinctions. And now I can take another shot at explaining why we use $n-1$ as the denominator in the sample variance. If, as you do in inferential statistics, you're using a sample to draw conclusions about the population it's a part of, you'll get a more accurate estimate of the population standard deviation by using the $n-1$. Using the entire sample size will cause you to underestimate the population standard deviation and hence the degree to which numbers in the population are spread out.

And one last thing to mention: the standard deviation is the **square root** of the variance, and the variance is the **square** of the standard deviation. It may seem redundant to have two such closely-related entities, but that's how it is. To find the standard deviation using the table method you have to find the variance first and then take its square root. When your calculator works with a data set it just reports the standard deviation, and if you need to find the variance for any reason you have to square the standard deviation.

Activity #4 & 5: Finding measures of central tendency and variation

Using the first five women's heights in the Class Data Base, find

1. The mode
2. The median
3. The mean
4. The midrange
5. The range
6. The sample variance (to the nearest tenth)
7. The sample standard deviation (to the nearest tenth)
8. The coefficient of variation (to the nearest whole percent)

Use the table method for the sample variance and standard deviation

Assignment #2

1) This small spreadsheet gives data for the five people in the Class Data Base who have six pets:

Shoe Size	Age	Height in Inches	Number of Pets
13	20	71	6
12	18	72	6
8.5	18	68	6
9.5	50	66	6
8.5	20	63	6

For each variable (shoe size, age, height, and number of pets) find the

a) Mode
b) Median
c) Mean
d) Midrange
e) Range
f) Variance (to the nearest tenth)
g) Standard Deviation (to the nearest tenth)
h) Coefficient of Variation (to the nearest whole percent)

Use the table method in finding the variance, and include the tables in your homework paper.

2) Using the notation in Lecture #4, the following table describes the variables in the Class Data Base for shoe size, age, and number of pets:

		Shoe Size	Age	Number of Pets
Men:	$n_M = 53$	$\bar{x}_M = 10.708$	$\bar{x}_M = 21.623$	$\bar{x}_M = 1.830$
Women:	$n_F = 67$	$\bar{x}_F = 7.955$	$\bar{x}_F = 26.060$	$\bar{x}_F = 2.701$

Use the weighted mean formula to find the mean shoe size, age, and number of pets for the entire sample, to the nearest thousandth.

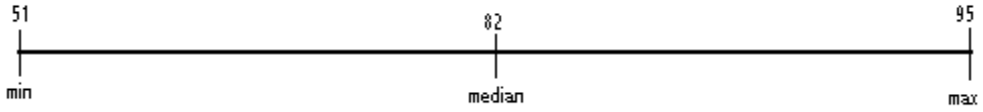
Lecture #6: Measures of Position

In addition to the measures of central tendency and variation which we've already discussed, there is a third kind: **measures of position**. They fall into two categories depending on whether the measure of central tendency being used is the median or the mean.

Median, Quartiles, Deciles, and Percentiles

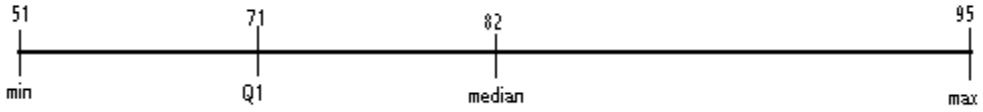
Imagine that you had a data set of 500 scores on a test, and that the smallest one was 51 and the largest was 95. You could write the 500 scores in a line from smallest to largest. Let's say that the median of the scores was 82. So half the scores would be 82 or below, and half would be 82 or above. As you can see, there would be quite a few scores the same in order to get 500 numbers.

The line could be envisioned looking something like this:

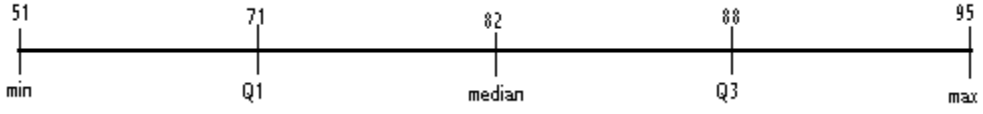


There would be 250 numbers between the min (51) and the median (82), and another 250 numbers between the median and the max (95).

Now look at the lower (left) half of the scores. They would have a median of their own. Let's say it's 71. So half of this lower half would be between 51 and 71, and the same between 71 and 82. Half of a half is a quarter, 125 scores in this case. We have a name for this median of the lower half: the **first quartile**, or Q_1. It's greater than or equal to one quarter of the scores:



We could do the same thing with the upper (right) half of the scores: find its median. Let's say this is 88. Three-quarters of the scores would be less than or equal to 88 – the half from 51 to 82 and the quarter (half of a half) from 82 to 88. So we call the score in this position the **third quartile**, or Q_3. (Here I might add that another name for the median could be Q_2, because it's greater than or equal to two-fourths of the scores.) Here's what we have so far:



These five scores and their positions, min, Q_1, median, Q_3, and max, are called the **five-number summary** of the data set. They divide the set into four groups, or quarters, of equal size.

Can you tell by looking at them that the low scores are much more spread out than the upper ones? It takes from 51 to 71, 20 points, to fit in the 125 lowest scores, but it takes only 11 $(82-71)$ for the next 125, then 6, and then 7. (Remember, I just made up all these numbers.)

In fact, half of all the scores (250 in this set) fall between Q_1 (71) and Q_3 (88). These scores are the middle half of the scores. We have a special name for the range of this middle half of scores: the **interquartile range**, symbolized IQR. Its formula is

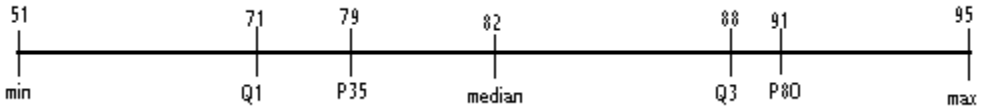
$$IQR = Q_3 - Q_1$$

and in our example it is $88-71=17.$ The middle half of the scores are within 17 points of each other. The interquartile range is a measure of variation: how tightly packed the middle half of the scores are.

Let's go on a bit with our example. Let's say that one-tenth, or 50, of the scores are between 51 and 60. We would call 60 the **first decile**, or D_1, because it's greater than or equal to one-tenth of the scores. Can you see where D_6 would fall? Between the median (which could be thought of as D_5) and Q_3, because $\frac{6}{10}$ is between $\frac{5}{10}$ and $\frac{3}{4}$ (think $0.5 < 0.6 < 0.75$). Deciles aren't used much, but the next "-ile" is.

And that's **percentiles**, which you've no doubt heard of. "Per cent" means "out of one hundred." Imagine the score which is greater than or equal to only one-hundredth of the scores. It's called the **first percentile**, or P_1. So $D_1 = P_{10}$, $Q_1 = P_{25}$, med $= P_{50}$, $Q_3 = P_{75}$, and so on.

Can you see where P_{35} would fall? Between Q_1 and the median. Let's say $P_{35} = 79$. And what about P_{80}? Between Q_3 and the max. Let's say $P_{80} = 91$. I'll put them on our line:



I hope it's obvious to you that this is not an ordinary scale, where distances between numbers have meaning (from 79 to 82 is a longer stretch than between 71 and 79, for instance), but rather the visualization of 500 numbers written in order.

Here are some questions to see if you understand the concepts:

1) How many scores are between 71 and 95? (375, or ¾ of 500)

2) How many scores are at least 79? (325, which can be done two ways: 65% of 500, or 500 minus 35% of 500)

3) How many scores are between 82 and 91? (150, or 30% of 500, 30 because $80 - 50 = 30$)

4) How many scores are between 79 and 91? (225, or 45% of 500, 45 because $80 - 35 = 45$

Outliers

One difference between measures of position and the other measures is that you can talk about the position of a single datum in the data set, but you can't talk about the central tendency or variation of a single datum. Each member of the set can be located, for instance, according to its percentile – what percent of the set consists of numbers less than or equal to it. This is familiar to us from standardized tests and from pediatricians' offices, where parents are told the percentile of their child's height and weight.

The fact that individual data points have position leads to the idea of an **outlier** – a number which is so different in size from the others that you might wonder if it's a mistake or simply unusual. We have a way of deciding whether a number is an outlier or not that utilizes the concept of quartiles.

There are two ways of being an outlier – being very small or being very large. The criterion for the former is that the datum, *x*, must satisfy this condition:

$$x < Q_1 - 1.5(IQR)$$

In other words, the datum must be more than one and a half interquartile ranges below the first quartile. For the heights in the Class Data Base, for which $Q_1 = 64$ and $Q_3 = 70.5$, so that the interquartile range is $70.5 - 64 = 6.5$, this amounts to

$x < 64 - 1.5(6.5)$, or $x < 54.25$ inches.

The corresponding criterion for the large outliers is to go one and a half interquartile ranges above the **third** quartile and consider numbers **larger** than that:

$x > Q_3 + 1.5(IQR)$, or in our case
$x > 70.5 + 1.5(6.5)$, or $x > 80.25$ inches.

The 54.25 is the **lower** limit for outliers, and the 80.25 is the **upper** limit. So a height of 54 inches or less is an outlier on the small end, and 81 inches or more is an outlier on the large end. Some data sets have outliers, and some don't. The height data set had no outliers, since the minimum height was 57 inches and the maximum 77 inches. As mentioned above, outliers might be mistakes or simply unusual values of the variable; it's up to the researcher to investigate which category the outliers fall in.

To use a different category from the Class Data Base, consider the number of pets. Here $Q_1 = 1$ and $Q_3 = 3$, so the interquartile range is $3-1=2$. Small outliers would be numbers of pets, *x*, for which

$$x < Q_1 - 1.5(IQR) = 1 - 1.5(2) = -2,$$

certainly a very unlikely, actually impossible, number of pets to have.

Unusually large numbers of pets would be those for which

$$x > Q_3 + 1.5(IQR) = 3 + 1.5(2) = 6,$$

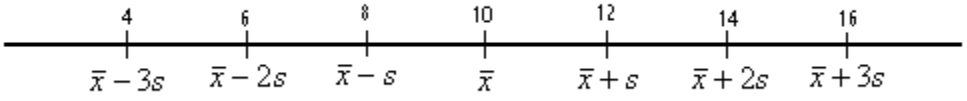
in other words 7 or more pets, unusual maybe, but not impossible. And in fact our set has six outliers, 10, 10, 9, 7, 7, and 7, as can best be seen by sorting the data from highest to lowest.

So data sets can have no outliers at all, or a couple, or a few, but not too many or they wouldn't be unusual. Also, if the data are generally widely spread, this will make the IQR large, which will make the lower limit small and the upper limit large, and these limits will be hard to get by.

Z-scores or Standard Scores

So far we've talked about measures of position that depend on the median and, by extension, the quartiles and percentiles. Now we'll consider a measure of position determined by the mean, and its little friend the standard deviation.

The basic approach is to make a scale in which the mean is at the middle and units are marked off using the standard deviation. Say what? Well, let's take an example with simple numbers, say $\bar{x} = 10$ and $s = 2$. Then $\bar{x} + s = 10 + 2 = 12$, and we say that 12 is **one standard deviation above the mean**. Likewise $\bar{x} - s = 10 - 2 = 8$, so 8 is **one standard deviation below the mean**. And we can go on like that. Here's a labeled number line:

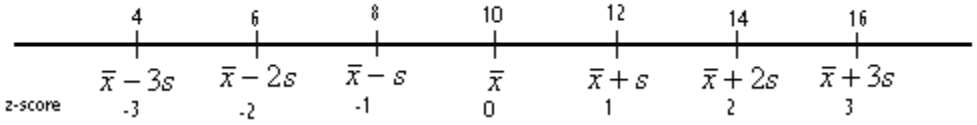


This is totally different from the example in which we pretended we had written 500 numbers in order from smallest to largest. There, the space between numbers on the line represented frequency – how many data points fell between two values – but here the space between numbers is a true scale, with the units being the standard deviation of the set.

We can talk about the **interval of data within one standard deviation of the mean**. This is, mathematically speaking, the interval $(\bar{x}-s,\ \bar{x}+s)$, using interval notation which you may remember from algebra. In the example above, this amounts to the interval $(8,12)$; in other words all the numbers between 8 and 12 are **within one standard deviation of the mean**. I know that in algebra we make a big distinction between **open** and **closed** intervals, ones that don't contain the endpoints and ones that do, and that we use square brackets $[\ ]$ for the latter, but here we're just going to use parentheses and kind of gloss over the distinction.

Intervals of data within two standard deviations of the mean look like $(\bar{x}-2s,\ \bar{x}+2s)$, in our example $(6,14)$. **Intervals of data within three standard deviations of the mean** are $(\bar{x}-3s,\ \bar{x}+3s)$, or $(4,16)$. Numbers between 4 and 16 are said to be **within three standard deviations of the mean**.

Each member of the data set has associated with it something called a ***z*-score** (you'll see what the *z* means later) or a **standard score**. It tells how many standard deviations the datum is from the mean. Here I've put the *z*-scores beneath the line:



All members of the data set have a *z*-score, and here's the formula for a datum, *x*:

$$z=\frac{x-\bar{x}}{s}$$

See how it works, say for 14:

$$z=\frac{14-10}{2}=2\,.$$

If a datum is **smaller than** the mean, its *z*-score is negative; if it's **bigger than** the mean, its *z*-score is positive, and if it's **equal to** the mean, its *z*-score is equal to 0, because $z=\frac{\bar{x}-\bar{x}}{s}=\frac{0}{s}=0$. Just be sure that you enclose the numerator in parentheses if you're calculating the *z*-score on your calculator, because you want to subtract **before** you divide. If you try $z=\frac{14-10}{2}$ and get 9, it's because you didn't do that; the calculator

thought you wanted to divide 10 by 2, which gives 5, and then subtract the 5 from 14 (because that's what you told it to do).

Just like giving the percentile of a member of a data set is a way of measuring its position, giving the *z*-score is another way of doing so. Let's look at the height data set. $\bar{x} = 67.242$ inches, and $s = 3.998$ inches. Again, I'm using more decimal places of these statistics than just one, and that's because I'm using them in further calculations and want to be as accurate as possible.

What's the *z*-score of a person who is 72 inches tall?

$$z = \frac{x - \bar{x}}{s} = \frac{72 - 67.242}{3.998} \approx 1.190 \approx 1.19$$

We rounded the *z*-score to the nearest hundredth because that's how it's done – a hangover from the olden days when we had to use tables of *z*-scores (don't worry about why) which listed the *z*-scores that way.

At any rate, a person whose height is 72 inches has a height which is **1.19 standard deviations above the mean**. How about a person who is 63 inches tall?

$$z = \frac{x - \bar{x}}{s} = \frac{63 - 67.242}{3.998} \approx -1.061 \approx -1.06$$

This person's height is **1.06 standard deviations below the mean**. The *z*-score came out negative because the person's height is shorter than the mean height.

Outliers Revisited

Can we define outliers in terms of their *z*-scores, in other words use the mean and standard deviation to set up some formula the way we did with Q_1 and Q_3? I'd say that anything with a *z*-score of less than -3 or more than 3 is an outlier, but this isn't a hard-and-fast rule. There's something called Chebychev's Theorem, named after the Russian mathematician with the marvelous first name of Pafnuty (1821-1894), which sets limits on what fraction of a data set can lie outside an interval of data within a certain number of standard deviations of the mean, but his result is weak compared to what happens when a data set is **normally distributed** (you may know this as "bell-shaped"), which we'll spend a huge amount of time on later in this course. In that case, only three-thousandths of the data points can have a *z*-score of less than -3 or more than 3, so I'd certainly call those numbers outliers, but maybe that's too strict a definition. Take it down to a *z*-score of less than -2 or more than 2 and you're up to 4.6% of the data points outside this interval. Maybe that's too inclusive. Enough of this for now.

Activity #6: Outliers, intervals of data, and *z*-scores

Using the age data from the Class Data Base,

1) Use the formula involving Q_1 and Q_3 to find the limits for outliers. How many outliers does the data set have?

2) State to the nearest tenth the interval of data within 3 standard deviations of the mean.

3) To the nearest hundredth find the *z*-scores for these ages. Use values for $\bar{x}$ and s correct to the nearest thousandth in making your calculations, or use the Variables function on your calculator.

    a) 18 years old

    b) 29 years old

    c) 43 years old

Exam #1 – Descriptive Statistics – Sample

For a study of climates, 200 days in September during the past century are selected at random and the high temperature in Ukiah for the day is recorded. For these days, $\bar{x} = 81.9^\circ$ F and $s = 7.4^\circ$ F. The lowest high temperature was 52° F and the highest was 107° F. (Use in Problems #1-5.)

1) Write a paragraph with complete sentences correctly using each of the following words that can be used to describe the situation. Underline the words you use from this list

Experimental Study
Observational Study
Population
Sample
Parameter
Statistic
Quantitative Variable
Qualitative Variable
Discrete Variable
Continuous Variable
Nominal Level of Measurement
Ordinal Level of Measurement
Interval Level of Measurement
Ratio Level of Measurement
Measure of Central Tendency
Measure of Variation
Measure of Position
Mean
Median
Mode
Midrange
Range
Standard Deviation
Variance
Interquartile Range
Coefficient of Variation

2) Find the z-score to the nearest hundredth of a temperature of 63°.

3) Find the interval of temperatures that are within two standard deviations of the mean.

4) If $Q_1 = 75^\circ$, on how many of the days was the high temperature between 75° and 107°?

5) If $P_{15} = 64^\circ$, on how many of the days did the temperature rise to 64° or above?

6) For the data set below, the ages in years of five homes sold during one month in a certain city, find the sample standard deviation to the nearest tenth **using the table method**.

34
11
29
6
10

7) For the data set below, the ages of homes sold during one month in a different city, **use your calculator** to find

a. the sample mean, and
b. the sample standard deviation
c. the median
d. the interquartile range
e. the sample variance
f. the midrange, and
g. the sample coefficient of variation

Round (a) through (f) to the nearest tenth if necessary, and round (g) to the nearest hundredth.

38
40
4
25
29
33

8) Using the data set in Problem #7, find the limits for the outliers and list all outliers.

9) If a group consists of 18 part-time students, with an average GPA of 2.948, and 47 full-time students, with an average GPA of 3.445, what is the average GPA of the whole group to the nearest thousandth?

Exam #1 – Descriptive Statistics – Sample – Solutions

1) An observational study is conducted in which a sample of 200 high temperatures in Ukiah during September is selected from the population of high temperatures in Ukiah during September. Temperature is a quantitative, continuous variable at the interval level of measurement. Statistics reported include the mean, which is 81.9 º F and the standard deviation, which is 7.4 º F. The variance is 54.76, the square of 7.4, and the coefficient of variation is 9%, obtained by dividing 7.4 by 81.9. The midrange is 79.5º F, obtained by taking half the sum of the minimum, 52º F, and the maximum, 107º F. The range is 55º F, the difference of the maximum and the minimum. The mean and the midrange are measures of central tendency. The standard deviation and the range are measures of variation. The minimum and the maximum are measures of position.

2) $z = \dfrac{x - \bar{x}}{s} = \dfrac{63 - 81.9}{7.4} \approx -2.554$. To the nearest hundredth: -2.55

3) $(\bar{x} - 2s,\ \bar{x} + 2s) = (81.9 - 2(7.4),\ 81.9 + 2(7.4)) = (67.1, 96.7)$. $(67.1^\circ F, 96.7^\circ F)$

4) One-fourth of the temperatures are **less** than or equal to Q_1, so three-fourths are between Q_1 and the max (between 75º and 107º. ¾ of 200 is **150**.

5) Fifteen percent of the temperatures are less than or equal to P_{15}, so 85% are equal to or greater than P_{15} (64º). 85% of 200 is **170**.

6)

x	$x - \bar{x}$	$(x - \bar{x})^2$
34	16	256
11	-7	49
29	11	121
6	-12	144
10	-8	64

$\bar{x} = \dfrac{90}{5} = 18$

$s = \sqrt{\dfrac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\dfrac{634}{5 - 1}} = \sqrt{\dfrac{634}{4}} \approx 12.5897$. To the nearest tenth, **12.6**.

7) a) 28.2

b) 13.1

c) 31

d) $38-25=13$

e) $(13.0754)^2 \approx 171.0$

f) $\frac{4+40}{2}=22$

g) $\frac{13.0754}{28.1667} \approx 0.46 = 46\%$

8) Limits on outliers: $x < Q_1 - 1.5(IQR)$, or $x < 25-1.5(13)=5.5$

$x > Q_3 + 1.5(IQR)$, or $x > 38+1.5(13)=57.5$

4 is an outlier.

9) $\frac{(18\times 2.948+47\times 3.445)}{(18+47)} \approx 3.307$

Lecture #7: Probability: Sample Spaces and Contingency Tables

The way we approach probability was thought up by an Italian mathematician named Gerolamo Cardano (1501-1576), who invented it to help himself make a living by gambling. His theory was further developed by the French mathematicians Pierre de Fermat (1601 or a few years later-1665) and Blaise Pascal (1623-1662). They figured out a way to divide the stakes fairly in a game of chance which is interrupted before it can be finished. I'm telling you this because it puts a human face on the whole subject.

The first idea is that of a **probability experiment**, which is a general term for something you do that results in a certain **outcome.** It's easier to understand with an example: You flip a coin. It lands with heads or tails up (we don't consider it ever landing on an edge). It's one or the other, each time you flip it. Landing with heads up is an **outcome**, and so is landing with tails up. Let's label these outcomes H and T. Then the set {H, T} is called the **sample space**, *S*, of the experiment. The number of outcomes in the sample space is its **size**, and our notation for this is the function notation $n(S)$, pronounced "n of S." So *S* = {H, T} and $n(S) = 2$ in this experiment. Not so hard.

How about if your experiment is rolling a die (the singular form of "dice")? Then there are six possible outcomes – the 1 is facing up, the 2 is facing up, etc. – and we can write *S* = {1, 2, 3, 4, 5, 6}, and $n(S) = 6$.

What about picking a card from a standard 52-card deck? Well, here you have to specify what the experiment **is**. Are you looking at which suit you got? If so the sample space is *S* = {Club, Diamond, Spade, Heart} and $n(S) = 4$. If you're looking at what rank you got, *S* = {2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, King, Ace} and $n(S) = 13$. If you're looking at both aspects, suit **and** rank, then *S* = {2 of Clubs, 3 of Clubs, … , King of Hearts, Ace of Hearts} and $n(S) = 52$. Or you could be looking to see if the card is red or black, in which case *S* = {Red, Black} and $n(S) = 2$. Or you could be looking to see if you got a face card (jack, queen, or king) or not, and *S* = {Face Card, Not Face Card} and $n(S) = 2$. We could probably go on like this for quite a while, but you get the idea. The sample space and its size depend on what aspect of the card you're focusing on.

One further concept is called an **event**. An event, *E*, is some set or collection of the outcomes in a sample space (or maybe none of the outcomes). For example, with rolling a die, we could talk about the event "rolling an even number." Then *E* = {2, 4, 6}. We call its size $n(E)$. In this case, $n(E) = 3$. Here's another event: "rolling at least a 5." Here *E* = {5, 6} and $n(E) = 2$. The event could encompass the entire sample space – "rolling less than a 7," which is *E* = {1, 2, 3, 4, 5, 6} and $n(E) = 6$, or could have absolutely no outcomes – "rolling more than a 6," which is *E* = { }, or ϕ, which is the mathematical symbol for the **empty set**, and $n(E) = 0$. Anything you can describe could

be an event, or you could just list the outcomes in the event without describing what they have in common.

Now we'll move on to the probability experiment of flipping **two** coins and seeing what sides are face up. We have to have some way of distinguishing the two coins. Maybe one is a dime and the other is a quarter. Or one could be gold and the other could be silver. Or one is named Coin #1 and the other Coin #2. When I refer to the outcome HH, I mean both were heads. When I mention HT, I mean that the first coin came up heads and the other one tails, but when I write TH, I mean that the first coin came up tails and the other one heads. This is an important distinction to remember.

So what's the sample space? S = {HH, HT, TH, TT}, and $n(S) = 4$. We could talk about the event "getting **exactly** two heads." Let's give this event the notation $2H$. Then $2H$ = {HH}, and $n(2H) = 1$. Using similar notation, we get $1H$ = {HT, TH}, and $n(1H) = 2$. And, of course, $0H$ = {TT}, and $n(0H) = 1$.

How about flipping three coins? Well, the sample space is S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}, so $n(S) = 8$. (If you're seeing a pattern of the sizes of the sample spaces here, good for you! Yes, for one coin it's $2^1 = 2$, for two coins $2^2 = 4$, and for three coins $2^3 = 8$.) Here are the events of getting various numbers of heads, and their sizes:

$3H$ = {HHH}, $n(3H) = 1$
$2H$ = {HHT, HTH, THH}, $n(2H) = 3$
$1H$ = {HTT, THT, TTH}, $n(1H) = 3$
$0H$ = {TTT}, $n(0H) = 1$

Having set up the structure of the sample space, outcomes, and events, we can finally get to probability. And this is **classical probability**, which simply means that **all outcomes are equally likely to occur each time you perform the experiment**. The coins are **fair**, which means that they're equally likely to land with the heads up or the tails up each time they're flipped; the dice are fair, which means….well, you get the idea. It doesn't mean that on every two flips you'll get one head and one tail, or that every six rolls of a die will yield a 1, a 2, etc. It just means that in the long run the outcomes will even out in frequency. (Much later in the course we'll talk about how you can show that certain dice **aren't** fair, or if you can't show that then you'll conclude that maybe they **are** fair.)

We talk about the **probability** of an event, E, occurring. This is a number between 0 and 1 – always! We write $P(E)$, pronounced "P of E," in the manner of function notation. So we can write

$$0 \le P(E) \le 1$$

to express the fact that the probability of an event is between 0 and 1. It's 0 if the event is impossible (like rolling a 7 with one die), it's 1 if it's a sure thing, and if it might or might not happen (what we call a **conditional** event), it's bigger than 0 and less than 1. And the closer the probability is to 1 the more confident we are of it happening.

So here's the formula for determining the probability of an event *E*:

$$P(E) = \frac{n(E)}{n(S)}.$$

In other words, the probability is the fraction of the outcomes in the sample space that are also in the event.

Let's go back to rolling the die. The probability of rolling a 4 is $\frac{1}{6}$, because there's one outcome in the event "rolling a 4" and 6 in the sample space of "rolling a die." The probability of rolling at least a 5 is $\frac{2}{6}$. The probability of rolling an even number is $\frac{3}{6}$.

Let's look at the coin flipping. Refer to the sample spaces and their sizes, and to the events of getting a certain number of heads and their sizes to get these results:
One coin:

$$P(1H) = \frac{1}{2}$$

$$P(0H) = \frac{1}{2}$$

Two coins:

$$P(2H) = \frac{1}{4}$$

$$P(1H) = \frac{2}{4} = \frac{1}{2}$$

$$P(0H) = \frac{1}{4}$$

Three coins:

$$P(3H) = \frac{1}{8}$$

$$P(2H) = \frac{3}{8}$$

$$P(1H) = \frac{3}{8}$$

$$P(0H) = \frac{1}{8}$$

The numerators of these probabilities form some amazing patterns, which you may be familiar with if you know about Pascal's Triangle.

Empirical Probability and Contingency Tables

Think about how the probability formula is dependent on the notion of fairness, or equally-likely outcomes. If a coin weren't fair, we couldn't say that $P(1H) = \frac{1}{2}$. We would have no idea what the probability of getting a head would be, except by flipping the coin many, many times and seeing what fraction of the flips come up heads.

In other words, we'd have to observe what actually happens. This is the basis for **empirical probability**, in which probabilities are determined not by sample spaces with equally-likely outcomes but by observing the fraction of times the probability experiment (like flipping the coin) comes out one way or another. The word "empirical" comes from a Greek word meaning experienced. Empirical probability is the result of experience, not deduction.

This might lead you wonder how we can call a coin fair if not by using empirical probability, by flipping it lots of times. It would be possible to make a coin without any markings, except maybe for labels of heads and tails written in very light-weight ink, and with uniform density and symmetry, and then we could assume that it is fair because of the physics of flipping it. (Although I once saw a remarkable video in which a person was able to flip a supposedly fair coin in such a way that it always came up heads – but he always started flipping with the coin in the same position, and he was able to impart the same momentum each time, so I don't think this is really a counterexample.)

Later in the course we'll see how we can conclude that certain processes (rolls of dice is what we'll use) are most likely fair or not using empirical probability, but for now we'll just say that the assumption of fairness is an abstract one that enables us to develop the classical theory of probability using equally-likely outcomes.

Now we're going to look at one example of empirical probability, and we're just going to make a start on it. It's called a **contingency table**, and it's a way of tabulating two pieces of data for each subject in the sample, where both variables are categorical. (If you're tabulating two pieces of data in which the variables are quantitative, you use a scatter plot.) For instance, looking at the Class Data Base we might wonder whether there are differences between men and women in their attitudes toward social media. You might get some idea of this by just looking at the numbers listed for the men and for the women, but it would be hard to draw conclusions from just the lists. So we categorize each person in the data base by their sex and by their attitude, and we present the results in a neat table.

Immediately we encounter a problem and need to establish **protocols**, sets of rules to determine our classification system. We could have five categories of attitudes, but I think it makes more sense to classify them as Negative, Neutral, and Positive. A rating of 1 or 2 will count as Negative, 3 will be Neutral, and 4 and 5 will be Positive.

		Attitude toward Social Media		
		Negative	**Neutral**	**Positive**
Sex	**Male**			
	Female			

Now we're ready to put numbers in the six cells in the table. We could start with tally marks, looking at each line in the Class Data Base and putting a tally mark in the correct box. For instance, Person #1, being a male whose attitude was a 2, would be tallied in the upper left box. When we finish, the table would look like this:

		Attitude toward Social Media		
		Negative	**Neutral**	**Positive**
Sex	**Male**	12	18	23
	Female	5	14	48

That's all we'll do with the contingency table this time, just tabulate it. I hope you agree that it makes the patterns of attitudes toward social media of the two sexes a lot more apparent. I see some differences in the patterns. This is just the beginning of the kind of analysis we'll be doing.

Activity #7: Sample space and probabilities for tossing four coins; making a contingency table

1) Write the sample space for flipping four coins; then find $P(4H)$, $P(3H)$, $P(2H)$, $P(1H)$, $P(0H)$. Remember that $n(S) = 2^4 = 16$, so there should be 16 outcomes in the sample space.

2) Make a contingency table with the following categories:

		Number of Pets			
		No Pets	**1 or 2 Pets**	**3 or 4 Pets**	**At Least 5 Pets**
Sex	**Male**				
	Female				

Lecture #8: Rules of Probability

Before we tackle the rules of probability, let's look at how we apply the concept of probability to an empirical probability situation like the contingency table we developed in the last lecture.

To begin with we add up the frequencies in the rows of cells to get the **row totals** (for instance, $12 + 18 + 23 = 53$) and the frequencies in the columns of cells to get the **column totals** (for instance, $12 + 5 = 17$), and then add the row totals – or the column totals, because you'd better get the same result – to get the **grand total**, which is 120.

		Attitude toward Social Media			
		Negative	**Neutral**	**Positive**	**Sum**
Sex	**Male**	12	18	23	**53**
	Female	5	14	48	**67**
	Sum	**17**	**32**	**71**	**120**

The calculation of probabilities is based on the situation that one person out of the 120 in the group is selected at random. The probability that the person selected has whatever characteristics are specified is given by the formula $\frac{f}{n}$, where *f* is the frequency or number of people having the characteristics and *n* is, as usual, the size of the sample, or in this case the grand total 120. So the probability that a randomly selected person is male is $\frac{53}{120}$. We could use the notation P(Male) = $\frac{53}{120}$. What's the probability that the person selected has a negative attitude toward social media? P(Negative) = $\frac{17}{120}$. A neutral attitude? P(Neutral) = $\frac{32}{120}$. (Let's leave these fractions unreduced; I'm sure you won't mind.)

What about the probability that the person selected has an attitude which is **not** positive? I'm sure you can see that there are two ways to go about finding it: either total the frequencies for the negative and neutral attitudes to use as *f*, or subtract the frequency for the positive attitude from the grand total. Either way, P(Not Positive) = $\frac{49}{120}$.

Events like "having a positive attitude" and "having an attitude which isn't positive" are called **complementary events**, because they **complement**, or complete, each other. Watch out for the spelling – this is not the word for saying nice things about someone or something. That word is spelled "compl**i**mentary." Every member of the group who is not represented in one of the events is represented in the other; there are

only two groups. If we call one of the events E, then its complement is represented by $\overline{E}$, pronounced "the complement of E."

There's a rule about E and $\overline{E}$. As you can see, P(Positive) + P(Not Positive) = $\frac{71}{120} + \frac{49}{120} = \frac{120}{120} = 1$. (You can see the advantage of not reducing the fractions.) So the general rule is P(E) + P($\overline{E}$) = 1, or P($\overline{E}$) = $1 - $P($E$), which was the second way we used to calculate P(Not Positive).

One more rule which I mentioned last time is that probabilities are always numbers between 0 and 1, inclusive (inclusive means that they **could** equal 0 and they **could** equal 1). You can see why that's true in the empirical case where the probability of an event equals $\frac{f}{n}$ -- f is a number between 0 and n, since f is how many of the outcomes of the sample of size n are in the event. (In classical probability, the probability of an event was $\frac{n(E)}{n(S)}$, so the same reasoning applied.)

And, Or, and Given

Now we're going to examine closely three very common little words which have precise meanings in probability.

The first is **and**. You probably think you know what it means. If I say I'm going out to dinner **and** to a movie, the only way this statement can be true is if I do go out to dinner and I do go to a movie. If I go out to dinner but don't go to a movie, or if I don't go out to dinner but I do go to a movie, or if I don't go out to dinner and I don't go out to a movie, it's not true.

To belabor the point a little, label the statement "I'm going out to dinner" A, and the statement "I'm going out to a movie" B. Then we can make a truth table:

A	B	A and B
T	T	T
T	F	F
F	T	F
F	F	F

I think its meaning is pretty obvious.

Referring now to the contingency table, if we want to know the probability that the person we select randomly is both male **and** had a positive attitude, we see that there

are 23 people in the upper-right cell, so P(Male **and** Positive) = $\frac{23}{120}$. It should be pretty clear that P(Female **and** Neutral) = $\frac{14}{120}$.

You probably could have gotten these without all the discussion of dining out and movies and truth tables, but when we get to the next little word, **or**, it turns out to be more complicated, and the discussion might help. That's because there are two ways in which **or** is used, and in statistics we use it in the less common way. If I say I'm going out to dinner **or** to a movie, you might be surprised if it turns out that I went out to dinner **and** to a movie, because you're used to what's called the **exclusive or**, or the **either/or**. This is where you can do one or the other but not both. Its truth table looks like this:

A	B	A or B
T	T	F
T	F	T
F	T	T
F	F	F

But we're going to use **or** in the **inclusive** sense, the **and/or**. (You may know about this distinction from computer science.) Here it's okay to do both, and the only way the statement would be false if I did neither:

A	B	A or B
T	T	T
T	F	T
F	T	T
F	F	F

Back to the contingency table. What is the probability that the person we select randomly is male **or** had a positive attitude? The group in question would include the 12 males who had a positive attitude, the 18 males who were neutral, the 23 males who had a negative attitude, and the 48 females who had a positive attitude: $12 + 18 + 23 + 48 =$ so P(Male **or** Positive) = $\frac{101}{120}$.

But aren't you tempted just to add the row total of the males, 53, and the column total of those with a positive attitude, 71, and get $\frac{124}{120}$? What's wrong with doing that? Well, you've counted some people twice, namely those who qualify both because they're male and because they had a positive attitude. This group got double-counted. We could correct this by subtracting 23 from our *f* to eliminate the double-counting, and since $124 - 23 = 101$, our answer would now match the one we got by adding up the frequencies of the four cells.

There's a formula for that, again using A and B as the events:

P(A **or** B) = P(A) + P(B) – P(A **and** B)

In our example that comes to

P(Male **or** Positive) = P(Male) + P(Positive) – P(Male **and** Positive)

$$= \frac{53}{120} + \frac{71}{120} - \frac{23}{120} = \frac{101}{120}$$

You can do these **or** problems either way, by adding up the frequencies of the relevant cells or by adding up the row and column totals and then subtracting the frequency of the cell that was counted twice.

By the way, the formula P(A **or** B) = P(A) + P(B) – P(A **and** B) works for both empirical and classical probability.

How about this one: What is the probability that the person we select randomly had a positive attitude **or** a negative attitude? I think you know intuitively that you'd add the column total for Positive, 71, and the column total for Negative, 17, and get an answer of $\frac{88}{120}$. But if you wanted to use the formula for P(A **or** B), what happened to the part where you subtract P(A **and** B)? I don't think it's a mystery, because since nobody was allowed to pick more than one number for his/her attitude, the probability of a single person having both a positive attitude **and** a negative attitude is 0, and you would just have been subtracting 0, which makes no difference.

We have a name for events whose **joint probability** (the **and** probability) is 0: They're called **mutually exclusive**. If one happens, the other can't, and *vice versa*. Neither can happen, but both can't.

Now for **given**. This word is used in what's called **conditional probability**, which we're just going to touch on. If I ask what the probability is that the person we select randomly had a positive attitude **given** that the person is female, I'm giving some information about which person was selected by the phrase that comes **after** the **given**: it was a woman. So the *n* in the formula $\frac{f}{n}$ is no longer 120, but rather 67, the number of women. We were able to reduce the denominator to the frequency of the group following the **given**, because of our extra knowledge. So

P(Positive **given** Female) = $\frac{48}{67}$.

The numerator refers to those women who had a positive attitude, but now they're being viewed as part of the group consisting solely of women, because we know that's where they're being chosen from.

Compare this to what happens if we reverse what comes before and after the **given**. What is the probability that the person we select randomly is female **given** that the person had a positive attitude? Now we know we picked one of the people with a positive attitude, so the denominator is 71. The numerator remains the same, because it's the same 48 women who had a positive attitude. So

P(Female **given** Positive) = $\frac{48}{71}$.

Just remember that the denominator is given by the row or column total of the characteristic that comes **after** the **given** and you can't go wrong. Here's a general formula):

P(A **given** B) = P(A **and** B) / P(B).

This formula also works for both empirical and classical probability.

What's the probability that the person we select randomly is male, **given** that the person had a negative attitude? P(Male **given** Negative) = $\frac{12}{17}$. That the person we select randomly had a negative attitude, **given** that the person is male? P(Negative **given** Male) = $\frac{12}{53}$.

Independent Events

Since we have a formula for P(A **or** B) and for P(A **given** B), you might wonder whether there's one for P(A **and** B). It turns out that there is such a formula, and it is P(A **and** B) = P(A) x P(B), but it works only if the events are what are called **independent events**, which roughly means that whether one happens has no influence at all on whether the other does or does not happen. Mutually exclusive events, for instance, are never independent, because if one happens the other can't, a rather extreme form of influence.

Probability Rules for Classical Probability

Now we're going to return to dice rolling. You may have been surprised to find that the sums from 2 to 12 occurred with such different frequencies, that in fact you most likely ended your rolling because you ran out of room for 6's, 7's or 8's.

Let's see why that is. Here's the chart for rolling two dice:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	2,6
	3	3,1	3,2	3,3	3,4	3,5	3,6
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

Of the 36 outcomes, how many result in a sum of 2? Only one, the 1,1. Same for 12, 6,6. So assuming the dice are **fair**, which means that all outcomes are equally likely to occur (the premise of classical probability), P(2) = P(12) = $\frac{1}{36}$. Dice are arranged such that the opposite sides add up to 7: the 1 and 6 are on opposite sides, as are the 2 and 5, and the 3 and 4. So getting 1,1 facing up means you've got 6,6 facing down. No wonder P(1) = P(12).

Here's the dice table with the outcomes replaced by the sums of the two dice:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

From this you can see that P(3) = P(11) = $\frac{2}{36}$, P(4) = P(10) = $\frac{3}{36}$, P(5) = P(9) = $\frac{4}{36}$,P(6) = P(8) = $\frac{5}{36}$, and, finally, standing alone on the **minor diagonal** (upper right to lower left), P(7) = $\frac{6}{36}$. Notice that the numerators in these probabilities increase by one as the diagonals get longer from 2 to 7 and then decrease by one as the diagonals get shorter from 7 to 12.

You can see why 7 should be the most common sum: more outcomes yield a sum of 7 than any other sum. But 6 and 8 are not far behind.

(Let me just explain why I think it unlikely that the dice you used **are** fair. The pips – or dots – are really little craters; these are dollar-store dice. In the dice used in

casinos, the craters are back-filled with a material of a different color but the same density as the rest of the die. But in these cheap ones, the sides with more pips are less dense and therefore lighter than the sides with fewer pips. So the force of gravity acts unevenly on the die and makes it more likely to end up with a larger number facing up and a smaller one facing down – more 6's up than 1's, to a lesser extent more 5's than 2's, and to an even lesser extent more 4's than 3's. But later in the course we'll see if the craters make enough of a difference to justify calling the dice unfair.)

Now we'll apply the rules of probability governing **and** and **or** to the dice situation. Let's start with the probability that you roll a 6 **and** a 7 (on one roll of two dice). Obviously that's impossible – two numbers can't simultaneously add up to 6 and to 7. So P(6 **and** 7) = 0. How about the probability of rolling a 6 **or** a 7? Let's use the formula, calling rolling a 6 event A and rolling a 7 event B:

P(A **or** B) = P(A) + P(B) – P(A **and** B):

P(6 **or** 7) = P(6) + P(7) – P(6 **and** 7) =

$$\frac{5}{36} + \frac{6}{36} - 0 = \frac{11}{36}.$$

A simpler method is just to count how many outcomes have sums of 6 or sums of 7:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	1,4	**1,5**	**1,6**
	2	2,1	2,2	2,3	**2,4**	**2,5**	2,6
	3	3,1	3,2	**3,3**	**3,4**	3,5	3,6
	4	4,1	**4,2**	**4,3**	4,4	4,5	4,6
	5	**5,1**	**5,2**	5,3	5,4	5,5	5,6
	6	**6,1**	6,2	6,3	6,4	6,5	6,6

What's the probability that you'll roll at least an 8? Well, that would mean rolling an 8 or 9 or 10 or 11 or 12. I've put those outcomes in bold type:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	**2,6**
	3	3,1	3,2	3,3	3,4	**3,5**	**3,6**
	4	4,1	4,2	4,3	**4,4**	**4,5**	**4,6**
	5	5,1	5,2	**5,3**	**5,4**	**5,5**	**5,6**
	6	6,1	**6,2**	**6,3**	**6,4**	**6,5**	**6,6**

There are 15 of them, so P(At least 8) = $\frac{15}{36}$.

What's the probability of at most a 4? That would be a 2, 3, or 4:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	**1,1**	**1,2**	**1,3**	1,4	1,5	1,6
	2	**2,1**	**2,2**	2,3	2,4	2,5	2,6
	3	**3,1**	3,2	3,3	3,4	3,5	3,6
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

P(At most 4) = $\frac{6}{36}$.

At least and **at most** are quantitative phrases that can give people a lot of trouble, so make sure this doesn't happen to you. Think of practical situations: You must be **at least** 21 to buy alcohol; you must be **at most** 5 feet tall to go on this ride.

Now let's talk about a different kind of use of the outcomes of the dice rolls. Instead of caring about the sum of the dice, we're going to care about how the first die or the second die landed. What's the probability that the first die lands on 3?

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	2,6
	3	**3,1**	**3,2**	**3,3**	**3,4**	**3,5**	**3,6**
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

P(3 on 1st die) = $\frac{6}{36} = \frac{1}{6}$, which is just what you'd expect.

What's the probability that the second die lands on 4?

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	**1,4**	1,5	1,6
	2	2,1	2,2	2,3	**2,4**	2,5	2,6
	3	3,1	3,2	3,3	**3,4**	3,5	3,6
	4	4,1	4,2	4,3	**4,4**	4,5	4,6
	5	5,1	5,2	5,3	**5,4**	5,5	5,6
	6	6,1	6,2	6,3	**6,4**	6,5	6,6

P(4 on 2nd die) = $\frac{6}{36} = \frac{1}{6}$, which is also just what you'd expect.

What's the probability that the first die lands on 3 **and** the second die lands on 4?

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	2,6
	3	3,1	3,2	3,3	**3,4**	3,5	3,6
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

P(3 on 1st die **and** 4 on 2nd die) = $\frac{1}{36}$. (Notice that this tells us that the outcome of the first die and the outcome of the second die are independent events, because

P(3 on 1st die **and** 4 on 2nd die) = $\frac{1}{36} = \frac{1}{6} \times \frac{1}{6}$ = P(3 on 1st die) x P(4 on 2nd die).)

Now let's find the probability that the first die lands on 3 **or** the second die lands on 4. First we'll do it directly by counting the outcomes that satisfy the condition:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	**1,4**	1,5	1,6
	2	2,1	2,2	2,3	**2,4**	2,5	2,6
	3	**3,1**	**3,2**	**3,3**	**3,4**	**3,5**	**3,6**
	4	4,1	4,2	4,3	**4,4**	4,5	4,6
	5	5,1	5,2	5,3	**5,4**	5,5	5,6
	6	6,1	6,2	6,3	**6,4**	6,5	6,6

Notice that there are only 11 of them. Don't count the 3,4 twice. Yes, it gets in on both parts of the **or**, but it's only one outcome! So P(3 on 1st die **or** 4 on 2nd die) = $\frac{11}{36}$. And this agrees completely with our formula:

P(3 on 1st die **or** 4 on 2nd die) =

P(3 on 1st die) + P(4 on 2nd die) – P(3 on 1st die **and** 4 on 2nd die) =

$$\frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}.$$

Activity #8: Rules of probability

Use the sex/attitude-toward-social-media contingency table to find the probabilities:

		Attitude toward Social Media		
		Negative	**Neutral**	**Positive**
Sex	**Male**	12	18	23
	Female	5	14	48

1) P(Female **and** Negative)

2) P(Female **or** Negative)

3) P(Neutral **given** Male)

4) P(Male **given** Neutral)

Use the two-dice chart to find the probabilities:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	2,6
	3	3,1	3,2	3,3	3,4	3,5	3,6
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

5) P(Odd on 1st die **and** Even on 2nd die)

6) P(Odd on 1st die **or** Even on 2nd die)

Assignment #3

		Number of Pets			
		No Pets	**1 or 2 Pets**	**3 or 4 Pets**	**At Least 5 Pets**
Sex	**Male**	13	24	12	4
	Female	8	34	12	13

Use the contingency table to answer these questions. Give answers as unreduced fractions. If one person is selected randomly from the survey group, find the probability that the person

1) Is female

2) Has one or two pets

3) Has no pets

4) Has fewer than five pets

5) Is female **and** has one or two pets

6) Is female **or** has one or two pets

7) Has at least one pet

8) Is male **given** that the person has no pets

9) Has no pets **given** that the person is male

10) Is male **and** has three or four pets

11) Is male **or** has three or four pets

12) Is female **and** has fewer than five pets

13) Is female **or** has fewer than five pets

14) Is female **given** that the person has one or two pets.

15) Has one or two pets, **given** that the person is female.

DIE #1 \ DIE #2	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

These questions deal with the rolling of two fair dice. Give answers as unreduced fractions.

Find the probability that

16) their sum is 5

17) their sum is not 5

18) their sum is an even number

19) they are both even numbers

20) one is even **and** one is odd

21) die #1 has a smaller number than die #2

22) their product is a perfect square

23) their sum is at least 9

24) their sum is at most 8

25) the first is 2 **and** the second is 5

26) the first is 2 **or** the second is 5

27) the first is even **and** the second is odd

28) the first is even **or** the second is odd

29) the absolute value of their difference is 3

30) their product exceeds 12

Lecture #9: Counting Rules and Probability

You probably think you already know how to count, and of course you do in the usual sense of the word, but when we talk about counting in the context of probability and statistics we mean something quite specific and quite different, as you will see. There are three rules of counting. They are related to each other and flow from each other. We'll develop these rules first and then relate them to probability.

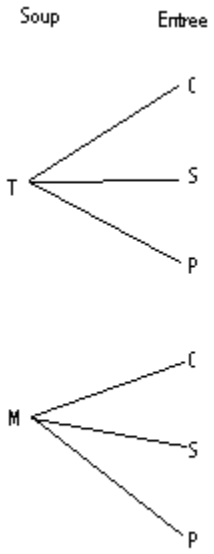
The Multiplication Rule of Counting

The first is the **multiplication rule of counting**. It applies when you're doing things in a certain order and at each stage you know how many possible actions you can take.

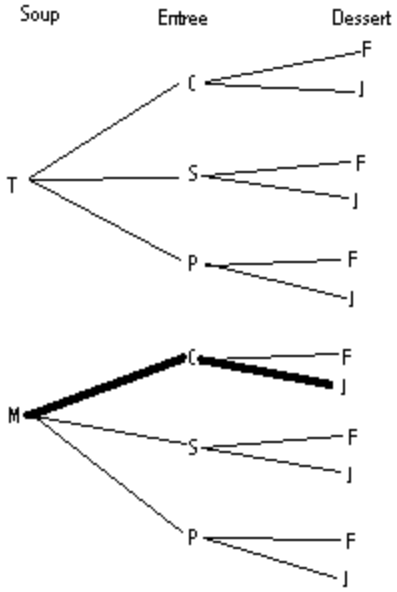
Say you're going out to dinner, and the menu contains two different soups, tomato and minestrone, three entrées (main courses), which are chicken, salmon, and pasta, and finally two different desserts, flan and jello. How many different meals can you order, if being different means that at least one of the choices differs from another meal?

One way (a very cumbersome one) to figure this out is to make a **tree diagram**. First you pick the soup (we'll use initials for all the choices):

Then you pick the entrée. For each choice of soup, there are three choices of entrée:



Finally, you choose the dessert, and for each two choices so far, either flan or jello can be chosen:



This looks kind of like a tree, if you tilt your head to the right. Each possible path from left to right gives a different meal. For instance, the thick line means you chose minestrone, chicken, and jello.

To find out the total number of meals, you have only to count the number of letters in the dessert column, because each one can be traced backwards, through the entrée and ending at the soup, to produce a unique meal. As you can see, there are 12 of these, so there are 12 possible meals.

This is obviously a very cumbersome method, and the multiplication rule of counting gives a shortcut: just take the number of choices at each stage of the meal, and then multiply these numbers together: $2 \times 3 \times 2 = 12$. Note that since multiplication is commutative, it wouldn't matter if you actually **chose** the entrée, then the dessert, and then the soup. You'd still have $3 \times 2 \times 2 = 12$ possible meals.

How many different meals can be ordered if there are four soups, five entrees, and three desserts available? $4 \times 5 \times 3 = 60$ meals.

This rule can be applied to various other real-life situations like figuring out how many license plate numbers are available in a certain format (the current California format is a digit, then three letters, then three more digits, for a total of $10 \times 26 \times 26 \times 26 \times 10 \times 10 \times 10 = 175{,}760{,}000$ possible license plates, minus a few because of unacceptable three-letter words – you can supply them), and in figuring out how many phone numbers are available in each area code (it would be 10^7 except not all digits can be used in certain positions). What happens when an area code runs out of phone numbers? Ten million might seem like a lot of phone numbers, but not when there are fax machines, pagers, cell phones, land lines, etc. being used all over. There are two alternatives. One is to split the area into two new areas, which means that many existing phones numbers have to change area codes, and this creates extra expenses for both the businesses and the residents that have to change. The other is the overlay, in which everybody keeps their area code, but new numbers in the area get a different area code, which means you might have to put in 10 digits to call your next-door neighbor.

Here's an example that we'll follow and modify all the way through the three counting rules. Say you're taking the letters *A*, *B*, *C*, and *D* and making two-letter "words" from them (in quotes because they don't really have to be words, just sequences of letters). Here's the sample space:

AA	AB	AC	AD
BA	BB	BC	BD
CA	CB	CC	CD
DA	DB	DC	DD

It's got 16 outcomes. But we could have figured this out without writing them all out by noticing that there are four choices for the first letter and four choices for the second letter, and because each choice for the first letter can be paired with each of the four choices for the second letter to make the word, there are $4 \times 4 = 16$ such words. So the multiplication rule of counting, as the name implies, tells us to take the number of possibilities at each stage of our process, whatever it is, and multiply them to find the total number of possibilities.

The rule also explains the size of the sample spaces for our coin flips. There are two possible ways each coin flip can turn out. So for two coins we get $2 \times 2 = 4$ possible outcomes, for three coins $2 \times 2 \times 2 = 8$, for seven coins $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^7 = 128$ and for *n* coins 2^n.

Permutations

The next counting rule involves what are called **permutations**, and it's best explained by going back to our example with the four letters and the two-letter "words." What if, after we pick a letter to start the word, we can no longer use that letter again in that particular word? Now the sample space would be

AB AC AD
BA BC BD
CA CB CD
DA DB DC

The words with a double letter, like AA, would be eliminated. We call this **without repetition**.

Note that $n(S)$ is 12 in this situation, and we could have found it by using the multiplication rule of counting, because there are still 4 choices for the first letter, but once it has been picked there are only 3 choices left, regardless of which letter we chose to begin the word, and $4 \times 3 = 12$. But we have another term and notation for this situation, namely **the number of permutations of 4 items taken 2 at a time**, or ${}_4P_2$. In general, if you have *n* choices and want to make a string using *r* of them, and you can't use them more than once in each string, and you want to know how many such strings are possible, you're talking about **the number of permutations of *n* items taken *r* at a time**, or ${}_nP_r$.

Permutations are a kind of special case of the multiplication rule of counting, in which at each stage of choosing we're selecting from the same pool we started with, but we can't pick any item twice. What if you had nine paintings and wanted to put four of them on your wall, and it really matters to you which position the four occupy? You'd have nine choices for the left-most one, eight for the next, seven for the third, and six for the right-most. Five paintings would be left out entirely. We call the number of possibilities **the number of permutations of 9 items taken 4 at a time**, or ${}_9P_4$. Even though it's not too hard just to multiply out $9 \times 8 \times 7 \times 6$ and get 3024, if the *r* gets much

bigger it's hard to keep track and know when to stop counting down, so we can just let the calculator, which has the ${}_nP_r$ function on it, do the work for us.

What if we had room for all nine paintings, but we still care about their exact arrangement on the wall? Then we'd have $9\times8\times7\times6\times5\times4\times3\times2\times1=362{,}880$ possible arrangements! By the time we got to the last painting we'd have no choice, because we'd already used the other eight. That's the meaning of the 1. We could symbolize this as ${}_9P_9$, but it is such a common and useful mathematical idea that we have another name and notation for it, namely **9 factorial**, or simply 9!. The exclamation point tells you to start with the integer to its left and keep reducing it by 1 until you get to 1 and take all the numbers and multiply them together.

Permutations come into play in problems where things have to be put into a certain order. How many 9-hitter batting orders can be made from a 25-person team? (A player can't appear more than once in the batting order.) ${}_{25}P_9$, or almost 750,000,000,000. How many slates of officers (president, vice-president, treasurer, secretary) can be chosen from a committee containing twenty people? (A committee member can't hold more than one office.) ${}_{20}P_4=116{,}280$.

Combinations

But what if you don't care who runs for president as opposed to vice-president, etc.? You simply want to pick a four-person subcommittee. Well, ${}_{20}P_4=116{,}280$ overcounts by a lot, because each four-person subcommittee has been included in the 116,280 as many times as those four people can be arranged in an order, as president, vice-president, etc. We call this situation a **combination** as opposed to a permutation, the difference being that in a permutation the order of the choices matters, but in a combination it doesn't, because you're just picking a group.

Our notation for this is ${}_nC_r$, or **the number of combinations of *n* items taken *r* at a time**, In the case of the four-person subcommittee, we'd write ${}_{20}C_4$, which comes out to 4845 on the calculator, and it would be called **the number of combinations of 20 items taken 4 at a time**.

Going back to our original situation with the four letters, a combination problem would be to ask how many **groups** of two letters could be formed from A, B, C, and D. Of course this is a without-repetition situation, but it also doesn't matter which letter is chosen first. These are the possible groups

A,B A,C A,D
B,C B,D C,D

We would write this as ${}_4C_2$, and it indeed equals 6.

For the baseball problem, we could ask how many different squads of nine players could be selected to take the field out of a team with 25 players. The answer would be ${}_{25}C_9 = 2{,}042{,}975$. Still quite a few possibilities! Luckily for the manager, not every group of nine has to be considered, because players play only certain positions.

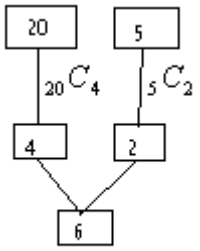
Sometimes it's hard to tell if a situation is a permutation or a combination. You have to ask yourself: is the order of choosing important or not? If it is, you have a permutation; if it isn't, it's a combination. But sometimes it's hard to tell **if** the order of choosing is important or not, and then you'd need more information to decide.

Let's say you're ordering a two-scoop ice cream cone, and you've got nineteen flavors to choose from. Well, maybe it matters to you which flavor is on the bottom and which one is on top. Maybe the chocolate needs to be on the bottom because you want to save it for last. In this case the situation would be a permutation; otherwise it would be a combination. And this doesn't even address the additional possibility that you want both scoops to be the same flavor, which would add another nineteen possibilities to the total.

And how about pizza? If you can choose three toppings from twelve, does it matter to you which one goes on first? Probably not, but it's something to consider.

Using More than One Rule in a Situation

Sometimes a counting situation requires you to use two or three of the rules of counting. Let's expand the example of the four-person subcommittee of the committee of twenty. There were ${}_{20}C_4$, or 4845, such subcommittees. But now let's say you're making a larger group, consisting of four people from the twenty-person committee and two more chosen from a committee of five. There are ${}_5C_2 = 10$ ways to pick the two new members of the enlarged group. Now we use the multiplication rule of counting to find the total number of ways the new group of six can be formed. First we chose the four out of twenty and then the two out of five (or we could do it in reverse order), and then multiply the number of choices for the two selections. Here's a little diagram:



So the answer is that there are $_{20}C_4 \cdot {}_5C_2 = 4845 \cdot 10 = 48{,}450$ ways to select the group of six.

Using Counting Rules to Find Probabilities

Here's an example of using the counting rules in probability to lend support to an accusation of discriminatory practices. Let's go back to the committee with 20 members, from which we want to select a four-person subcommittee. Add in the fact that 10 of the members are from an underserved minority and 10 aren't. It turns out the subcommittee contains not a single person from the minority. Is this discrimination?

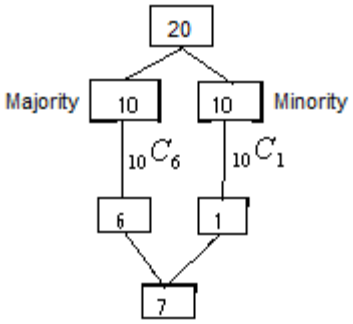
The way to find out is to begin by calculating the probability that such a subcommittee would contain no minority members. Remember, if each person on the committee were equally likely to be selected, i.e. if there were no discrimination and thus the assumptions of classical probability were valid, $P(E) = \dfrac{n(E)}{n(S)}$. In this case *S* is the sample space consisting of all possible four-person subcommittees of a 20-person committee, so $n(S) = {}_{20}C_4$. *E* is the event consisting of all possible four-person subcommittees made up exclusively of non-minority members, of whom there are 10, so $n(E) = {}_{10}C_4$. So $P(E) = \dfrac{{}_{10}C_4 \cdot {}_{10}C_0}{{}_{20}C_4} = \dfrac{210}{4845} \approx 0.043$. In other words, such an unbalanced subcommittee would occur just over 4 times out of a 100 simply by chance, even if no discrimination existed.

Is this unlikely enough for someone to decide that there **was** discrimination? It's a judgment call. But what if the subcommittee were to have six members, and it turned out that none were from the minority group? In that case, $P(E) = \dfrac{{}_{10}C_6 \cdot {}_{10}C_0}{{}_{20}C_6} = \dfrac{210}{38{,}760} \approx 0.005$, a small enough probability to make me think that the determination of discrimination is no longer a judgment call. I mean, once out of 200

times you would get such a make-up of the subcommittee by chance if there were no discrimination – give me a break!

Did you notice that the numerators in the two cases above were the same, 210? That's because if you pick 4 people out of 10, you leave out 6, and *vice versa*. In general, ${}_nC_r = {}_nC_{n-r}$, and in particular ${}_{10}C_4 = {}_{10}C_6$.

Let's end with a more complicated example. What if the subcommittee had some members of each group? Let's expand the subcommittee to seven members and calculate the probability that only one of them is a member of the minority group. This means that the other six of them aren't. First, how many such seven-member subcommittees are possible? Well, you'd have to pick six from the ten who aren't in the minority group. We already know that this is ${}_{10}C_6 = 210$. Then we'd have to pick one from the ten in the minority group: ${}_{10}C_1 = 10$. (Of course this equals ten, because there are ten different individuals who could be selected.) Now we use the multiplication rule of counting to figure out the total number of these groups consisting of six not in the minority group and one who is: ${}_{10}C_6 \cdot {}_{10}C_1 = 210 \cdot 10 = 2100$. The diagram illustrates this process:



The probability of picking such a group at random (without discrimination) is $P(E) = \frac{{}_{10}C_6 \cdot {}_{10}C_1}{{}_{20}C_7} = \frac{2100}{77{,}520} \approx 0.027$, or just under three chances out of a hundred. Is this a small enough chance to support the claim of discrimination? Again, this is a judgment call. It's not as small as 0.005, but it is smaller than 0.043, the number in our previous judgment call. Alas, there is seldom a definitive answer to such a question in statistics.

Activity #9: Rules of counting

1. In how many ways can 7 different DVDs be arranged on a shelf?

2. How many different 4-digit identification tags can be made if the digits can be used more than once? If the first digit must be a 3 and repetitions are not allowed?

3. How many ways can 3 cards be selected from a standard deck of 52 cards, if it doesn't matter which one is picked first, second, etc.?

4. How many ways can 5 tickets be selected from 50 tickets if each ticket wins a different prize and no ticket can win more than one prize?

5. How many different tests can be made from a test bank of 18 questions if the test consists of 6 questions? The order of the questions doesn't matter.

6. How many ways can 4 girls and 3 boys be selected from 12 girls and 9 boys?

7. If a committee has 8 women and 6 men, what is the probability to the nearest thousandth that a subcommittee of 5 people will contain

    a) All men?

    b) 1 woman and 4 men?

    c) 2 women and 3 men?

Assignment #4

1. A garage sale offers 10 different novels, 8 different CDs, and 3 different DVDs. You want to buy one of each. How many different ways can this be done?

2. How many different ways can 9 children line up?

3. At a carnival, you may pick 3 rides to go on and the order in which to ride them. You don't want to repeat any of the rides. There are 11 rides altogether. How many ways can you pick the rides?

4. How many different ways (orders) can you visit 6 tourist attractions during your vacation?

5. You want to study for 4 of your 6 courses tonight, and it matters to you which you study first, second, etc. How many different study schedules can you choose from?

6. How many ways can a diner select 3 appetizers and 2 vegetables if there are 7 appetizers and 5 vegetables on the menu?

7. How many ways can 6 sopranos and 7 altos be selected from 12 sopranos and 15 altos?

8. A baseball team has 7 pitchers: 4 are left-handed (lefties), and 3 are right-handed (righties). Three of them are injured. Find these probabilities, and give them as decimals rounded to the nearest thousandth:

    a. All three injured ones are lefties

    b. All three injured ones are righties

    c. 2 are righties and 1 is a lefty

    d. 1 is a righty and 2 are lefties

9. A large litter consists of 10 kittens, 3 of which are striped. If 4 are selected, find the probability of getting

    a. No striped ones

    b. 1 striped one

    c. 3 striped ones

    Express your answers as decimals rounded to the nearest thousandth.

Lecture #10: Discrete Probability Distributions

Some of this lecture should be very familiar to you, just old topics with new words attached to them. Other parts will expand the concepts we've been developing.

The first concept is the **random variable**. This just means assigning a number to each outcome of a probability experiment. We've already done this for rolling two dice: the sum of the upward-facing pips is the random variable. Flipping four coins: the number of heads is the random variable.

But there are two kinds of random variables, **discrete** and **continuous**. We used these words in the very first lecture in the course. Discrete random variables take on counting numbers as values: 0, 1, 2, and so on. Maybe there's a definite stopping place, like 12 for rolling two dice, or 4 for the number of heads in flipping four coins, but sometimes there isn't, or we don't know what it is. For instance, if the random variable is the number of phone calls a business receives in a given hour of the day, we don't know the highest possible value of the random variable, but we know the value will be 0, 1, 2, or a larger whole number.

Today we're going to talk only about discrete random variables and their **probability distributions**. We'll take on the **continuous** ones, like giving the height of a person when the probability experiment is to measure him or her, next time.

A Probability Distribution from Classical Probability

What if you're going to have four children, and each one will be either a boy or a girl, and the likelihood of either sex is the same for each birth? We can say that the sex of each child is independent of the sex of the others. This is the same as flipping four fair coins. (Some people don't believe any of this. They think that some people are more likely to produce boys than girls, or *vice versa*, or that once having produced a child of one sex you go on producing this sex. I'm not disputing these beliefs, but to have an example from classical probability we can't employ them.) Here's the sample space, using F for a girl and M for a boy, and listing the sexes in the order of birth:

FFFF

FFFM
FFMF
FMFF
MFFF

FFMM
FMFM
FMMF
MFFM
MFMF
MMFF

FMMM
MFMM
MMFM
MMMF

MMMM

The size of the sample space ($n(S)$) is 16, as you know from the multiplication rule of counting: $2 \times 2 \times 2 \times 2 = 2^4 = 16$. I've put boxes around the outcomes which result in the same number of boys, which is what we're going to use for the random variable, which we will call X. As you can see, there is one outcome resulting in 0 boys, four resulting in 1 boy, six resulting in 2 boys, four resulting in 3 boys, and one resulting in 4 boys.

Using the assumption of classical probability, that any outcome is as likely to occur as any other outcome, we can use the formula $P(E) = \frac{n(E)}{n(S)}$, where the event E is that you have X boys, to determine the probability that you have 0, 1, 2, 3, or 4 boys. We put this information into a table which is called a **probability distribution**:

X	P(X)
0	1/16 or 0.0625
1	4/16 or 0.25
2	6/16 or 0.375
3	4/16 or 0.25
4	1/16 or 0.0625

The left-hand column lists all possible values of the random variable X, and the right-hand column lists the probability that the value X will occur. So P(0) = 0.0625, P(1) = 0.25, etc. The P(X)'s have to add up to one, since one of the values of X has to occur each time the experiment (in this case, having four children) is performed.

As with any data set, we want to know two things: a measure of central tendency and a measure of variation. In a probability distribution, this will be the population mean, μ, and the population standard deviation, σ. Note that these are parameters, numbers

describing a population, because the probability distribution describes the total behavior of the random variable, not just a sample of it.

Here's how you find μ: first you multiply each value of the random variable by its associated probability. Since $0 \cdot 0.0625 = 0$, you get 0 for the first entry. The second is $1 \cdot 0.25 = 0.25$, and so on. Here's the completed column added on:

X	P(X)	X·P(X)
0	1/16 or 0.0625	0
1	4/16 or 0.25	0.25
2	6/16 or 0.375	0.75
3	4/16 or 0.25	0.75
4	1/16 or 0.0625	0.25

Now you add up the last column, whose sum can be expressed as $\Sigma X \cdot P(X)$, using our symbol for the sum. As you can see, $\Sigma X \cdot P(X) = 2$, and this is μ, the population mean. dxSo the formula for the population mean is $\mu = \Sigma X \cdot P(X)$, and it is our measure of central tendency.

Another name for this μ is the **expected value**, because it's the value of the random variable which you would most expect to find. In the case of having four children, you would expect on average that half, or 2, would be boys. But it's neat to see that it works out this way also by adding up the $X \cdot P(X)$ values. It works here because the probabilities are symmetrical around P(2) – in other words P(0) = P(4) and P(1) = P(3).

In general expected value is just a synonym for the mean, but it's a useful concept for answering questions like how many boys one would **expect** to find in 20 families with four children. Since you expect 2 per family, the answer is $20 \cdot 2 = 40$. (Of course, in this particular situation you could simply reason that since the 20 families have a total of 80 offspring, and boys and girls are equally likely to occur, you'd expect half of the 80, or 40, to be boys.)

How about σ? This is a somewhat more complicated calculation, but you'll get plenty of practice, and you can check your work on the calculator. First, you find $X - \mu$, the deviation from the mean, for each value of X. For 0, this would be $0 - 2 = -2$, for 1 it would be $1 - 2 = -1$, and so on:
K'

X	P(X)	X·P(X)	X - μ
0	1/16 or 0.0625	0	-2
1	4/16 or 0.25	0.25	-1
2	6/16 or 0.375	0.75	0
3	4/16 or 0.25	0.75	1
4	1/16 or 0.0625	0.25	2
		μ = Σ X·P(X) = 2	

This is somewhat reminiscent of how we found the standard deviation of a data set using the table method, and so is the next step, in which we square the $X - \mu$'s, remembering that squares are never negative:

X	P(X)	X·P(X)	X - μ	$(X - \mu)^2$
0	1/16 or 0.0625	0	-2	4
1	4/16 or 0.25	0.25	-1	1
2	6/16 or 0.375	0.75	0	0
3	4/16 or 0.25	0.75	1	1
4	1/16 or 0.0625	0.25	2	4
		μ = Σ X·P(X) = 2		

But the similarity stops here, because now we have to take into account the different probabilities, or weights, of the different values of X. What we do is multiply the $(X - \mu)^2$'s by the probability of the X's that produced them. The first one is $0.0625 \cdot 4 = 0.25$, the next $0.25 \cdot 1 = 0.25$, and so on (be careful to multiply by P(X) and **not** X·P(X)):

X	P(X)	X·P(X)	X - μ	$(X - \mu)^2$	$(X - \mu)^2 \cdot P(X)$
0	1/16 or 0.0625	0	-2	4	0.25
1	4/16 or 0.25	0.25	-1	1	0.25
2	6/16 or 0.375	0.75	0	0	0
3	4/16 or 0.25	0.75	1	1	0.25
4	1/16 or 0.0625	0.25	2	4	0.25
		μ = Σ X·P(X) = 2			

Isn't it interesting that the last column consists entirely of 0.25's, except for the 0?

Anyway, finally we've finished the table, and now we do something that is again reminiscent of the table method for finding the standard deviation of a data set: add up the last column: $\Sigma(X - \mu)^2 \cdot P(X) = 1$. However, we don't divide by n -1 or by N – this **is** the population variance, σ^2 (remember that in the corresponding process for data sets, first we got s^2, the sample variance).

So the formula is $\sigma^2 = \Sigma(X-\mu)^2 \cdot P(X)$, and from this we get the formula for the population standard deviation, σ: $\sigma = \sqrt{\Sigma(X-\mu)^2 \cdot P(X)}$. So in the case of the four children, $\sigma = \sqrt{1} = 1$, or 1.0, since we would round to the nearest tenth if the number didn't terminate by then, and it most likely **won't**, because it's a square root. This was a special case – not only is the standard deviation a rational (in fact, a whole) number, but it's numerically equal to the variance.

A Probability Distribution from Empirical Probability

In the example of the four-children families, we could figure out the probabilities for the different values of the random variable by looking at the sample space, because it was a classical probability situation. Now let's look at a random variable for which you have to be told the probabilities, because there's no way you could deduce them on your own.

Here's the (fictional) situation: Every day I buy lottery tickets, but not always the same number of tickets. I buy either one, or two, or three, or four tickets. There's no pattern or rotation to the number I buy on a given day, but over time I notice that I buy one ticket on half the days, two tickets on a fifth of the days, and three tickets also on a fifth of the days.

That's the story. The first task is to find the fraction of the days on which I buy four tickets, the only other possible number of tickets that I buy. Remember that since I **always** buy at least one ticket, the fractions of days I buy the different numbers of tickets must add up to one. To find the fraction remaining for buying four tickets, subtract the other fractions from 1:

$$1 - \frac{1}{2} - \frac{1}{5} - \frac{1}{5} =$$
$$\frac{10}{10} - \frac{5}{10} - \frac{2}{10} - \frac{2}{10} = \frac{1}{10}$$

So I buy four tickets on a tenth of the days.

We can interpret these fractions as probabilities, since how likely it is on any given day that I buy a certain number of tickets has to be interpreted as the fraction of days on which I **do** buy that many. Calling the number of tickets I buy on a given day the random variable X, we can make a probability distribution table for my ticket-buying activity:

X	P(X)
1	0.5
2	0.2
3	0.2
4	0.1

Now we're going to follow the same steps we did with the four-children family example and find the mean and standard deviation of the distribution. First the mean:

X	P(X)	X·P(X)
1	0.5	0.5
2	0.2	0.4
3	0.2	0.6
4	0.1	0.4

$\mu = \Sigma\, X.P(X) = 1.9$

So on an average day I buy 1.9 lottery tickets. Of course I **can't** do that, but **don't round this number**! If you round it, your answers to questions about how many tickets I can expect to buy over a certain period of time will be inaccurate. Take six weeks. That's 42 days. Since $42\,(1.9) = 79.8$, on the average I buy just under 80 tickets in six weeks. But if you round the mean of 1.9 to 2, you'd get $42\,(2) = 84$ tickets purchased, clearly an overstatement.

Here's a note of explanation about the term **expected value**. As stated before, expected value is just a synonym for mean, but consider this: What if, in the course of a ten-day period, my ticket-buying behavior exactly conforms to the pattern I explained above? Well, on half of those ten days, namely five, I'd buy one tickets, on a fifth, or two days, I'd buy two tickets, same for three tickets, and on a tenth, or one of the days, I'd buy four tickets. Here's a little table to show my ticket-buying behavior on the ten days:

# of Days	# of Tickets Bought per Day	Total # Bought
5	1	5
2	2	4
2	3	6
1	4	4
		19

(The Total # Bought is found by multiplying the # of Days by the # of Tickets Bought per Day.)

So I would have bought 19 tickets during those ten days, and since $\frac{19}{10} = 1.9$, I would expect to average 1.9 tickets bought per day.

Now on to the standard deviation, σ. Here I show the completed table:

X	P(X)	X·P(X)	X - μ	$(X - \mu)^2$	$(X - \mu)^2 \cdot P(X)$
1	0.5	0.5	-0.9	0.81`	0.405
2	0.2	0.4	0.1	0.01	0.002
3	0.2	0.6	1.1	1.21	0.242
4	0.1	0.4	2.1	4.41	0.441
		μ = Σ X.P(X) =1.9			

(Let me point something out here. Remember when we were finding the standard deviation of a data set, and when we got to the deviations from the mean column, $x - \bar{x}$, you saw that these deviations had to add up to 0? In fact this was a check to show you had the correct mean. In the four-children example, the $x - \mu$'s also added up to 0, but that was because the probabilities are symmetrical. In this lottery ticket example, the $x - \mu$'s clearly **don't** add up to 0, nor should they, because the mean is heavily weighted in the direction of X = 1, buying one ticket on a given day.)

Here are the calculations for the row X = 3, in case you're having trouble seeing where the numbers come from: $3(0.2) = 0.6;\ 3 - 1.9 = 1.1;\ 1.1^2 = 1.21;\ 1.21(0.2) = 0.242.$

The next step is to add the $(X - \mu)^2 \cdot P(X)$'s, and this gives us $\sigma^2 = \Sigma(X - \mu)^2 \cdot P(X) = 1.090$. So the population standard deviation is $\sigma = \sqrt{\Sigma(X - \mu)^2 \cdot P(X)} = \sqrt{1.090} \approx 1.0440$, or, to the nearest tenth (since we **have** to round this time), 1.0. (Remember to keep the 0 after the decimal point to show that you rounded to the nearest tenth, not to the nearest whole number.)

Finally, here's a little critical-thinking exercise using the lottery ticket example. Remember the phrase "the interval of data within one standard deviation of the mean"? When we encountered it, we used the notation $(\bar{x} - s, \bar{x} + s)$ to denote it. Now, since we're dealing with a population and parameters (a probability distribution **always** describes a complete population, not a sample), we'd call it $(\mu - \sigma, \mu + \sigma)$.

Here's the question: What is the probability that on any given day I buy a number of tickets within one standard deviation of the mean?

There are three steps to do. First find the interval $(\mu-\sigma,\mu+\sigma)$ for this example: $(\mu-\sigma,\mu+\sigma)=(1.9-1.0,1.9+1.0)=(0.9,2.9)$.

Next, find the values of X that actually fall within this interval. Well, 1 does, and so does 2, but when you get to 3 it's too big, because $3>2.9$. So buying one ticket or buying two tickets are the only ways of buying a number of tickets within one standard deviation of the mean.

Finally, find the probability of buying one or two tickets. Since I can't do both on one day (the two events are **mutually exclusive**), you find P(1 **or** 2) by adding P(1) and P(2).

X	P(X)
1	**0.5**
2	**0.2**
3	0.2
4	0.1

And since $P(1)+P(2)=0.5+0.2=0.7$, the answer to the question is that the probability that on any given day I buy a number of tickets within one standard deviation of the mean is 0.7, or 70%.

Activity #10: An empirical discrete probability distribution

Here's a situation very similar to the lottery ticket example in the lecture, only this time I could also buy five tickets on one day, and the fractions of days are as follows: a tenth of the days I buy one ticket; a quarter of the days I buy two tickets; a fifth of the days I buy three tickets, and a fourth of the days I buy four tickets.

1) On what fraction of the days do I buy five tickets?

2) Make a probability distribution for the random variable X, the number of lottery tickets that I buy on a given day.

3) Find the mean of the distribution.

4) Find the standard deviation of the distribution, to the nearest tenth.

5) What is the probability that on any given day I buy a number of tickets within one standard deviation of the mean?

Assignment #5

1) A couple plans to have three children. The sex of each child is independent of the sex of the other children, and boys and girls are equally likely to be born.
   a) Using M for a boy and F for a girl, list all possible birth-order sequences for the three children.
   b) Let x be the random variable denoting how many boys are in the family.
      i) Make a probability table for the random variable x
      ii) Find the mean of the probability distribution of x. Show your work
      iii) Find the standard deviation for the probability distribution of x to the nearest tenth. Show your work.
   c) How many boys would you expect in 30 such families?

2) Do the same things and answer the same questions as in #1, only this time the couple plans to have five children.

3) Jon eats anywhere from one to five candy bars per day, a decision he makes in a random way. One-tenth of the days he eats one candy bar. One-fifth of the days he eats two candy bars. Two-fifths of the days he eats three candy bars, and one-tenth of the days he eats four candy bars.
   a) What fraction of the days does he eat five candy bars?
   b) Let the random variable x be the number of candy bars Jon eats during a day.
      i) Make a probability table for the random variable x
      ii) Find the mean for the probability distribution of x. Show your work
      iii) Find the standard deviation for the probability distribution of x to the nearest tenth. Show your work.
      iv) What is the probability that on any given day Jon will eat a number of candy bars that is within one standard deviation of the mean? Show your work.
      v) What is the expected value of the probability distribution of x?
   c) How many candy bars would Jon average in 8 days?
   d) How many candy bars can Jon expect to eat in 5 weeks?
   e) How many candy bars can Jon expect to eat in a year?

Lecture #11: Continuous Probability Distributions

You've seen now how to handle a discrete random variable, by listing all its values along with their probabilities. But what if you're dealing with a continuous random variable, like height or weight or duration – something measured – and you want to talk about the probability of the random variable taking on different values?

Clearly you can't just list all the possible values. You'd have to spend the rest of your life doing it, and even then you wouldn't make a dent. Say you were weighing something, and the random variable is the weight. Even if you could give a probability for, say, 42.783 g and 42.784 g, you'd still have to list, for instance, 42.7835 g. Between each two rational numbers there is another one, and so on and so on. We say that these numbers are **dense**. We simply can't list them all.

So with continuous random variables a whole different approach to probability is used. First, we don't speak of the probability that the random variable takes on an individual value. Instead we deal with the probability that the random variable falls **within a certain range of values**. The probability that the variable takes on an individual value is 0. **Nothing** weighs 42.783 g on the nose, but there may be a positive probability that whatever it is we're weighing will weigh **between** 42.783 g and 42.784 g.

And how do we find these probabilities of ranges of values? Not by adding up individual probabilities, as you can see, but by using a concept from calculus (don't let that word scare you – you'll find that it's surprisingly easy to grasp): We look at what's called the **area under the curve**.

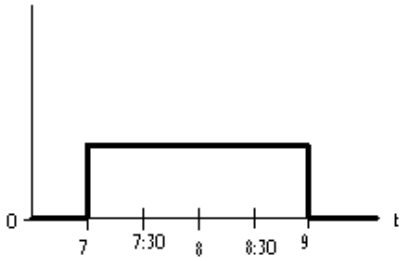
Probability Density Functions

Let me explain this using a really simple example. Let's say that I arrive at work every morning between 7 a.m. and 9 a.m. I never come before 7 or after 9. It isn't that I mostly arrive pretty near 8 a.m. I am as likely to arrive at 7:19 as at 7:58 as at 8:13 as at 8:45. In mathematical terms, my arrival times are **uniformly distributed** from 7 a.m. to 9 a.m.

What fraction of the mornings will I arrive between 7:30 and 8? Another way to ask this question is: What is the probability that on a given morning I will arrive between 7:30 and 8? The answer is $\frac{1}{4}$ or 0.25 or 25%. This is because since I arrive 100% of the days between 7 a.m. and 9 a.m., and because I'm equally likely to arrive at any time between 7 and 9, and because the half hour from 7:30 to 8 is one-fourth of this two-hour

interval, the probability that I'll arrive between 7:30 and 8 is one-fourth of 100%, or 25%. I hope that seems obvious.

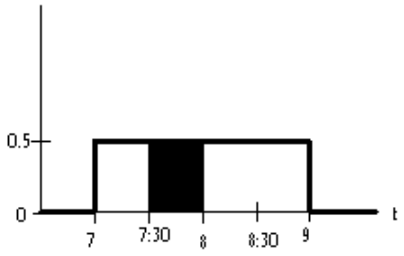
But here's how we'd do this using the probability distribution of a continuous random variable. Look at this graph:



The *t*-axis represents my time of arrival at work. The thick line, which we call a **probability density function**, represents the probability of my arriving at work. The probability is 0 before 7 a.m. (the thick line coincides with the *t*-axis), shoots up to a certain level at 7 and maintains that level, and then drops back down to 0 at 9 a.m.

What is this certain level? To find that, think back to discrete probability distributions. There, the P(X)'s all had to add up to 1. One of the X values was sure to occur. It's the same thing in the continuous case, only now we talk about the **total area under the curve** (and above the *t*-axis) equaling 1. In fact, that's part of the definition of a probability density function: the total area under it must equal 1. In our case the area is a rectangle bounded by the vertical lines at 7 and at 9, the *t*-axis, and our probability density function. This area is 1. The area of a rectangle is the product of its height and its width. The width is 2 hours, from 7 to 9. So the height must be $\frac{1}{2}$, or 0.5, because $\frac{1}{2} \cdot 2 = 1$.

Now I'll label the vertical axis with the 0.5 and shade in the area we're interested in:



In this system, the probability that I'll arrive between 7:30 and 8 is equal to the shaded area. It too is a rectangle, with width $\frac{1}{2}$ (the half hour from 7:30 to 8) and height $\frac{1}{2}$, so

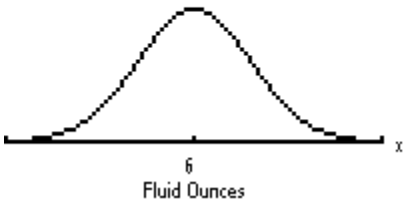
its area is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = 0.25 = 25\%$. This is the same answer we got using common sense. I just wanted to illustrate the concept of the probability density function (pdf) and area under the curve, and especially to emphasize the defining characteristic of a pdf as having a total area underneath it (and above the horizontal axis) of 1.

The Normal Distribution

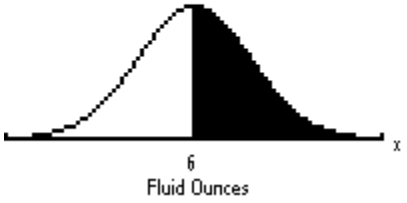
The most famous pdf is what we call the **normal distribution**. You've probably heard it called the bell curve. It is indeed bell-shaped. It quite accurately describes an enormous range of real-life phenomena where the final value of the random variable depends on many, many very small occurrences which themselves are random. If a random variable conforms to a normal distribution, we call it **normally distributed**. Heights of adult women in the U. S. are approximately normally distributed, as are the lengths of their feet and the circumferences of their heads. Same for adult men.

Think of those vending machines that dispense terrible coffee and are supposed to put 6 fluid ounces in a paper cup. If you actually measured the volume of many of these filled cups, made a histogram of the volumes, and let your eyes get out of focus so you saw the tops of the bars as a curved line, that curve would be bell-shaped. Most of the cups would have close to 6 fluid ounces in them; there would be the same number with a certain amount less as there are with that certain amount more (the histogram would be symmetrical), and as you got further and further away from 6 the bars would be shorter and shorter.

Here's a representation of this normal distribution:

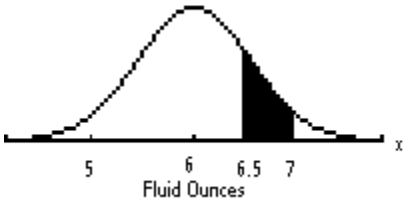


It's symmetrical around 6, the mean of the distribution of X, the random variable that represents the number of fluid ounces deposited in a cup, and in theory the pdf never touches the *x*-axis (though of course you couldn't have a negative number of fluid ounces in a cup, for instance), and because it's a pdf the total area under the curve is 1. So if you want to know what fraction of the cups have at least 6 fluid ounces of coffee, you represent the situation this way:



and the answer, of course, would be $P(X > 6) = 0.5$. You'll be delighted to learn that for the purposes of continuous random variables, it makes no difference whether we say "greater than" or "at least," because there is zero probability associated with X=6.

If we wanted to know $P(6.5 < X < 7)$, the probability that a cup has between 6.5 and 7 fluid ounces, we would make this drawing:



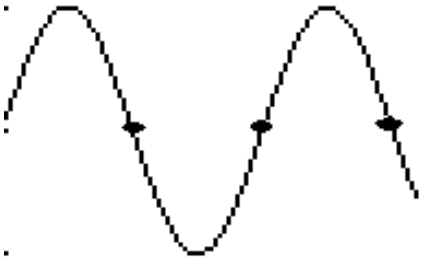
There are an infinite number of normal distributions. The shape and location of their pdf's depends entirely on the mean and the standard deviation of the distribution. The mean determines the central point of the distribution. Here are two normal pdf's with different means but the same standard deviation:

And here are two normal pdf's with the same mean but different standard deviations:

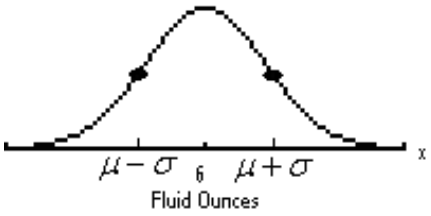
The smaller the standard deviation (i.e. the less varied the values of the random variable), the pointier the pdf.

The mean is apparent from looking at a normal pdf – it's the value of the random variable directly under the hump. How about the standard deviation? It's more subtle, and you don't have to worry about it, but if you want your drawing to be accurate, then at one standard deviation below the mean, $\mu - \sigma$, and at one standard deviation above the mean, $\mu + \sigma$, the pdf has what we call in calculus **inflection points**.

Inflection points are points where a curve changes what we call its concavity – roughly, it goes from being cupped upward to being cupped downward, or *vice versa*. Here's an example with a sine curve, which changes it concavity an infinite number of times throughout its span. I've marked the inflection points:



Let me do the same with our original normal curve:



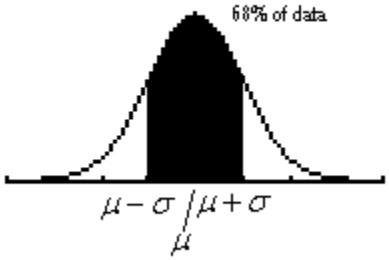
$\mu + \sigma$ looks to be located approximately at 6.7 fluid ounces, which would mean that $\sigma = 0.7$, which is larger than it should be (that machine needs servicing – it's way out of whack!!).

(For those of you interested in the math behind this, here is the formula for a normal pdf:

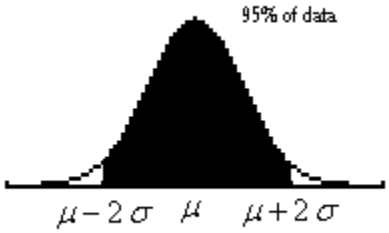
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma}}$$

where μ is the mean of the normal distribution and σ is its standard deviation. It contains two of the most famous numbers in mathematics, π, which, as you remember, is the ratio between the circumference and diameter of all circles, and *e*, which you might remember from algebra as being the natural logarithmic base.)

The great thing about normal distributions is that we know exactly what fraction of the data falls within any interval, in particular within one standard deviation of the mean,



within two standard deviations of the mean,



and a whopping 99.7% of the data lie within three standard deviations of the mean.

We can also find the percent of data that lie in any interval at all. We used to do this by means of what's called the standard normal table, an ingenious but cumbersome chart that requires lots of preliminary calculations and lacks the level of accuracy we would like to achieve. If you want to look at one, try this:
http://www.sjsu.edu/faculty/gerstman/EpiInfo/z-table.htm

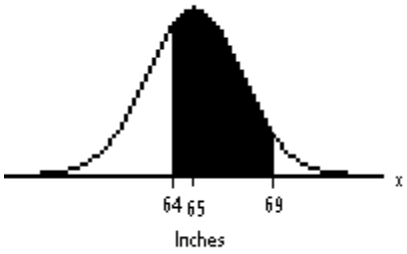
With graphing calculators and computer math packages we can do much better, much more easily. I'll do the following problems as they would be done on the TI-83 or TI-84.

Practice Problems: Finding Probabilities for Normal Distributions

For all these problems, we're going to assume that women's heights are normally distributed with a mean of 65 inches and a standard deviation of 3 inches.

1) What is the probability that a woman is between 64 inches and 69 inches tall (5'4" to 5'9")? Put another way, what fraction of women's heights are in this range? We would write this P(64 < X < 69).

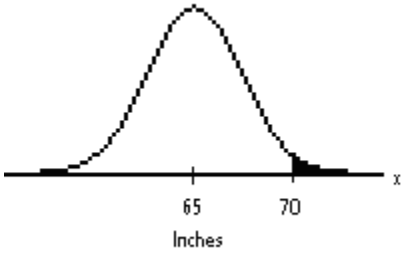
First, draw a horizontal axis and label it *x*, write the units (inches) below it, and draw a normal pdf above the axis. Then mark and label 65 on the axis underneath the highest point of the pdf, mark and label 64 to the left of 65 and 69 to the right of 65, draw vertical lines from the 64 and the 69 to the curve and shade the part between them, above the *x*-axis, and under the curve:



Using the normalcdf (**not** normalpdf) function, enter the number on the left where the shading begins, the number on the right where it ends, the mean of the distribution, and its standard deviation, all separated by commas, normalcdf (64, 69, 65, 3), and you will get 0.539347. Round this to the nearest ten-thousandth (four places after the decimal point), or equivalently to the nearest hundredth of a percent, and you come up with the correct answer: **0.5393, or 53.93%**. (These instructions, and others in this text, are about the older operating system of the TI's. If you have a newer operating system, you'll see that it prompts you for the numbers you need to input, so it's easier than having to remember "left, right, mean, standard deviation.")

2) What is the probability that a woman is taller than 5 feet, 10 inches, or 70 inches? Put another way, what fraction of women are taller than 70 inches? This would be written as P(X > 70).

Start the same way as in Problem 1, but you have to mark and label only one number besides the mean, the 70. Then shade **to the right of** the 70, because that's where the taller heights are:



The only complication using normalcdf is that there **is** no number on the right where the shading ends, so put in a big one, and if you're not sure if it's big enough put in a bigger one and see if it changes your answer, at least to the nearest ten-thousandth. normalcdf (70, 1000, 65, 3) ≈ 0.04779, so the rounded answer is **0.0478, or 4.78%**.

3) What is the probability that a woman is shorter than 67 inches, or what fraction of women are shorter than 67 inches, written P(X < 67)?

This time you shade to the left:

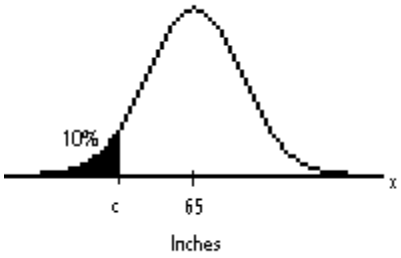
And, since the shading goes all the way to the left, in theory anyway, there is no smallest shaded number, so just proceed the way you did in Problem 2, testing with a smaller number if you are not sure yours was small enough:
Normalcdf (0, 67, 65, 3) ≈ 0.7475075, making the answer **0.7475, or 74.75%.**

Practice Problems: Finding Cut-offs for Normal Distributions

In the problems above, we found the probability that the random variable falls within a certain range. Now we're going to reverse the process. We'll start with the probability of a certain range, and then we'll have to find the values of the random variable that determine that range. I'll call these values cut-offs.

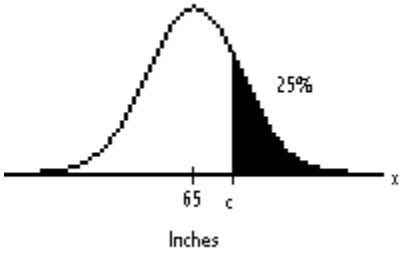
In these three problems, we'll use the same situation as above: Women's heights are normally distributed with a mean of 65 inches and a standard deviation of 3 inches.

1) How short does a woman have to be to be in the shortest 10% of women? If we call this cut-off *c*, this could be written as finding *c* such that $P(X < c) = 0.10$. We'll do the same kind of diagram as before, but this time we'll label the known probability, 10%, and we do this above the shaded area, definitely not on the *x*-axis, because it's an area, not a height. The hardest part of the diagram is deciding which side of the mean to put the *c* on and which side of the *c* to shade. You really have to think about it. In this case, since by definition 50% of women are shorter than the mean, the cut-off for 10% has to be less than the mean:



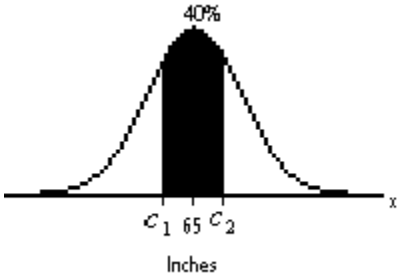
Using the calculator, which will find a cut-off using the invNorm function, followed by the percent of data under the normal curve **to the left of** (always to the left of, no matter which side of *c* the shading is on) the cut-off, then the mean and standard deviation, separated by commas, we get invNorm (0.10, 65, 3) ≈ 61.1553, or, to the nearest tenth of an inch, 61.2 inches. So about **10% of women are shorter than 61.2 inches**. You can check this using normalcdf, and you might as well use more of the cut-off than we rounded to, for greater assurance that your check shows you got the right answer. You get normalcdf (0, 61.1553, 65, 3), which come to 0.0999997, or 10%.

2) How tall does a woman have to be to be in the tallest fourth of women? (What is the cut-off for the tallest 25% of women?) If we call this height *c*, we want to find the value of *c* such that $P(X > c) = 0.25$. Here's the diagram:



Even though we shaded the area to the **right** of the cut-off, because that's where the tallest 25% of women are, when we use invNorm we must put in 0.75, which is $1-0.25$, because the calculator finds cut-offs for areas **to the left** only: invNorm (0.75, 65, 3) ≈ 67.02, or **67.0 inches**, to the nearest tenth. Checking with normalcdf (67.02, 1000, 65, 3), we get about 0.25037, pretty close to 25%.

3) What if we're interested in finding cut-offs for a **middle** group of women's heights, say the middle 40%? Obviously, we're looking for two numbers here, one on either side of the mean. Call them c_1 and c_2:



To use invNorm, we must find out how much area is under the curve to the **left** of c_1. Well, if 100% of area is under the entire curve, then what's left over after taking away the middle 40% is $1-0.40=0.60$, and since that 60% is split evenly between the two tails (the parts at the sides), that gives 30% for each tail. So c_1 is the number such that P(X < c_1) = 0.30, and invNorm (0.30, 65, 3) ≈ 63.4268, or 63.4 inches. How much area is there under the curve to the **left** of c_2? Either subtract the 30% to the right from 100%, or add up the 30% in the left tail and the 40% in the middle, and you'll get 70% either way. So c_2 is the number such that P(X < c_2) = 0.70, and invNorm (0.70, 65, 3) ≈ 66.5732, or 66.6 inches. So to the nearest inch, **the middle 40% of heights go from 63.4 to 66.6 inches**. The check is normalcdf (63.4268, 66.5732, 65, 3) ≈ 0.3999968, or 40%.

Activity #11a: Normal distribution problems

For all these problems, assume that women's heights are normally distributed with a mean of 65 inches and a standard deviation of 3 inches. Include a diagram for each problem. Round probabilities to the nearest ten-thousandth and cut-offs to the nearest tenth of an inch.

1) What is the probability that a woman is between 60 inches and 63 inches tall?

2) What is the probability that a woman is taller than 5 feet 4 inches, or 64 inches?

3) What is the probability that a woman is shorter than 62 inches?

4) How short does a woman have to be to be in the shortest 20% of women?

5) How tall does a woman have to be to be in the tallest fifth of women?

6) What are the cut-offs for the heights of the middle 50% of women?

Activity #11b: Normal distribution practice

For these problems, assume that the lengths of human pregnancies are approximately normally distributed with mean 266 days and standard deviation 16 days. Include a diagram for each problem. Round probabilities to the nearest ten-thousandth, or hundredth of a percent, and round number of days to the nearest whole number.

1) What fraction of pregnancies last less than 250 days?

2) What fraction of pregnancies last between 240 and 280 days?

3) Find the cut-off for the longest 30% of pregnancies.

4) Find the cut-offs for the middle 80% of lengths of pregnancies.

Lecture #12: The Central Limit Theorem

Now that you've learned how to determine probabilities and cut-offs for normal distributions, you might wonder how you can be (reasonably) sure that a distribution **is** normal. After all, the normalcdf and invNorm functions are valid only for normal distributions.

There are various sophisticated techniques for making this determination, but your calculator has a relatively simple one called NormProbPlot.

But more importantly, there's a way in which **every** distribution can be turned into a normal one, allowing us to find probabilities and cut-offs, and this way is part of what we call the **Central Limit Theorem**, a result from advanced calculus (don't worry, though), which we will use throughout the inferential statistics part of this course.

What I want to do here is to give you a sense of it, and give you an important formula from it, and let it sit on the back burner until we begin to use it.

It's based on the concept of a **sampling distribution**. Think about women's heights. We take a woman's height; maybe she's shorter than average, maybe she's average, maybe she's taller. We have assumed that these heights, taken as a population, are normally distributed with a certain mean ($\mu = 65$ inches) and a certain standard deviation ($\sigma = 3$ inches). We called the random variable for height *X*. Instead of $\mu = 65$ inches, we could write more precisely $\mu_x = 65$ inches, and we could also write $\sigma_x = 3$ inches.

Now imagine that we form groups, or samples, of ten women, many, many such samples, the members of which are randomly selected from women as a whole, and for each sample we look at the sample **mean** and make a data set of those means rather than the individual heights. This set of sample means is called a sampling distribution. Now we might want to find the usual things about the set of means – a measure of central tendency and a measure of variation.

You can think of this measure of central tendency as the **mean of the means**, a kind of a second-level mean. Does it make sense to you that this mean, which we would label $\mu_{\bar{x}}$ instead of μ_x, would have to be the same as μ_x? You're taking the same population, taking samples from it, and looking at their means – how could this set of means have a different mean than the population it came from, which we now call the **parent population**? It couldn't. And it doesn't matter how big the samples are which

you're taking to make the sampling distribution. The mean of these $\bar{x}$'s will be the same as the mean of the *x*'s no matter how many are in the samples (i.e. how big *n* is). Symbolically, $\mu_{\bar{x}} = \mu_x$.

But the standard deviation, the measure of variation, is a different story. It might not be that unusual to find a woman who is at least six feet tall. According to my – admittedly made up – parameters, normalcdf (72, 1000, 65, 3) ≈ 0.00982, or approximately 1% of women fall in this category. But how unusual it would be to find a randomly-selected sample of ten women whose **average** height is at least six feet! If some of them were under six feet, there would have to be some very tall ones to average six feet or more as a group. And the likelihood of all ten being at least six feet tall is miniscule (in fact, 0.01^{10}, or 1×10^{-20}, a decimal point followed by 19 zeros and then a one).

Can you also see that the larger the sample size *n* that is used for the sampling distribution, the more unlikely it is that samples will have means very different from the mean of the parent population? So it's not true that $\sigma_{\bar{x}} = \sigma_x$. In fact, the formula looks like this: $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. The bigger *n* is, the larger number you're dividing into the standard deviation of the parent population, and the smaller the quotient, $\sigma_{\bar{x}}$. We'll see a demonstration of this formula in a little while.

The standard deviation of the sampling distribution, $\sigma_{\bar{x}}$, has a special name. It's called the **standard error of the mean**.

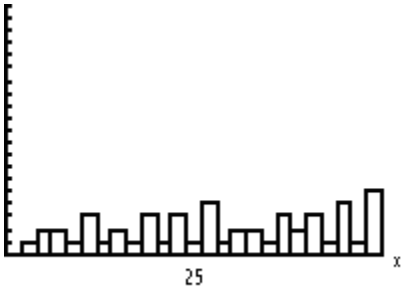
So a sampling distribution, while not changing the mean of the parent distribution, tightens it up and draws it together, and the larger the sample size the greater this effect. But that's not all it does. Remember how I said that every distribution could in some sense become a normal one? That's the other part of what the Central Limit Theorem does for us.

First of all, if the parent distribution is itself a normal one, then the sampling distribution is also normal, no matter what the sample size, *n*, is. However, for any parent distribution, even the most un-normal ones, as *n* gets bigger, the sampling distribution looks more and more normal, and at a certain point you might as well just consider it normal for the purposes of finding probabilities and cut-offs. And what is that point? It turns out that if *n* is at least 30, in other words if the sampling distribution is made up of samples of size 30 or more, then the distribution may be considered approximately normal.

The upshot is that if we examine a sample of size 30 or more for some distribution, no matter what kind it is, we can view the sample as part of a sampling distribution of all samples of its size, and we can use normalcdf and invNorm to our hearts' content.

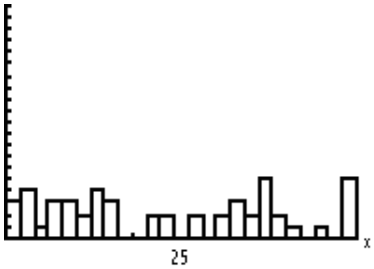
To illustrate all parts of the Central Limit Theorem, let's consider the distribution of a random variable in which an integer between 0 and 50, inclusive (meaning that 0 and 50 are part of the group), is selected randomly. So each of the 51 integers is equally likely to be chosen on any turn. This is just like the example of my coming to work at any time between 7 a.m. and 9 a.m., except that was a continuous random variable and this one is discrete. But it's still called a **uniform distribution** because the probabilities are uniformly distributed. It is very definitely **not** a normal distribution.

In this example I had the calculator generate 50 values of this random variable and do the one-variable stats on them. I called this Parent Distribution #1. Its mean was 28.12 and its standard deviation about 14.245. Here's a histogram of Parent Distribution #1 (the class width is 2, so each bar represents 2 values):



It looks like the skyline of a small city with no proper city planning. There's no pattern to it, and there shouldn't be one in a uniform distribution.

I then had the calculator generate another 50 values of the random variable, and I called this set Parent Distribution #2. Its mean was 22.50, and its standard deviation about 15.712. Here's its histogram:



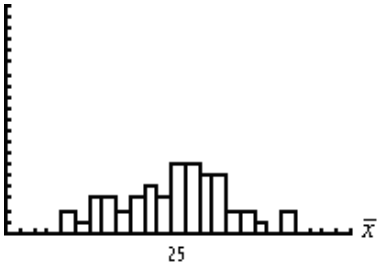
It's definitely different from the histogram of Parent Distribution #1, but it has no pattern either.

You probably figured out that the mean of the random variable is 25, the integer halfway from 0 to 50. And that mean is the population mean, μ, whereas the 28.12 and the 22.5 are $\bar{x}$'s, or sample means, because they were computed from samples, in this case of size $n = 50$.

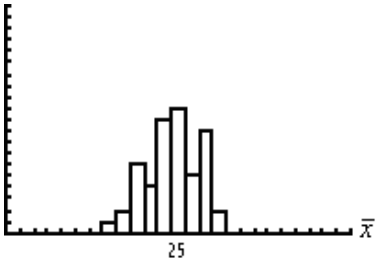
What about the standard deviations? Our examples gave us values of s = 14.245 and 15.712. It takes advanced algebra to figure out what the population standard deviation σ_x is: approximately 14.72. You can see that our samples yielded values fairly close to this.

On to sampling distributions. First we generate 50 samples each of size $n = 4$, then of size $n = 16$, and finally of size $n = 64$. Look at the histograms of their $\bar{x}$'s:

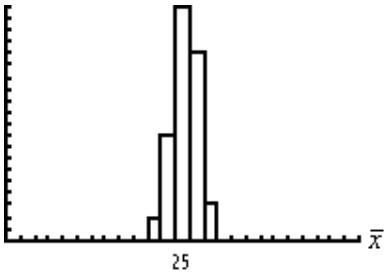
***n* = 4:**



***n* = 16:**



***n* = 64:**



There are several things to notice about these histograms. First, the axes are labeled $\bar{x}$ rather than plain *x*, because these are sampling distributions. Second, all three are centered around 25, the mean of the parent distribution. Third, they tighten up as *n* gets bigger, and they look more and more like normal distributions. Even when *n* = 4, the histogram has begun to look almost normal.

This table displays the means and standard deviations of the three sampling distributions. Notice that the means are labeled $\bar{x}_{\bar{x}}$, and the standard deviations $s_{\bar{x}}$, because we're looking at samples of the sampling distributions.

	***n* = 4**	***n* = 16**	***n* = 64**
$\bar{x}_{\bar{x}}$	24.445	24.275	25.245
$s_{\bar{x}}$	7.813	3.902	1.845

Finally, let's see if the three samples support the formula for the standard error of the mean, $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. For *n* = 4, we would expect $\sigma_{\bar{x}} = \frac{14.72}{\sqrt{4}} = \frac{14.72}{2} = 7.36$. Our sample yielded 7.813. Pretty close. For *n* = 16, we would expect

$\sigma_{\bar{x}} = \frac{14.72}{\sqrt{16}} = \frac{14.72}{4} = 3.68$. Our sample yielded 3.902. For n = 64, we would expect $\sigma_{\bar{x}} = \frac{14.72}{\sqrt{64}} = \frac{14.72}{8} = 1.84$, and we got 1.845. Very close indeed.

This enlarged table shows the comparison:

	***n* = 4**	***n* = 16**	***n* = 64**
$\bar{X}_{\bar{x}}$	24.445	24.275	25.245
$s_{\bar{x}}$	7.813	3.902	1.845
$\sigma_{\bar{x}}$	7.36	3.68	1.84

That's about all there is to say now. Just remember, when we start doing inferential statistics, that the sampling distribution has the same mean as the parent distribution, that its standard deviation (standard error of the mean) is the standard deviation of the parent distribution divided by the square root of the sample size, and that the larger n is the closer the sampling distribution is to a normal distribution, with a sample size of n = 30 being the threshold for considering the distribution to be approximately normal.

Exam #2 – Probability – Sample

For partial credit, show your work.

I am going to use the first eight letters of the alphabet, A, B, C, D, E, F, G, and H, in problems 1-3.

1) How many different three-letter "words" (they don't have to make sense) can I make if I am not allowed to use the same letter more than once?

2) How many groups of three letters can I pick, if I don't care what order I pick them in?

3) How many different three-letter "words" can I make if I am allowed to use the same letter more than once?

4) A co-ed softball team has 16 members, 7 men and 9 women. If all players are equally likely to be in the starting lineup, what is the probability that a starting lineup of 10 players will have 5 men and 5 women? Round your answer to the nearest thousandth.

In Problems 5-12, use this situation: You have a fair coin and a spinner that has a pointer that can point to 1, 2, 3, 4 or 5. A probability experiment is performed in which you flip the coin once and spin the spinner once. The spinner is equally likely to point to any number. In Problems 5-11, leave answers as unreduced fractions.

5) List all the elements of the sample space for this experiment. For instance, call getting a head on the flip and getting a 2 on the spin H2.

6) What is the probability of getting a tail **or** spinning a 5?

7) What is the probability of getting a tail **and** spinning a 5?

8) What is the probability of **not** spinning a 3?

9) What is the probability of getting a tail **or** spinning an odd number?

10) What is the probability of getting a tail **and** spinning an odd number?

11) What is the probability of spinning **at most** a 4?

12) What is the probability of spinning **at least** a 4?

This table gives entertainment preferences for 300 people. Use this table for Problems 13-18. Leave answers as unreduced fractions.

	UNDER 30	30 OR OVER
CLUB	86	36
THEATER	37	41
NEITHER	52	48

If a person is randomly selected from the group, what is the probability that

13) The person does **not** prefer the theater?

14) The person prefers clubs **or** the theater?

15) The person prefers clubs **given that** the person is under 30?

16) The person is under 30 **given that** the person prefers clubs?

17) The person is 30 or over **or** prefers theater?

18) The person is 30 or over **and** prefers theater?

A public speaker gives anywhere from 0 to 5 speeches each week. Here is a probability distribution table for *X*, the number of speeches given during any given week:

X	***P(X)***
0	0.05
1	0.49
2	0.23
3	0.12
4	0.08
5	0.03

19) What is the mean of the distribution? Show your work.

20) To the nearest hundredth, what is the standard deviation of the distribution? Show your work.

21) What is the probability that in any given week the speaker will give a number of speeches that is within one standard deviation of the mean?

Problems 22-25 refer to weighing bags of M&Ms. The weights are normally distributed, with mean 67.28 grams and standard deviation 1.53 grams. Make a diagram for each problem and answer the question. In Problems 22 and 23, give answers to the nearest ten-thousandth.

22) What fraction of bags weigh between 66.12 grams and 68.06 grams?

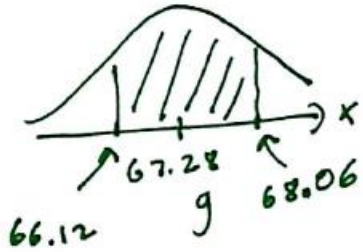
23) What fraction of bags weigh less than 66.92 grams?

24) To the nearest hundredth, what is the cut-off for the weight of a bag of M&Ms which is heavier than 75% of the bags?

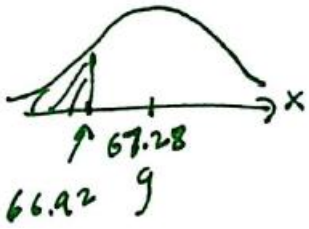
25) To the nearest hundredth, what are the cut-offs for the weights of the middle 30% of the bags?

Exam #2 – Probability – Sample – Answers

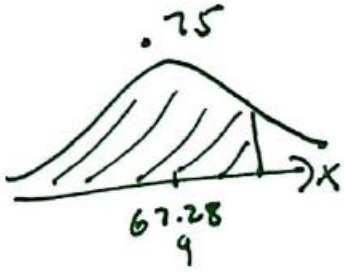
1. 336
2. 56
3. 512
4. 0.330
5. H1, H2, H3, H4, H5, T1, T2, T3, T4, T5
6. 6/10
7. 1/10
8. 8/10
9. 8/10
10. 3/10
11. 8/10
12. 4/10
13. 222/300
14. 200/300
15. 86/175
16. 86/122
17. 162/300
18. 41/300
19. 1.78
20. 1.16
21. 0.72
22. 0.4707



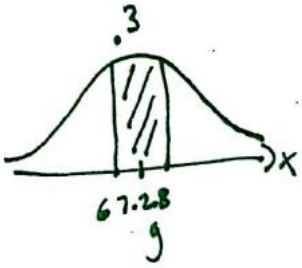
23. 0.4070



24. 68.31 g



25. 66.69 g and 67.87 g



Lecture #13: Confidence Intervals for the Mean

Inferential Statistics

Having covered descriptive statistics and then probability, we are now embarking on the most significant part of the course – inferential statistics. There are two main ways in which we make inferences (draw conclusions about populations from samples). The first is **estimation**, which we'll cover this lecture and next, and the second is the **testing of claims**, which we'll do on and off for the rest of the course.

Point Estimate for the Mean

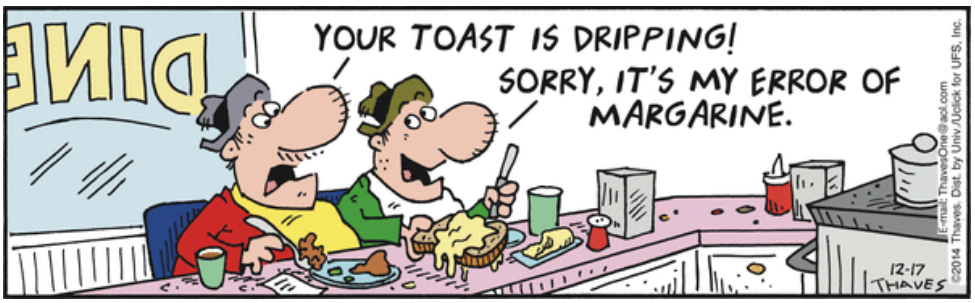
Let's say we're interested in guessing (inferring) the population mean shoe size for Mendocino College students from our Class Data Base. For that group of 120, $\bar{x} \approx 9.171$. What should we guess for μ, the population mean? I don't see how we could do better than 9.171. It's not as though the 120 students in the class data base were chosen for the size of their feet. In fact, to do all of the analyses we're going to make, we **must** assume that the Class Data Base represents a random sample of Mendocino College students.

When we guess a parameter (μ) by picking a single number like $\bar{x} \approx 9.171$, we are making what's called a **point estimate**. So we say that $\bar{x}$ is the point estimate for μ. If we're not going to use $\bar{x}$ (because maybe we have some other, better information), why did we bother to collect the data in the first place?

The Margin of Error

But what if some other group of 120 made up the Class Data Base? It's very unlikely that this group would have a sample mean of 9.171, although we would expect this mean to be reasonably close to 9.171. Think of a sampling distribution with samples of size 120, each having its own sample mean. Our Class Data Base happens to contain one such sample.

So what we do, instead of limiting ourselves to a point estimate, is to create what's called a **confidence interval** by employing a number called ***E***, and saying that we are **confident** to a certain degree (which we will go into in great detail) that the population mean falls between $\bar{x} - E$ and $\bar{x} + E$. We could also use the compound inequality notation of algebra: $\bar{x} - E < \mu < \bar{x} + E$. This expresses the idea that μ has been captured, or pinpointed, **between** the two extremes of $\bar{x} - E$ and $\bar{x} + E$. (Another way to express this is to say that μ is contained in $\bar{x} \pm E$. This has the advantage of including the values of the sample mean and the margin of error, but it lacks the limits of the interval.)



E is called the **maximum error of estimate** or the **margin of error** and is sometimes abbreviated *ME*. In a bit I'll give you a formula for the margin of error, but first I'd like to try to give you a feel for it and the concept of confidence.

What if I just pick 5 as the margin of error? In the shoe size example, that would mean $9.171 - 5 < \mu < 9.171 + 5$, or $4.171 < \mu < 14.171$. In other words, the average shoe size of Mendocino College students is somewhere between a 4½ and a 14. How confident am I that this statement is true? One hundred percent confident! I would stake my life on it, if need be. But the statement is of absolutely no use. It gives no information. Let's say you were marketing a new kind of slipper with college logos, and you wanted to know what sizes to make. The population mean might turn out to be 6, or 13.

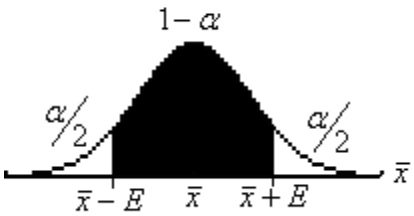
I'd like to use a smaller *E*. How about 0.2? That would mean $9.171 - 0.2 < \mu < 9.171 + 0.2$, or $8.971 < \mu < 9.371$. Much better! This interval contains only one actual shoe size, 9. So I would produce a lot of size 9 slippers and some in sizes slightly larger and smaller than 9. But how confident could I be that I had really captured μ? Not anywhere near as confident as I was about the wider interval. I hope you get the sense that the narrower the interval (the smaller the margin of error, *E*) the more useful it is, but the less confident you become that you've actually captured the population mean. It's a trade-off. Do you want to be surer (or at least less unsure) that you've bracketed μ? Widen the interval; increase *E*.

To get a formula for *E*, we need to develop some new notation. We call the number that expresses our confidence that we captured the mean the **confidence level**, or CL. We can pick this level. There are three levels that are usually used: 99%, if the matter is really important and we want to be quite certain; 90%, if it's not a big deal; 95% if it's in between. Can you see that *E* will be largest for a 99% CL and smallest for 90%?

Now subtract CL from 1, or 100%. For the 99% confidence level, this would give $1 - 0.99 = 0.01$. We call this α, the Greek 'a,' pronounced alpha. You'll be seeing a lot of α. For the 95% confidence level it would be $1 - 0.95 = 0.05$, for 90% $1 - 0.90 = 0.10$. Rearranging the equation, we see that the confidence level CL is actually $1 - \alpha$.

Think now of the Central Limit Theorem, which says that the sampling distribution of samples of size 120 will be normal, with the same mean as the parent population. I want to go left and right from this mean a certain distance along the $\bar{x}$ axis

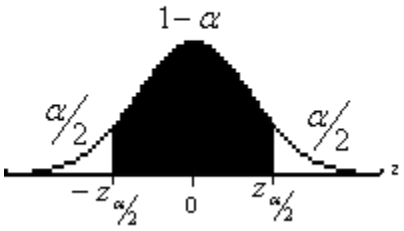
so that the area under the normal curve is equal to the confidence level. That leaves α of the area outside my limits. Since the curve is symmetrical, that gives an area of $\alpha/2$ under the curve on either side of the limits:



(If you're very perceptive, you might notice that we're playing fast and loose with $\bar{x}$ and μ here, but it's better if you just accept it.)

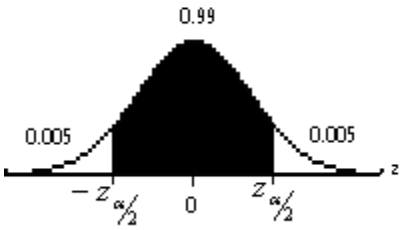
At this point I have to introduce you to a certain special normal distribution. We call it the *z*-distribution, or the **standard normal distribution**. It has a mean of 0 and a standard deviation of 1. It's the first of the four great statistical distributions, or pdfs, that we'll encounter in this course. You've been able to avoid it up until now because we used a calculator to find probabilities and cut-offs for normal distributions. Of course you **have** seen the letter *z* used in the *z*-scores or standard scores introduced in Lecture #6. The standard normal distribution is centered around 0, so half its values are negative and half are positive. We now give a special name to the limits of the confidence interval in the *z* distribution. We call them $z_{\alpha/2}$ and $-z_{\alpha/2}$. They are the cut-offs in the *z* distribution for the middle CL = $1-\alpha$ of the area.

Let me redo the figure above for the *z*-distribution



In order to get a formula for *E*, we need to be able to calculate $z_{\alpha/2}$. We'll do this for our three confidence levels, 90%, 95% and 99%, but once having done this we will be able just to use the three values of $z_{\alpha/2}$ whenever the need arises.

Let's start with the 99% confidence level, for which $\alpha = 1-0.99 = 0.01$, and thus $\alpha/2 = \dfrac{0.01}{2} = 0.005$. Here's the diagram:



We can calculate $-z_{\alpha/2}$ by using invNorm (0.005, 0, 1) $\approx$ –2.5758, or -2.576. Of course $z_{\alpha/2}$ = invNorm (0.995, 0, 1) $\approx$ 2.5758, or 2.576. Either way, for a confidence level of 99%, we will use 2.576 for $z_{\alpha/2}$.

Similar calculations (invNorm (0.975, 0, 1) for a 95% confidence level, and invNorm (0.95, 0, 1) for a 90% confidence level) yield values for $z_{\alpha/2}$ of 1.960 and 1.645. These $z_{\alpha/2}$ values don't change; they're always the same.

Here's a table that summarizes what we've done:

Confidence Level	α	$\alpha/2$	$z_{\alpha/2}$
99%	0.01	0.005	2.576
95%	0.05	0.025	1.960
90%	0.10	0.05	1.645

In practice, all you need at your disposal are the first and last columns:

Confidence Level	$z_{\alpha/2}$
99%	2.576
95%	1.960
90%	1.645

Now at long last you're ready for the formula for *E*. Doing some mathematical rearranging, which, believe me, you don't want to see, we get $E = z_{\alpha/2}\sigma_{\bar{x}}$, where $\sigma_{\bar{x}}$ is the standard error of the mean, from Lecture #12, pertaining to the sampling distribution of the $\bar{x}$'s. The formula for the sampling error was $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. So $E = z_{\alpha/2}\frac{\sigma_x}{\sqrt{n}}$. Since σ_x is the population standard deviation of the parent population, which of course we don't

know (if we knew that we'd most likely know the population mean μ and wouldn't have to estimate it!), so we'll use *s*, the sample standard deviation, instead, which we can do if *n* is big enough, because of the Central Limit Theorem. This gives us our **final** version of *E*:

$$E = z_{\alpha/2} \frac{s}{\sqrt{n}}.$$

(It's final if we're estimating the **mean**. When we're estimating some other parameter, as you'll see in the next lecture, there's a different formula.)

Constructing Confidence Intervals

Using the formula, let's now make a 99% confidence interval for the mean shoe size of Mendocino College students. Remember, its form is $\bar{x} - E < \mu < \bar{x} + E$. In our case, this becomes $9.171 - 2.576 \cdot \frac{1.987}{\sqrt{120}} < \mu < 9.171 + 2.576 \cdot \frac{1.987}{\sqrt{120}}$. *E* itself comes out to $2.576 \cdot \frac{1.987}{\sqrt{120}} \approx 0.467$, and when the dust settles we have $8.704 < \mu < 9.638$, and that's the answer, the 99% confidence interval for the true population mean shoe size of Mendocino College students. As you can see, this interval contains two actual shoe sizes, 9 and 9½.

The calculator is a far easier way of constructing confidence intervals. We use TInterval. Here are the results for the three levels of confidence:

99%: $8.696 < \mu < 9.646$
95%: $8.812 < \mu < 9.530$
90%: $8.870 < \mu < 9.472$

The discrepancy between what we got when using the formula for *E* comes because we didn't actually use the *z*-distribution in TInterval, but rather the *t*-distribution. It is a distribution similar in shape to the *z*-distribution but easier to use and yielding results very close to what we get with the *z*-distribution, but erring on the side of caution. The *t*-distribution is the second of the four great distributions of statistics. Here's a brief explanation of what the *t*-distribution actually is.

The normal distribution covers situations where the distribution is normal (obviously) and we know the population standard deviation, σ. Or, if we don't know σ (and really, how could we?), at least we're dealing with a sampling distribution of size 30 or more, in which case we can use the sample standard deviation for σ.

If these conditions aren't fulfilled, we can use a distribution called the ***t*-distribution**, which was developed in the early 20th century by a person working for the Guinness Brewery in Dublin. To use the *t*-distribution, the population must be more or less normally distributed, the sample size doesn't have to be 30 or more, and we just get to use *s* as σ. The shape of the *t*-distribution depends on the sample size *n*, and the

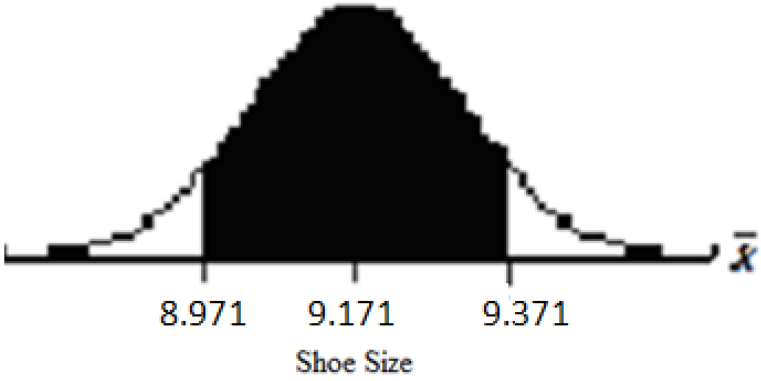
larger *n* is the more like the *z*- the *t*-distribution looks. The *t*-distribution is always a little flatter than the *z*-distribution: its tails are higher and its middle is lower.

To characterize a particular *t*-distribution, we refer, not to its sample size *n*, but to its **degrees of freedom**, which is one **less** than the sample size, or $n-1$. The concept of degrees of freedom will recur a few more times in this course. Basically, in this case, if you have *n* numbers and you know their sum, you're **free** to pick $n-1$ of them, as long as they don't add up to more than the sum, but when it comes to the last number you have no choice, because it's whatever is needed to make the agreed-upon sum.

Here is the pdf for a *t*-distribution with d.f. = 3 (three degrees of freedom and thus a sample size of only 4) along with the pdf for the *z*-distribution:

The *t* is the one that's higher on the sides and lower in the middle. The effect of this is that using the *t* in generating a confidence interval will give a wider one than using the *z*, because there is more area in the tails of the *t*-distribution.

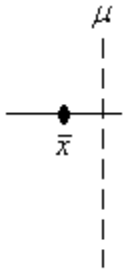
Notice again that the higher the level of confidence the wider the interval. As the level of confidence decreases, the range of possible means narrows. That reminds me of our made-up margin of error, 0.2, which made an interval containing only one size. What level of confidence did that *E* come from? This question turns out to be pretty easy. Look at the diagram:



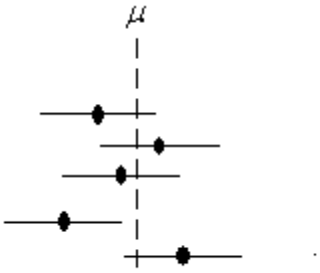
The confidence level is the shaded area, or normalcdf (8.971, 9.371, 9.171, $1.987/\sqrt{102} \approx 0.7298$. So the confidence level for this margin of error is about 73%, too low to be of any use.

What exactly does it mean to say that we are, for instance, 99% confident that the true mean shoe size of Mendocino College students is between 8.704 and 9.638, or 9.171 ± 0.467?

The population mean exists; we just don't know what it is. We take a sample, find its mean, and go down from it and up from it (left and right actually) a certain distance, *E*, to form an interval. This diagram shows one such interval:



This interval **did** include the population mean μ. Let's say we generate other samples of the same size and make intervals with them:



Some of the intervals contain μ, and some, like the fourth one down, don't. If you are generating a 99% confidence interval, this means that on the average 99 out of 100 such intervals **will** contain μ, and one won't. Of course, you have no idea if the one you generated got lucky or not. But you **do** know that the higher the level of confidence you use, the more likely it is that your interval does contain μ. And even if your interval missed μ, it probably didn't miss it by much.

The diagram also illustrates the reason why a higher confidence level means a wider interval: the longer the horizontal line, the more likely it is that the line crosses the dotted line for μ.

Sample Size

There are certainly reasons we might want to limit the margin of error to a certain size, as illustrated by our example, in order to pinpoint the population mean. But how can we do this without lowering the confidence level and hence decreasing $z_{\alpha/2}$? Our sample standard deviation, *s*, is what it is and can't be shrunk. Where else in the formula $E = z_{\alpha/2}\frac{s}{\sqrt{n}}$ can we look for relief? Clearly, our only other hope is *n*, the sample size. Since *n* is in the denominator, increasing it will decrease *E*. This leads us to the **sample size formula for estimating the mean.** In it, we know the confidence level and the

sample standard deviation, and we want to know what sample size will produce a confidence interval with the desired margin of error.

Here's a little algebraic manipulation, turning the formula $E = z_{\alpha/2}\frac{s}{\sqrt{n}}$, currently solved for *E*, into an equivalent formula solved for *n*:

Square both sides: $E^2 = \left(z_{\alpha/2}\frac{s}{\sqrt{n}}\right)^2$

giving $E^2 = \frac{\left(z_{\alpha/2}\cdot s\right)^2}{n}$.

Multiply both sides by $\frac{n}{E^2}$: $\frac{n}{E^2}\cdot E^2 = \frac{n}{E^2}\cdot\frac{\left(z_{\alpha/2}\cdot s\right)^2}{n}$

and simplify. Presto: $n = \left(\frac{z_{\alpha/2}\cdot s}{E}\right)^2$

For example, what if we want to estimate the population mean shoe size of Mendocino College students within half a size at the 95% confidence level?

With $E = 0.5$ the sample size formula $n = \left(\frac{z_{\alpha/2}\cdot s}{E}\right)^2$ then becomes $n = \left(\frac{1.960\cdot 1.987}{0.5}\right)^2 \approx 60.669$. We can't include 0.669 of a person, so we round our answer to 61. A sample size of at least 61 will ensure that we estimate the mean within half a size at the 95% confidence level.

Even when our sample size answer wouldn't normally round up, we do so anyway, because the *n* we get is the absolutely smallest number that will fulfill the condition of estimating within a certain margin of error at a certain level of confidence. If we were to round down we would be below this limit.

Activity #13: Finding confidence intervals for means

Use the height data from the Class Data Base to answer these questions.

1. Find the 99% confidence interval for the mean height of Mendocino College students. Round heights to the nearest tenth.

2. Find the 95% confidence interval for the mean height of Mendocino College students. Round heights to the nearest tenth.

3. Find the 90% confidence interval for the mean height of Mendocino College students. Round heights to the nearest tenth.

In #4-6, round the standard deviation to the nearest thousandth when using it in the sample-size formula.

4. Find the sample size needed to estimate the mean height of Mendocino College students within one-half inch at the 99% confidence level.

5. Find the sample size needed to estimate the mean height of Mendocino College students within one-half inch at the 95% confidence level.

6. Find the sample size needed to estimate the mean height of Mendocino College students within one-half inch at the 90% confidence level.

Lecture #14: Confidence Intervals for the Proportion

In the last lecture we covered estimating a population mean, μ, from a sample, first using a point estimate, $\bar{x}$, and then generating an interval, $\bar{x} - E < \mu < \bar{x} + E$, which we could state with a certain level of confidence contains the population mean. This time we'll deal with estimating a different parameter, called the **population proportion**, *p*.

When we have a quantitative data set at the interval or ratio level of measurement, we have a set of numbers for which we can calculate a mean and a standard deviation. But when we have a binomial variable, like a 'yes' or 'no,' or a 'male' or 'female,' the only thing we can calculate is the fraction of the set that are in one of these categories or the other.

Whichever of these categories we choose to concentrate on, we can look at a data set and calculate the **sample proportion** for this category. Let's say we're interested in the fraction of Mendocino College students who are teenagers. From the Class Data Base, we find that 52 out of 120 are teenagers. We call 52, the number in our category (teenagers), *x*. As usual, the size of the set is *n*. The fraction $\frac{x}{n}$, or $\frac{52}{120}$ in this case, is the sample proportion, and its symbol is $\hat{p}$, pronounced 'p-hat.' So the formula for the sample proportion is $\hat{p} = \frac{x}{n}$.

Point Estimate for the Population Proportion

When we use our sample to make inferences about the population, we begin, as we did in the case of the population mean, with a **point estimate**. Not surprisingly we use the sample proportion $\hat{p}$ as the estimate for the population proportion *p*. So our estimate for the proportion of Mendocino College students who are teenagers is $\frac{52}{120} \approx 0.433$. (If necessary, we round these proportions to the nearest thousandth, or tenth of a percent.)

The Margin of Error

But, just as with the mean, we have to consider what would happen if we chose a different sample of 120 Mendocino College students. Probably it would have a different number of teenagers, not exactly 52. So we want to go down from our point estimate $\hat{p}$ and up from it a certain amount, in an attempt to capture the population proportion *p*. We call this amount, just like with the mean, the **margin of error**, *E*. We use it to generate a confidence interval, which in this case will look like this: $\hat{p} - E < p < \hat{p} + E$. Take careful note of the parameter in the middle: it's *p*, not μ, which we are estimating here.

The formula we used for *E* in estimating the mean simply won't work here. There's no $\bar{x}$, and hence no *s*. Before I reveal it, there's one more symbol you have to

understand: $\hat{q}$ (q-hat). It's the fraction of the data set that is the **other** category than the one we used for $\hat{p}$. Here it's the fraction of the Class Data Base that **aren't** teenagers. If 52 out of 120 are teenagers, then $120 - 52 = 68$ aren't. The formula for $\hat{q}$ is $\hat{q} = 1 - \hat{p}$, or $\hat{q} = \frac{n-x}{n}$. The two formulas are the same because $\frac{n-x}{n} = \frac{n}{n} - \frac{x}{n} = 1 - \hat{p}$. So in this case $\hat{q} = 1 - \frac{52}{120} = \frac{120-52}{120} = \frac{68}{120} \approx 0.567$.

So here's the formula for the margin of error when estimating population proportions: $E = z_{\alpha/2}\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$. The $z_{\alpha/2}$ is the same as we used in estimating the mean:

Confidence Level	$z_{\alpha/2}$
99%	2.576
95%	1.960
90%	1.645

Let's find the margin of error for estimating the population proportion of students who are teenagers at Mendocino College at the 99% confidence level:
$E = z\alpha/_2\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}} = 2.576\sqrt{\frac{0.433 \cdot 0.567}{120}} \approx 0.117.$

Constructing Confidence Intervals

We've already done most of the work for constructing a confidence interval at the 99% confidence level for the population proportion of students who are teenagers at Mendocino College. Here it is: $\hat{p} - E < p < \hat{p} + E \rightarrow 0.433 - 0.117 < p < 0.433 + 0.117$, or $0.316 < p < 0.550$. We can be 99% confident that the percent of Mendocino College students who are teenagers is between 31.6% and 55.0%.

Let's have the calculator do the work now, using 1-PropZInt. Here are the results for our three levels of confidence:

99%: $0.317 < p < 0.550$
95%: $0.345 < p < 0.522$
90%: $0.359 < p < 0.508$

Notice again how the interval narrows as the level of confidence decreases; we estimate more closely but with less confidence that we've actually captured the population proportion.

Looking at Polls

Estimating the population proportion is probably the application of statistics that you are most likely to encounter in everyday life. It shows up all over in the form of opinion polls, in which random samples of people are asked their opinions about candidates and issues. For example, a Gallup poll from Dec. 26, 2017, showed that 56% of those polled disapproved of the job Donald Trump was doing as president.

There were 1500 respondents in the poll, and the reported margin of error (E) is $\pm 3\%$. The confidence level is not stated.

Let's see how this works out, in the formula $E = z_{\alpha/2}\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$. We'll use 0.56 for $\hat{p}$ and 0.44 for $\hat{q}$. Let's start by assuming that the 95% confidence level was used:

$$E = 1.960\sqrt{\frac{0.56 \times 0.44}{1500}} \approx 0.025$$

Close but a little smaller than 3%. Could they have used the 99% confidence level?

$$E = 2.576\sqrt{\frac{0.56 \times 0.44}{1500}} \approx 0.033$$

That's a little closer to 3%. But there could be other explanations besides the confidence level. For instance, it might be that not all 1500 respondents were counted in this question. If *n* were smaller, because *n* is in the denominator *E* would be larger, so the confidence level could have been 95%. I mention this because 95% is the customary confidence level used in polling.

Sample Size

Our goal here is to produce a formula for the sample size, given the desired margin of error, much as we did in the lecture on confidence intervals for the mean. The formula there was $n = \left(\frac{z_{\alpha/2} \cdot s}{E}\right)^2$. That won't work for proportions, because we don't have an *s*, and anyway we developed a different formula for the margin of error for estimating proportions. (Remember, it was rearranging the formula for *E* in estimating means that produced the formula for *n*.)

Perhaps you had a doubt about the formula $n = \left(\frac{z_{\alpha/2} \cdot s}{E} \right)^2$, namely, that if you were setting out to estimate the mean, how on earth would you already know the standard deviation of the sample whose size you are now determining? This is a valid objection; the books just say that you'd know it from other studies of the variable whose mean you want to estimate. This seems kind of glib.

The good news is that in estimating the population proportion, we can do better. It appears that if we're going to rearrange the formula $E = z_{\alpha/2}\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$ to solve for *n*, we'd have to know $\hat{p}$ (and thus $\hat{q}$) before we found the appropriate sample size. But let's take a look at $\hat{p}$ and $\hat{q}$ and more importantly their product $\hat{p} \cdot \hat{q}$:

$\hat{p}$	$\hat{q}$	$\hat{p} \cdot \hat{q}$
0.1	0.9	0.09
0.2	0.8	0.16
0.3	0.7	0.21
0.4	0.6	0.24
0.5	0.5	0.25
0.6	0.4	0.24
0.7	0.3	0.21
0.8	0.2	0.16
0.9	0.1	0.09

You can see that the largest value $\hat{p} \cdot \hat{q}$ ever takes on is 0.25, which happens when $\hat{p}$ and $\hat{q}$ are both 0.5 (when the sample is evenly divided between the two categories). So if we just replace $\hat{p} \cdot \hat{q}$ by 0.25, we'll be on the safe side – that is, we'll be using a sample that might be bigger than necessary for a certain margin of error, but it will never be smaller than necessary.

Time for a little algebra. First replace $\hat{p} \cdot \hat{q}$ by 0.25 in the formula $E = z_{\alpha/2}\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$: $\quad E = z_{\alpha/2}\sqrt{\frac{0.25}{n}}$.

Then square both sides: $\quad E^2 = z_{\alpha/2}^{\,2} \cdot \frac{0.25}{n}$.

Then multiply both sides by *n*: $\quad n \cdot E^2 = n \cdot z_{\alpha/2}^{\,2} \cdot \frac{0.25}{n} = 0.25 z_{\alpha/2}^{\,2}$.

Finally, divide both sides by E^2: $\quad \frac{n \cdot E^2}{E^2} = \frac{0.25 z_{\alpha/2}^{\,2}}{E^2}$.

So $n = \dfrac{0.25 z_{\alpha/2}^{\ 2}}{E^2}$, or $n = 0.25\left(\dfrac{z_{\alpha/2}}{E}\right)^2$.

The way we've done this formula, so that it doesn't depend on a particular $\hat{p}$, means that the sample size for a given confidence level and a given margin or error will be the same no matter what the binomial variable is – for or against Candidate X, male or female, whatever.

How large a sample should we use if we want to estimate a proportion at the 95% confidence level within 5%? E is 0.05, and $z_{\alpha/2}$ is 1.960. So $n = 0.25\left(\dfrac{1.960}{0.05}\right)^2 = 384.16$. Remember that we have to round up to the next whole number, despite the digit in the ten's place. So the sample size we should use is 385.

Comparing Estimating the Mean and Estimating the Proportion

Here is a side-by-side comparison of the confidence intervals covered in this lecture and the last one:

	Estimating the Mean	**Estimating the Proportion**
Parameter	μ	p
Point Estimate	$\bar{x}$	$\hat{p}$
Confidence Interval	$\bar{x} - E < \mu < \bar{x} + E$	$\hat{p} - E < p < \hat{p} + E$
Calculator	TInterval	1-PropZInt
Formula for *E*	$E = z_{\alpha/2}\dfrac{s_x}{\sqrt{n}}$	$E = z_{\alpha/2}\sqrt{\dfrac{\hat{p}\cdot\hat{q}}{n}}$
Sample Size Formula	$n = \left(\dfrac{z_{\alpha/2}\cdot s}{E}\right)^2$	$n = 0.25\left(\dfrac{z_{\alpha/2}}{E}\right)^2$

Activity #14: Finding Confidence Intervals for Proportions

In Problems 1-3, round answers to the nearest thousandth (tenth of a percent).

1. Find the interval to estimate, at the 99% percent confidence level, the proportion of Mendocino College students who are at least 24 years old.

2. Find the interval to estimate, at the 95% percent confidence level, the proportion of Mendocino College students who are at least 24 years old.

3. Find the interval to estimate, at the 90% percent confidence level, the proportion of Mendocino College students who are at least 24 years old.

4. What is the minimum sample size needed to estimate a proportion at the 90% confidence level within 7%?

Assignment #6

Pretend that the Class Data Base is a random sample of Mendocino College students. Round answers in Problems 1-3 to the nearest hundredth.

1) Find
   a) a 99% confidence interval
   b) a 95% confidence interval, and
   c) a 90% confidence interval
   for the average height of Mendocino College students.

2) Repeat #1 for the average age of Mendocino College students

3) Repeat #1 for the average number of pets owned by Mendocino College students.

In Problems 4-6, use *s* to the nearest thousandth.

4) Find the minimum sample size needed to be 90% confident of estimating the average height of Mendocino College students accurately within a quarter of an inch?

5) Find the minimum sample size needed to be 95% confident of estimating the mean age of Mendocino College students accurately within 6 months?

6) Find the minimum sample size needed to be 99% confident of estimating the average number of pets owned by Mendocino College students accurately within 0.4 pets?

Round answers in Problems 7 and 8 to the nearest thousandth.

7) Find
   a) a 99% confidence interval,
   b) a 95% confidence interval, and
   c) a 90% confidence interval
   for the true proportion of Mendocino College students who are natives of Mendocino or Lake County.

8) Repeat #7 for the true proportion of Mendocino College students who wear shoes that are at least size 10.

9) Find the minimum sample size needed for estimating, at the 90% confidence level, a proportion with an accuracy of $\pm 4\%$.

10) Find the minimum sample size needed for estimating, at the 95% confidence level, a proportion within 0.02.

11)Find the minimum sample size needed for estimating, at the 99% confidence level, a proportion within 3%.

Lecture #15: Introduction to Testing Claims

We've covered one use of inferential statistics – estimating parameters of populations using statistics from samples. We did this in two ways: first using a point estimate, a single guess, and then generating confidence intervals, ranges of guesses.

The other common use of inferential statistics is **testing claims**, also sometimes referred to **hypothesis testing** or **significance testing**. Whatever it's called, it's used in the social and biological sciences and in industrial and business settings. It's a fairly complicated process, one we'll be using for most of the rest of the course, and I'm going to attempt to break it down for you.

Stating Hypotheses and Labeling the Claim

Let me describe three different situations that someone might be interested in studying:

Situation	The life expectancy of professional athletes	How many miles a brand of tires lasts	The speed with which a new pain reliever takes effect

People could make different guesses about each of these situations. The guess is the **claim**. In the case of the athletes, you might think because they're so fit and so rich that they live longer than the average person, but you might also think that because they use their bodies so roughly and are exposed to all kinds of temptations they might have a shorter life than the average. The tires – well, if you manufactured them you might want to boast about how much better they are (longer they last) than other brands, but if you're a rival the opposite might be the case. The pain reliever – the quicker the better if you're the manufacturer, the slower the better if you have a rival product.

I'm going to pick a position for each of these situations and make a claim:

Situation	The life expectancy of professional athletes	How many miles a brand of tires lasts	The speed with which a new pain reliever takes effect
Claim	Professional athletes have the same life expectancy as other people.	This brand of tires is no better than average.	This pain reliever acts faster than the average painkiller.

Now each claim has to be translated into a mathematical sentence. This lecture and the next one will cover testing claims about the population mean, so the sentences will begin with μ, which will be followed by one of six symbols: $=, <, >, \neq, \leq,$ or $\geq$.

Sometimes it's tricky to see which of the six symbols the claim implies. Not so for the athletes; it's =.

How about the tires? If a tire is no better than average, either it **is** average or it's worse than average – it doesn't go as far before needing to be replaced. The one thing it **doesn't** do is go more miles than the average before needing to be replaced. The symbol we use is $\leq$.

And the pain reliever? Looking at the speed with which it takes effect, if we're saying this one acts faster, we're saying that its time into the bloodstream (which is how we'd measure speed in this case) is **shorter** than average, so the appropriate symbol is <.

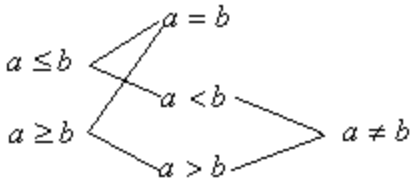
What are we comparing μ to? In each case, we give this a special symbol, μ_0, pronounced 'mu sub zero,' which simply means the baseline average. I'm going to pick a value for μ_0 for each of our three situations.

Let's just say that the mean life expectancy of all people is 79 years, that the average distance that tires can drive before needing to be replaced is 30,000 miles, and that the average time it takes pain relievers to reach the bloodstream is 14.9 minutes. I'm not saying that these numbers are accurate, but we need to use some number for μ_0.

Now I can expand the table to incorporate the claims written as mathematical sentences:

Situation	The life expectancy of professional athletes	How many miles a brand of tires lasts	The speed with which a new pain reliever takes effect
Claim	Professional athletes have the same life expectancy as other people.	This brand of tires is no better than average.	This pain reliever acts faster than the average painkiller.
Claim as a Mathematical Sentence	$\mu = 79$ years	$\mu \leq 30{,}000$ miles	$\mu < 14.9$ minutes

Now I have to explain a perfectly obvious but very important mathematical fact, called the **Trichotomy Property**. You may know that a **di**chotomy is a situation in which there are two choices or alternatives. So a **tri**chotomy has three choices or alternatives. If you have two numbers, *a* and *b*, then **one and only one** of these three alternatives is true: either $a = b$, or $a < b$, or $a > b$. One of these three statements **must** be true for any two numbers *a* and *b*, and the other two **must** be false. But any two of the statements can be combined in one sentence, as this table shows:



In other words, if you want to express the possibility that *a* could equal *b* **or** that *a* could be less than *b*, you write $a \leq b$. This is pronounced "*a* is less than or equal to *b*." (Note that this is an **either/or** statement – you're not saying that *a* equals *b* **and** *a* is also less than *b*. Couldn't happen.) If you want to express the possibility that *a* equals *b* **or** *a* is greater than *b*, you write $a \geq b$, which is read "*a* is greater than or equal to *b*." If you want to express the possibility that *a* is less than *b* **or** that *a* is greater than *b* (but not equal to *b*), you write $a \neq b$, read "*a* is not equal to *b*."

The point is that although the Trichotomy Property presents three alternative relationships for *a* and *b*, three other relationships are implied, and these are the ones in which two possible relationships are combined. Thus we get the six symbols referred to earlier in the lecture: $=, <, >, \neq, \leq,$ and $\geq$.

Having written the claims in mathematical sentences in the last version of the table, we now have to write which of the relationships in the Trichotomy Property have been excluded, and we have to do this in one mathematical sentence, using one of the six symbols between the μ and the μ_0. Sometimes this will involve a sentence with a symbol for **one** possible relationship ($=, <,$ or $>$), and sometimes it will involve a sentence with a symbol for **two** possible relationships ($\neq, \leq,$ or $\geq$). Let's call this sentence What the Claim Leaves Out:

Situation	The life expectancy of professional athletes	How many miles a brand of tires lasts	The speed with which a new pain reliever takes effect
Claim	Professional athletes have the same life expectancy as other people.	This brand of tires is no better than average.	This pain reliever acts faster than the average painkiller.
Claim as a Mathematical Sentence	$\mu = 79$ years	$\mu \leq 30{,}000$ miles	$\mu < 14.9$ minutes
What the Claim Leaves Out	$\mu \neq 79$ years	$\mu > 30{,}000$ miles	$\mu \geq 14.9$ minutes

Make sure you understand why What the Claim Leaves Out is what it is. For instance, in the case of the tires the Claim combines the $a = b$ option and the $a < b$ option, so all that's left is the $a > b$ option, but with the pain reliever the Claim has only one option, $a < b$, so the other two, $a = b$ and $a > b$, have to be combined to make up What the Claim Leaves Out.

Now we're ready for the true business at hand: stating the hypotheses. (I know this seems like a very long process indeed, but on the one hand we've only just started, and on the other hand pretty soon you'll feel like you've been doing this all your life.)

When we test a claim, we have to state **two** hypotheses. They are called the **null hypothesis** (symbol H_0, 'null' being British for 'zero') and the **alternative hypothesis** (symbol H_1, or sometimes H_A -- we'll use H_1). The Claim becomes one of the hypotheses, and What the Claim Leaves Out becomes the other, and **it's extremely important that you learn how to decide which is which!**

Either the Claim or What the Claim Leaves Out will contain the option of μ equaling μ_0 -- in other words it will have either an $=$, a $\leq$, or a $\geq$. Whichever of the two (the Claim or What the Claim Leaves Out) **has** this symbol containing the idea of 'equal' becomes the **null hypothesis,** H_0. Sometimes it's the Claim, and sometimes it's What the Claim Leaves Out, that becomes H_0. The remaining statement, the one containing either a $<$, a $>$, or a $\neq$, by default becomes the **alternative hypothesis**, H_1.

Returning to our three situations, you can see that in the case of life expectancy of professional athletes, the Claim has the equals, so it becomes H_0, and What the Claim Leaves Out is H_1. So, asked to state the hypotheses and label the claim, you write:

H_0: $\mu = 79$ years (Claim)
H_1: $\mu \neq 79$ years

Be sure to put 'Claim' in parentheses after the claim; you'll need to refer to this later on when we actually test a claim. You don't have to write all that information in the table when asked to state the hypotheses and label the claim; just cut to the chase if possible.

How about the tire situation?

H_0: $\mu \leq 30{,}000$ miles (Claim)
H_1: $\mu > 30{,}000$ miles

But don't get the idea that the claim is **always** H_0, because it isn't, as we see from the pain reliever situation:

H_0: $\mu \geq 14.9$ minutes

H_1: $\mu < 14.9$ minutes (Claim)

The alternative hypothesis gives each test one of three names, as follows:

If H_1 contains the symbol <, the test is **left-tailed**. 'Less than' means 'to the left of.'

If H_1 contains the symbol >, the test is **right-tailed**. 'Greater than' means 'to the right of.'

If H_1 contains the symbol $\neq$, the test is **two-tailed**. 'Unequal to' means 'less than' (to the left of) **or** 'greater than' (to the right of).

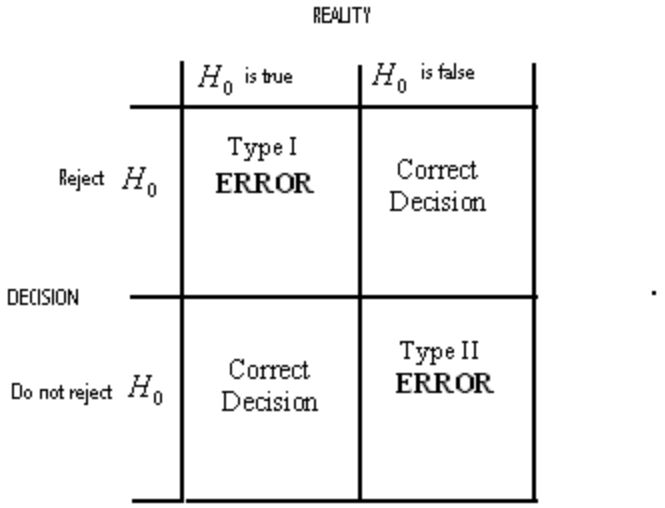
Many words and phrases describing the size relationship between two quantities are straight-forward, such as 'is less than,' 'is not equal to,' and 'is greater than.' Some are more subtle. 'Exceeds' is >. Watch out for 'at least' ($\geq$) and 'at most' ($\leq$). Here are two that imply =: 'hasn't changed from,' 'is the same as.' The following imply $\neq$: 'has changed from,' 'is not the same as,' and 'is different from.'

Deciding What to Do with the Null Hypothesis: Type I and Type II Errors

Translating a claim properly and situating it and what it leaves out correctly into the H_0 / H_1 structure is essential to performing the test of the claim correctly, which is why I've placed so much emphasis on it. The next step will be to look at the sample mean $\bar{x}$ of a random sample from the population about whose mean we want to test a claim and to calculate the probability that a population with mean μ_0 would produce a sample with sample mean $\bar{x}$ or a value even more different than μ_0. We'll do this in the next lecture.

For now we'll investigate what happens after this probability (called the ***p*-value**) is determined. Using the *p*-value, we make a decision. The decision is about the **null hypothesis**. It's **always** about the null hypothesis. It's **not** about the claim, unless the claim **is** the null hypothesis.

Not only is the decision always about the null hypothesis, there are only two possible decisions that you can make: reject the null hypothesis, or fail to reject it. The reason you'll make one or the other of these decisions is based on the *p*-value described above, but let's assume that you've made a decision and see what happens then. The decision could be correct or could be wrong, depending on the actual truth or falsehood of the null hypothesis. There are four possible scenarios when we look at the decision and at reality, neatly conveyed in this widely-used table:



The correct decisions are in the upper right cell, where we correctly rejected a false null hypothesis, and in the lower left cell, where we correctly failed to reject a true null hypothesis. These are good situations (although we can never be sure we have achieved one of them, because we **don't** in reality know the mean of the population, which is why we're testing claims about it).

We want to distinguish between the two kinds of errors, because they have vastly different implications, so we call them Type I and Type II. The error in the upper left cell occurs when we incorrectly reject a **true** null hypothesis. This is a **Type I error**. The error in the lower right cell occurs when we incorrectly fail to reject a **false** null hypothesis. This is a **Type II error**.

Let's take a specific claim, state the hypotheses, and say what would constitute making a Type I and a Type II error in this case. Let's claim that eating sardines for breakfast the day of a test will increase scores on the test. Because we're talking about increasing, the claim translates as $\mu > \mu_0$; in other words our new post-sardine mean is larger than our old pre-sardine mean. The claim, lacking the equals component, is thus H_1. What the Claim Leaves Out is thus $\mu \leq \mu_0$, which becomes H_0. Thus stating the hypotheses and identifying the claims comes out:

H_0: $\mu \leq \mu_0$
H_1: $\mu > \mu_0$ (Claim)

Written in words, this becomes

H_0: Eating sardines before a test doesn't raise test scores.
H_1: Eating sardines before a test raises test scores. (Claim)

A Type I error will mean **deciding** that eating sardines raises test scores when **in fact** it doesn't.

A Type II error will mean **deciding** that eating sardines doesn't raise test scores when **in fact** it does.

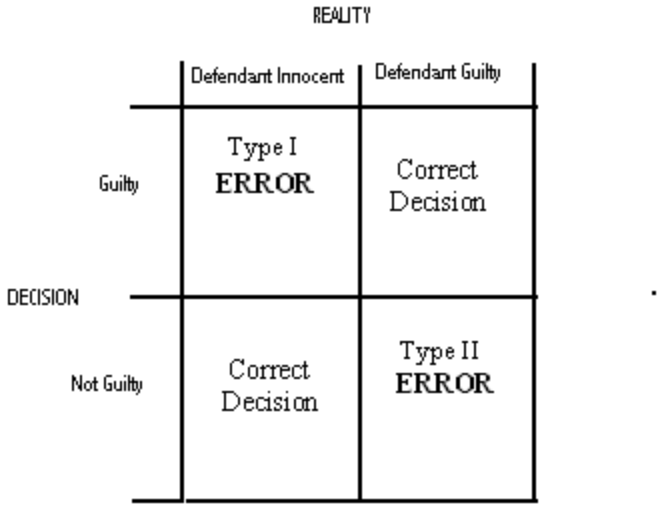
Back to the general discussion. Do you see that the fewer Type I errors you make the more Type II errors you'll make and *vice versa*? If you're the kind of person who goes around willy-nilly rejecting null hypotheses, you'll wind up rejecting a lot of true ones, but you'll seldom fail to reject a false one because you're so prone to rejecting them. But if you're the cautious kind who is very reluctant to reject a null hypothesis in case it's true, you'll fail to reject a lot of false ones, but you'll seldom reject a true one because you don't reject much of anything.

The probability that we made a Type I error in deciding about the null hypothesis in testing a certain claim is what we called the *p*-value, and we compare it to a quantity called α (the same old alpha we encountered in creating confidence intervals), which in this context is called the **significance level**, and represents the largest probability we're willing to risk of making a Type I error. We want to keep a tight lid on α. We don't want to make Type I errors! We would rather make Type II errors, if we have to make an error, and of course the more strictly we limit the likelihood of making Type I errors, the more probable it becomes that we're making Type II errors.

Well, we have to choose which kind of error to avoid more conscientiously, and the fact that it's the Type I error reflects the caution and modesty of the scientific method. In research, we're usually claiming that something has made a difference, either increasing or decreasing a quantity, or at least changing it. Thus the claim is most likely to be the alternative hypothesis. Rejecting the null hypothesis in these cases means accepting the alternative hypothesis (our claim), because of course the null and alternative hypotheses are complements of each other – if one is true the other isn't, and *vice versa*. So a Type I error, rejecting a true null hypothesis, usually means that we're making an unsubstantiated claim. We don't want to do that very often. We'd rather fail to support a true claim than support a false one.

Here's an analogy to our criminal justice system, in which a person is considered innocent until proven guilty (beyond reasonable doubt). This will also give you insight into why, in making the decision about the null hypothesis, we don't simply say 'reject the null hypothesis' or '**accept** the null hypothesis,' instead of 'reject the null hypothesis' or '**do not reject**' the null hypothesis.' Why the double negative?

In a jury trial, the jury is asked to make a decision, based on evidence, about whether the defendant committed the crime. In reality, the defendant either did or didn't commit the crime, but if reality were known there would be no need for a trial. Look at this revised table:



Let's look at each of the four cells. The upper right cell means that a guilty person was found guilty, so justice was done assuming the law was a just one. The lower left cell means that an innocent person escaped unjust punishment. These are good situations.

The lower right cell means a guilty person was not punished. Too bad for the victim, if it was a crime with a victim. The upper left cell means an innocent person is unjustly punished, maybe even killed. We want to avoid this last injustice as much as possible, and so our system calls for the presumption of innocence, even if the standards of 'beyond a reasonable doubt' mean that criminals go unpunished. We can't control both of these mistakes at the same time, and we'd rather have criminals go unpunished than have innocent people unjustly deprived of liberty and maybe life. At least that's the theory. And how about that double negative in 'do not reject the null hypothesis?' You've seen movies and TV. When the jury comes back and is asked its verdict, the verdict is stated either 'guilty' or 'not guilty,' rather than either 'guilty' or 'innocent.'

Activity #15: Stating hypotheses

State the null and the alternative hypothesis and identify the claim:

1) The average age of taxi drivers in New York City is 36.3 years.

2) The average income of nurses has changed from what it was in 1995, $36,250.

3) The average disk jockey is older than 27.6 years.

4) The average pulse rate of female joggers is not less than 72 beats per minute.

5) The average bowling score of people who enrolled in a basic bowling class is below 100.

6) The average cost of a DVD player is $297.75.

7) The average electric bill for residents of White Pine Estates exceeds $52.98 per month.

8) The average number of calories in Brand A's low-calorie meals is at most 300.

9) The average weight loss of people who use Brand A's low-calorie meals for 6 weeks is at least 3.6 pounds.

Lecture #16: Testing Claims about the Mean

This time we're actually going to test claims from start to finish, incorporating what we covered in the last lecture and filling in all the missing steps. The first test will take a lot of explanation, but afterwards things will go much more quickly.

Testing the First Claim

Here's the first claim: The average Mendocino College student is at least 26 years old.

When we make a claim, we have to specify just how willing we are to risk making a Type I error, rejecting a true null hypothesis, as explained in the last lecture. We have to pick a **significance level**, α, to accompany the test. We use one of the three α's that were associated with the 99%, 95%, and 90% confidence intervals, namely 1%, 5%, or 10%, depending upon how strongly we feel about not committing a Type I error. Let's use 10%, $\alpha = 0.10$, for this claim.

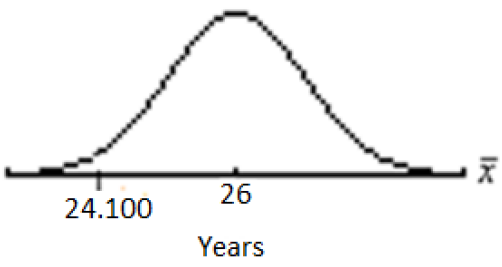
The first step is to state the hypotheses and identify the claim. Remember that 'at least' means greater than or equal to, so the claim translates to "$\mu \geq 26$ years." This is clearly the null hypothesis, H_0, because it contains the equals sign. Thus the alternative hypothesis, H_1, has only the 'less than' option and is "$\mu < 26$ years." Here's the first step:

H_0: $\mu \geq 26$ years (Claim)
H_1: $\mu < 26$ years

Clearly this is a left-tailed test, since 'less than' means 'to the left of.' The 26 years is the μ_0, the alleged parameter of the population.

To investigate the claim, we look at the age data from the Class Data Base. We find that for this sample, the sample mean, $\bar{x}$, is 24.100 years.

We combine all this information in a diagram of a normal distribution, which we are allowed to do even if age isn't normally distributed (and it isn't), because we're looking at a sampling distribution of the sample means of samples in which n = 120:



We are **assuming** that μ_0, 26 years, is in fact the population mean, and we're going to find out how likely it is, with this assumption, that a sample of size 120 would yield a sample mean as much smaller than 26 years as 24.100 years is, or even smaller. If that likelihood is very small, we'll decide that 26 years wasn't the population mean, because if it were our sample mean would be a very unlikely result. This kind of backwards reasoning is called **indirect proof** in mathematics.

The number we find to describe this likelihood is called the ***p*-value** (also known just as *p*). It is the probability that by rejecting the null hypothesis we would be making a Type I error, i.e. rejecting a true null hypothesis. We will then compare the *p*-value to the level of significance we've selected to see if we're willing to take that amount of risk of making a Type I error.

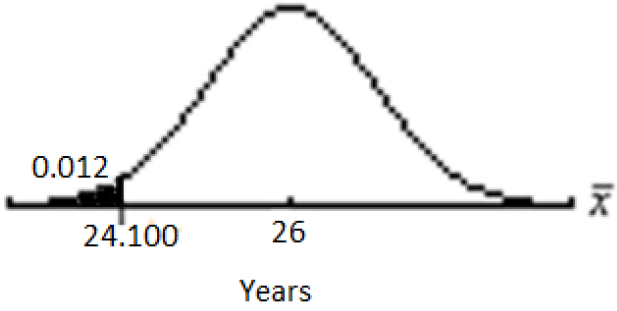
The *p*-value is actually a shaded part under the normal pdf we drew above. It is the shaded part to the left of (because we're doing a left-tailed test) the $\bar{x}$, 24.100 years. Here is the diagram with the area which is the *p*-value shaded:

What is the *p*-value equal to in this case? We already have the means to find this out. We use normalcdf, remembering that in a sampling distribution we must divide the standard deviation by the square root of the sample size to get the standard deviation (standard error of the mean) of the sampling distribution:

Normalcdf (0, 24.100, 26, $^{9.194}/_{\sqrt{120}}$) ≈ 0.01179

(We used the figure 9.194 because it is the sample standard deviation, *s*, of our sample.)

Now we label the shaded area which represents the *p*-value with what we found it to equal, to the nearest thousandth:



And that's our diagram. In practice, you can draw a single diagram like this last one for each claim you test.

Having calculated the *p*-value to be 0.012, we're in the position of making a decision about the null hypothesis. As discussed in the last lecture, there are only two possible decisions: reject the null hypothesis, or don't reject it. We reject it if the *p*-value is low enough that we don't mind taking that much of a risk of making a Type I error. We determine this by comparing the *p*-value to our significance level, α. If the *p*-value is **less than** the significance level, we can live with the risk, and we decide to reject the null hypothesis. If the *p*-value is **greater than** the significance level, there's an unacceptable level of risk, and we decide not to reject the null hypothesis. In short:

If $p < \alpha$, reject H_0.

If $p > \alpha$, don't reject H_0.

(Are you enough tuned in to the Trichotomy Property to wonder what happens if *p* **equals** α? No need to worry, because it never does. The *p*-value is a decimal that goes on far beyond the hundredths place.)

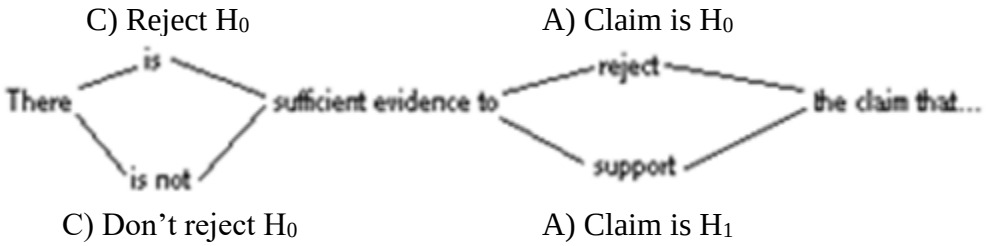
For our claim, since 0.012 < 0.10, or $p < \alpha$, we in fact decide to reject H_0.

You might think we're done testing, but how would you feel if, after someone tells you they're testing the claim that the average Mendocino College student is at least 26 years old, and you ask them how their research came out, they answer, "Oh, we rejected the null hypothesis"? You'd feel cheated and not at all sure about the outcome. Your response would probably be along the lines of, "Say what?"

So the final step in testing a claim is to state a conclusion in plain English, or to summarize the results. To do this, it's better if you abandon all attempts at creativity and simply follow these instructions. First let me give them in words, and then I'll present a couple of little diagrams which sum them up.

Begin with the word, "There". Follow with either "is" or "is not". Use "is" if you rejected the null hypothesis. Use "is not" if you didn't reject the null hypothesis. Regardless, then put "sufficient evidence to". Follow with "reject" if the claim turned out to be the null hypothesis (remember that you've labeled the claim and can just look back at the first step and see), or with "support" if the claim was the alternative hypothesis. Either way, then put "the claim that" and finish up by restating the claim.

Here's the sequence:



The ***is/is not*** part is determined by the decision of what to do about the null hypothesis, in other words by the size relation between the *p*-value and α. The ***reject/support*** part is determined by which hypothesis the claim was. If the claim is the null hypothesis, we either **reject** it or don't **reject** it. If the claim is the alternative hypothesis, we either **support** it or don't **support** it.

Here's a nice schematic way to summarize this:

	Reject H_0	Don't reject H_0
The claim is H_0	There **is** sufficient evidence to **reject** the claim that…..	There **isn't** sufficient evidence to **reject** the claim that…..
The claim is H_1	There **is** sufficient evidence to **support** the claim that…..	There **isn't** sufficient evidence to **support** the claim that…..

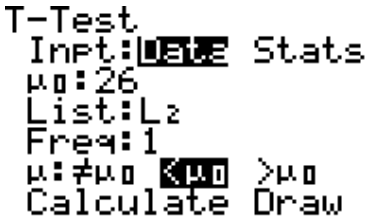
In testing our claim, the claim was the null hypothesis and we rejected the null hypothesis, so we're in the upper left of the four summary boxes: There **is** sufficient evidence to **reject** the claim that the average Mendocino College student is at least 24 years old.

We'll be streamlining this seemingly endless process in a number of ways, chiefly in how we find the *p*-value. We'll use the Stat Tests menu on the calculator and perform a *T*-Test.

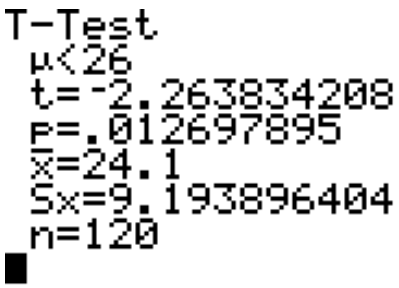
Since the data are in a list we can use the Data setting for Input. We then have to put in the μ_0 so the calculator will know where the distribution is supposed to be centered, and then the name of the list we want to analyze, and finally – most importantly – the form of the alternative hypothesis:

$\mu : \neq \mu_0 \quad < \mu_0 \quad > \mu_0$

You have to pick one of the three possibilities to highlight, and it has to match your alternative hypothesis, or the calculator will be finding some unwanted area for the *p*-value. Here's the *T*-Test screen:

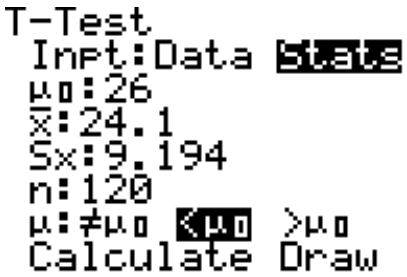


When you pick the Calculate option in the bottom row, you get this screen:

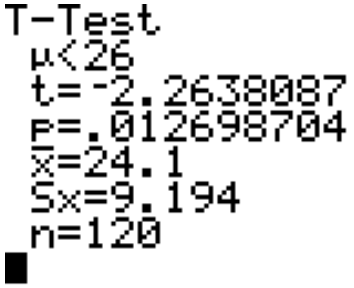


First, it reminds you what you said the alternative hypothesis was ($\mu < 26$ in this case). Don't worry about the next line, "*t* = …", which refers to the other, more cumbersome method of claims testing, which we skipped. The *p* in the next line is the *p*-value, which is what we were after all this time. Notice how close the *p*-value, 0.013, is to the *p*-value we got when we used normalcdf, just a little bigger, as I said it would be.

You can also choose the Stats option for the Input, and if you don't have the data set in a list, you **have** to use this option. You'll be asked to input the μ_0 and the *s*, $\bar{x}$ and *n* of the sample, and finally, the all-important alternative hypothesis:



Upon opting for Calculate, the display looks like this:



Testing the Second Claim

At last, here's the second claim: The average Mendocino College student is taller than 5'5". This time, let's use the 5% significance level ($\alpha = 0.05$).

In testing this claim let's follow the direction line you will use in the future:

STEPS IN TESTING CLAIMS

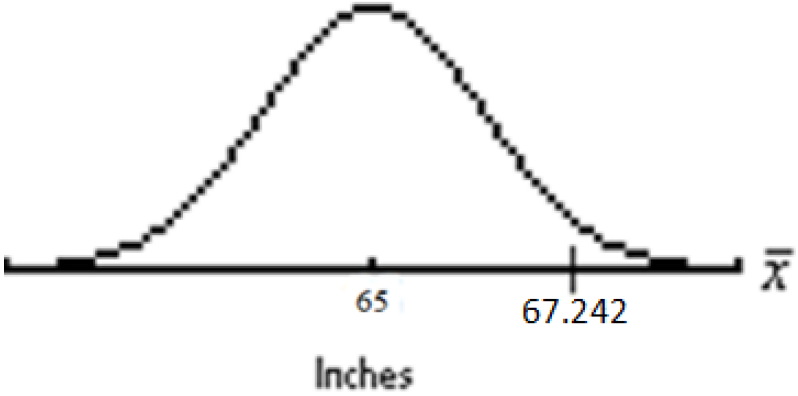
A) State the hypotheses and identify the claim.
B) Perform the test using the sample data or statistics, and make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value, which should be rounded to the nearest thousandth.
C) Decide whether or not to reject the null hypothesis.
D) Summarize the results.

The claim translates easily enough into the mathematical sentence "$\mu > 65$ inches." This is clearly the alternative hypothesis, H_1, because it lacks the equals sign. Thus the null hypothesis, H_0, combines the other two possibilities, the 'is less than' and the 'is equal to,' to become "$\mu \leq 65$ inches." Here's Step A:

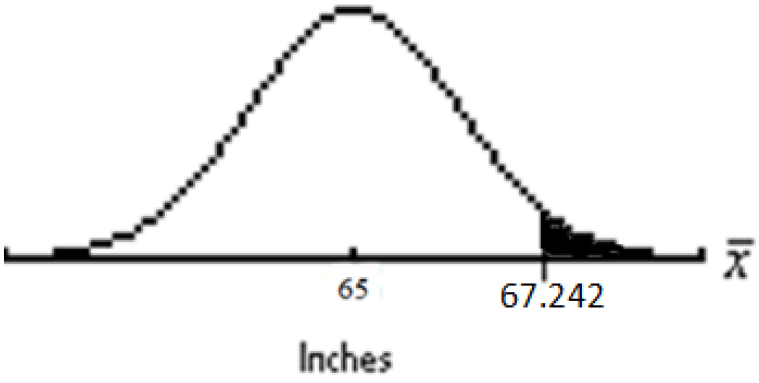
H_0: $\mu \leq 65$ inches
H_1: $\mu > 65$ inches (Claim)

Clearly this is a right-tailed test, since 'greater than' means 'to the right of.' The 65 inches is μ_0.

Performing 1-Variable Stats on the height list gives us the sample mean, 67.242 inches, so we can draw the initial diagram for Step B:



Here's the diagram after we've shaded the *p*-value:



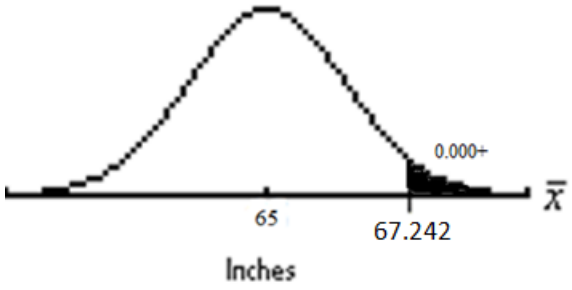
Let's find the *p*-value using *T*-Test:

```
T-Test
 Inpt:Data Stats
 μ0:65
 List:L3
 Freq:1
 μ:≠μ0 <μ0 >μ0
 Calculate Draw
```

Notice that μ_0 has been changed to reflect the new claim and that the alternative hypothesis now reflects the 'greater than' option. Calculating gives this screen:

```
T-Test
 μ>65
 t=6.142297292
 p=5.5544936E-9
 x̄=67.24166667
 Sx=3.99788985
 n=120
```

So what's the *p*-value? One thing it **isn't** is 5.554! Remember, the entire pdf has an area underneath it of 1, and the *p*-value is part of that area. Or, another way of knowing that the *p*-value can't be bigger than 1 is to remember that probabilities can never exceed 1. If you see a *p*-value that **appears** to be bigger than 1, remember to look at the right end of the number, where there will be an E $^-$ some-whole-number-or-other. The *p*-value is actually 5.554×10^{-9}, or 0.000000006. Rounded to the nearest thousandth, we get 0.000 (it's closer to zero thousandths than to one thousandth). But that seems a little harsh – after all there **is** some area there – so we adopt the following notation to show that while the *p*-value didn't even make it to 0.001, it is in fact greater than zero: 0.000^+. Here's the final diagram for Step B:



Since $0.000^{+} < 0.05$, making $p < \alpha$, the answer to Step C is "Reject H_0."

Step D finds us in the lower-left cell of the summary boxes (we **did** reject the null hypothesis, and the claim was the **alternative** hypothesis). So Step D reads, "There **is** sufficient evidence to **support** the claim that the average Mendocino College student is taller than 5'5"." If the claim weren't true, the likelihood of our group of 120 averaging so great a height is almost nil.

Testing the Third Claim

Here's the third claim: The average Mendocino College student has 2 pets. Here let's use the 1% significance level ($\alpha = 0.01$).

A) State the hypotheses and identify the claim.

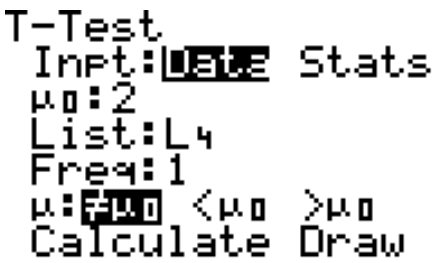
The claim has the 'equals' and is thus the null hypothesis, leaving the 'is less than' and the 'is greater than' to combine to make the alternative hypothesis 'doesn't equal':

H_0: $\mu = 2$ (Claim)

H_1: $\mu \neq 2$

B) Perform the test using the sample data or statistics, and make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value, which should be rounded to the nearest thousandth.

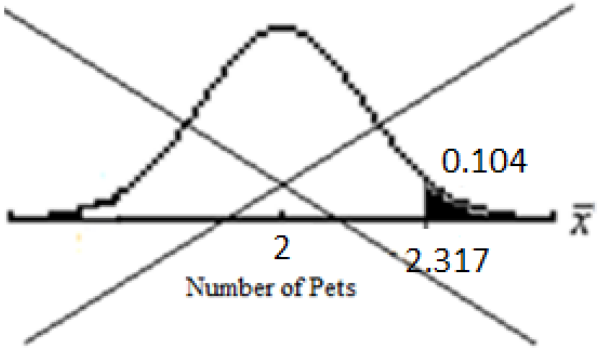
Here's the *T*-Test screen:



and the result:

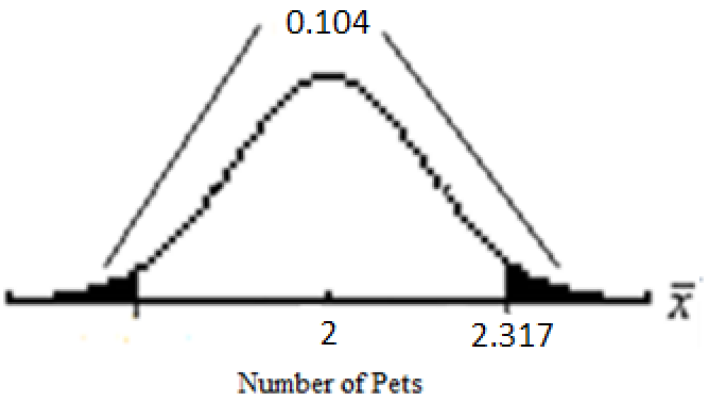
```
T-Test
 μ≠2
 t=1.63760585
 p=.1041456265
 x̄=2.316666667
 Sx=2.118281106
 n=120
```

You might think that the diagram should look like this, but as you can tell by the lines crossing it out, there's something wrong:



What's wrong is that we're doing a **two-tailed** test, and the *p*-value of 0.104 includes the possibility not only that the mean is **greater than** 2 by a certain amount but also the possibility that it's **less than** 2 by the **same** amount. When we claim that the average number of pets is 2, that claim could be refuted by a much larger sample mean or a much smaller one. We got a larger one, but we also have to include the corresponding possibility of our having gotten a smaller one.

So what you do is to shade in a mirror-image of the shaded area on the opposite side of the μ_0. But don't give it a number on the $\bar{x}$ axis, since we didn't get a sample mean less than μ_0. And then, to show that the *p*-value includes both options, write the *p*-value above the center of the pdf and draw lines from it to both of the areas that compose it (each of which by itself would be only $^{0.104}/_{2} = 0.052$).



C) Decide whether or not to reject the null hypothesis.

Since 0.104 > 0.01, or $p > \alpha$, we **do not** reject the null hypothesis.

D) Summarize the results.

The claim was the null hypothesis, and we didn't reject the null hypothesis, so we're in the upper-right rectangle: There **isn't** sufficient evidence to **reject** the claim that the average Mendocino College student owns 2 pets. It could be true, in the sense that about 10% of the time a sample of 120 such numbers of pets would have a sample mean as different from, or even more different from, 2, even if 2 were the true population mean.

Activity #16: Testing claims about the mean

STEPS IN TESTING CLAIMS

A) State the hypotheses and identify the claim.

B) Perform the test using the sample data or statistics, and make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value, which should be rounded to the nearest thousandth.

C) Decide whether or not to reject the null hypothesis.

D) Summarize the results.

1) The average Mendocino College student has a shoe size smaller than 9.5. Use the 5% significance level.

2) The average Mendocino College student is 5' 6". Let $\alpha = 0.10$.

3) The average Mendocino College student is older than 23. Use the 1% significance level.

4) The average Mendocino College student has at most two pets. Let $\alpha = 0.10$.

Assignment #7

STEPS IN TESTING CLAIMS

A) State the hypotheses and identify the claim.
B)Perform the test using the sample data or statistics, and make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value, which should be rounded to the nearest thousandth.
C)Decide whether or not to reject the null hypothesis.
D)Summarize the results.

1) It has been reported that the average annual amount that college seniors spend on cell phones is $1200. The student senate at a large university claims that their seniors spend much less than this, so it conducts a study of 75 randomly selected seniors and finds that the average spent is $1167, with a sample standard deviation of $106. Test the senate's claim using the 1% significance level.

2) Ten years ago, the average acreage of farms in a certain geographic region was 48 acres. A recent study included 25 farms with sample mean 52.5 acres per farm and sample standard deviation 18.9 acres. Test the claim, at the 10% significance level, that the average has changed.

3) The average yield of wine grapes in a certain valley in Northern California is 3000 pounds per acre. A new method of viticulture has been developed and is tested on 50 individual vineyards. The mean yield under the new system is 3128 pounds of grapes per acre with a standard deviation of 574 pounds. At the 5% significance level, test the claim that the average yield has increased.

4) A special cable is said to have a breaking strength of 750 pounds. A researcher selects a sample of 20 cables and finds that the average breaking strength is 735.8 pounds, with a standard deviation of 19.6 pounds. Test the claim that the average breaking strength is 750 pounds at the 1% significance level.

Lecture #17: Testing Claims about the Proportion

Last lecture was the hard part; now you know everything you need to know about testing claims and can apply your knowledge to testing claims about other parameters besides the mean. This lecture we'll test claims about the population proportion, but only a very small part of this process will be at all new to you.

You remember the basics from when we covered estimating population proportions in Lecture #14. If a variable is binomial, we call *p* the proportion in the population for which the variable was in one of the two categories. In a sample from that population, the sample proportion is called $\hat{p}$ and equals $\frac{x}{n}$, where *x* is the number of times the variable was in the category of interest and *n*, of course, is the sample size.

So the direction lines

STEPS IN TESTING CLAIMS

A)State the hypotheses and identify the claim.
B) State the values of *x* and *n*. Perform the test using the sample data or statistics, and make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value, which should be rounded to the nearest thousandth.
C)Decide whether or not to reject the null hypothesis.
D)Summarize the results.

are perfectly applicable to testing claims about the proportion, with the addition of stating the values of *x* and *n*. The only differences are that the parameter is *p* instead of μ, that the value of the parameter mentioned in the claim is p_0 instead of μ_0, and that the statistic is $\hat{p}$ instead of $\bar{x}$. Oh, and one more: we will **never** use the *t*-distribution in testing claims about the proportion, only *z*.

Here goes.

Testing the First Claim

Here's the claim: At most one-third of people stopped by the CHP for suspected DUI test above the legal limit. To test this claim, a random sample of 200 people stopped for suspected DUI is selected, and of these it turned out that 36% tested above the legal limit. Let's choose the 5% significance level.

A) Remembering to use *p* for the parameter, and that 'at most' encompasses the 'is less than' and the 'equals' options, the hypotheses become

H_0: $p \le 1/3$ (Claim)

H_1: $p > 1/3$

This is a right-tailed test.

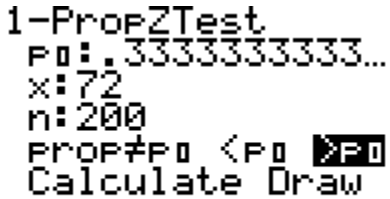
Let's skip the build-up and go directly to the Stat Test menu on the calculator. We'll be using 1-PropZTest.

Putting $1/3$ in for p_0 seems obvious, and when we enter that the calculator uses this repeating decimal in all its glory: .3333333333… But the next line asks for *x*, which is the number of people out of the 200 who tested above the legal limit, and that's information we weren't given directly. We'll have to use a little arithmetic, or a little algebra. With common sense, you'll realize that since 36% of the 200 people in the sample tested above the legal limit, to find out how many people that actually is you'll take 36% of 200 and come up with *x* = 72. Without common sense, you'll have to work harder. Take the equation $\hat{p} = \frac{x}{n}$ and substitute 0.36 for $\hat{p}$ and 200 for n:

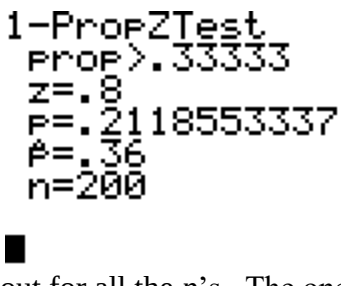
$$0.36 = \frac{x}{200}$$

Multiply both sides by 200 and again you'll get x = 72.

Either way, the 1-PropZTest screen looks like this:

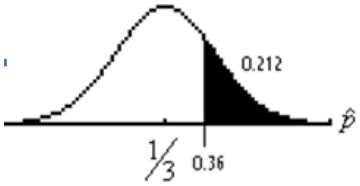


and, upon calculation, the screen becomes:



Watch out for all the *p*'s. The one with the hat is the sample proportion (which we didn't actually input – the calculator computed it from the *x* and the *n*), and the one without the hat is our *p*-value.

B) $x = 72; n = 200$



C) Here again is the way to decide whether or not to reject the null hypothesis:

If $p < \alpha$, reject H_0.

If $p > \alpha$, don't reject H_0.

In this case, 0.212 > 0.05, or $p > \alpha$, so **do not** reject H_0.

D) Here again are the rectangles for summarizing the results:

	Reject H_0	Don't reject H_0
The claim is H_0	There **is** sufficient evidence to **reject** the claim that…..	There **isn't** sufficient evidence to **reject** the claim that…..
The claim is H_1	There **is** sufficient evidence to **support** the claim that…..	There **isn't** sufficient evidence to **support** the claim that…..

Since we **didn't** reject H_0, and since the claim was H_0, we're in the upper-right rectangle: There **isn't** sufficient evidence to **reject** the claim that at most one-third of people stopped by the CHP for suspected DUI test above the legal limit.

Testing the Second Claim

Here's the second claim: The majority of Mendocino College students are female. Let $\alpha = 0.05$. Our sample is the Class Data Base, in which 67 of the 120 people in the sample are female.

What **is** a majority? It's more than half. If a group is split 50-50, there isn't a majority. Any time you see the phrase "A majority…" you don't have to read any further to translate it to a mathematical sentence. It's always $p > 0.5$. (Likewise, "A minority…" always translates as $p < 0.5$.) And of course it's always the alternative hypothesis. The null hypothesis has to include the 'is less than' and 'equals' options.

A) H_0: $p \leq 0.5$

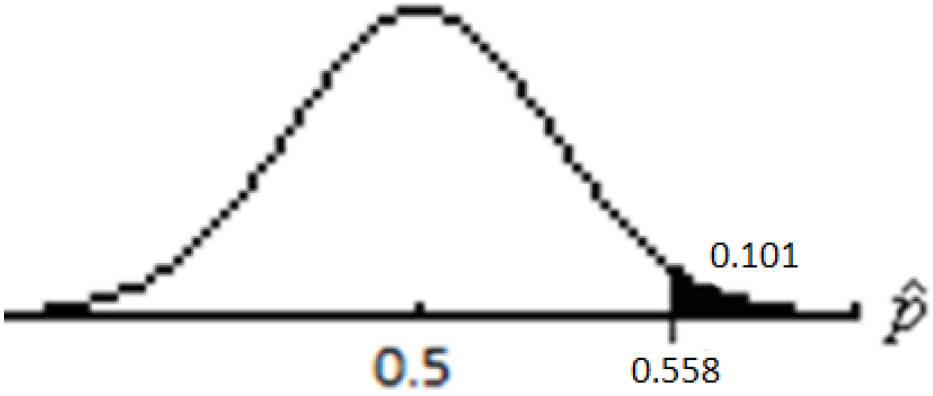
H_1: $p > 0.5$ (Claim)

B) $x = 67; n = 120$

```
1-PropZTest
 p0:.5
 x:67
 n:120
 prop≠p0 <p0 >p0
 Calculate Draw
```

```
1-PropZTest
 prop>.5
 z=1.278019301
 p=.100621376
 p̂=.5583333333
 n=120
```

Notice that the calculator found $\hat{p}$ for us. Here's the diagram:



C) 0.101>0.05, or $p > \alpha$, so **don't reject** H_0.

D) Since we **didn't** reject H_0, and since the claim was H_1, we're in the lower-right rectangle: There **isn't** sufficient evidence to **support** the claim that the majority of Mendocino College students are female.

Activity #17: Testing Claims about the Proportion

STEPS IN TESTING CLAIMS

A)State the hypotheses and identify the claim.
B)State the values of *x* and *n*. Perform the test using the sample data or statistics, and make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value, which should be rounded to the nearest thousandth.
C)Decide whether or not to reject the null hypothesis.
D)Summarize the results.

Use the 5% significance level in testing these claims.

1) It has been claimed that a minority of Americans are in favor of national health insurance. In a survey in which 400 people were interviewed, it was found that 47% were in favor of national health insurance.

2) A researcher studying demographics thinks that the majority of Mendocino College students are natives of Lake or Mendocino County. Test this claim.

3) Three-fourths of Mendocino College students own at least one pet.

Assignment #8

STEPS IN TESTING CLAIMS

A)State the hypotheses and identify the claim.
B)State the values of *x* and *n*. Perform the test using the sample data or statistics, and make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value, which should be rounded to the nearest thousandth.
C)Decide whether or not to reject the null hypothesis.
D)Summarize the results.

Test these claims.

1) It has been claimed that a majority of Americans have adequate health insurance. In a survey in which 200 people were interviewed, it was found that 57% have adequate health insurance. Test the claim at the 5% level of significance.

2) A journalist asserts that at most 40% of the population favors a national lottery. In a random sample of 400 adults, 184 expressed support for a national lottery. At the 1% significance level, what should be concluded about this claim?

In these problems, assume that the 120 people in the class survey constitute a random sample of Mendocino College students.

3) A researcher studying demographics thinks that a minority of Mendocino College students are **not** graduates of a Lake or Mendocino County high school. If 29 of the people surveyed are not graduates of a Lake or Mendocino County high school, test the claim at the 5% significance level.

4) Another researcher is studying consumer habits of community college students and claims that two-thirds of Mendocino College students own cars made in 2004 or later. Test the claim at the 10% significance level, given that 67 people surveyed had cars from 2004 or later.

Lecture #18: Testing Claims about the Standard Deviation

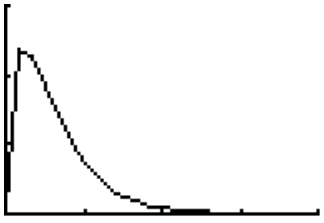
Now we are going to test claims about a third parameter, the population standard deviation. You might wonder why we'd want to do this. The population mean and the population proportion make sense – you'd want to be able to draw inferences about them. The standard deviation might seem irrelevant, but it isn't. There are situations in which it's important to be able to draw conclusions about this measure of variation.

Here's a situation in which it's important for the standard deviation to be small: You're manufacturing screws. The mean diameter of the screws is obviously important for the screws to fit, but if the diameters have a lot of variation many of the screws won't be usable. Variation needs to be kept very small. And here's a situation in which it's important for the standard deviation to be large: You're giving a standardized test. If all the scores are similar (with little variation), your test won't be useful for differentiating the test performances; everybody's score will be about the same.

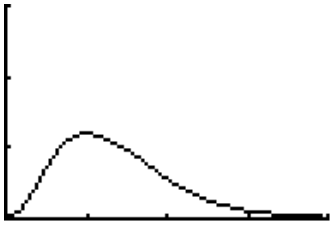
So we want to be able to test claims about σ, the population standard deviation, and to do so we have to know about the third of the four great statistical distributions, or probability density functions. The first two were the normal, in particular the *z*-distribution, and the *t*-distribution, with its degrees of freedom. Now we're going to use the **chi-square distribution**, or χ^2-distribution (pronounced with a hard 'k' – 'kigh' – if you say 'chee' or 'chai' you will sound uneducated, heaven forbid).

Whereas the *z*- and the *t*- distributions are symmetrical and theoretically extend forever in both directions, the χ^2-distribution has neither characteristic. It starts at 0, extends forever to the right, and is not symmetrical, though it becomes more so as the number of degrees of freedom increases. It has degrees of freedom, like the *t*-distribution, and you determine them the same way – by subtracting one from the sample size, *n*.

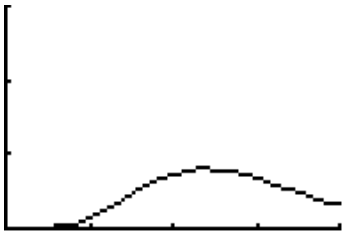
The reason that the χ^2-distribution starts at 0 is, as the name implies, because it involves adding things that are squared and thus cannot be negative. The more of these "things" that are added, the larger the number of degrees of freedom and the bigger the value of χ^2. Here's the χ^2 pdf for a χ^2 with three degrees of freedom ($n = 4$):



Its hump occurs over $\chi^2 = 1$. Note that we put a vertical axis at 0, because the pdf doesn't extend to the left of 0. Here's the χ^2 pdf for a χ^2 with seven degrees of freedom ($n = 8$):



Its hump occurs over $\chi^2 = 5$. Finally, here's the χ^2 pdf for a χ^2 with 14 degrees of freedom ($n = 15$):



Its hump occurs over $\chi^2 = 12$.

Here are all three χ^2's on one graph:

As you can see, the curves become lower and more symmetrical as the number of degrees of freedom increases, and the humps move to the right and are located over d.f. -2, or $n-3$.

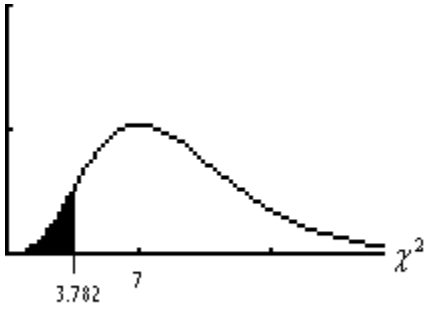
Testing claims about the population standard deviations involves finding areas under χ^2 pdf's. We can't use the calculator's test menu. (You might notice that there **is** a χ^2-Test listed on the menu, but it isn't meant for this purpose and can't be used for it.)

To determine what area to find under the pdf involves a process we haven't used yet in testing claims, though it's used in the old-fashioned, classical method of testing all

claims about parameters: finding the **test value**. The test value comes from a formula involving the parameter (σ_0 in this case) and the statistic (*s*, the standard deviation of our sample), and it is this test value which we place on the χ^2-distribution axis.

Before we get to that, though, let's practice finding areas under χ^2 pdfs.

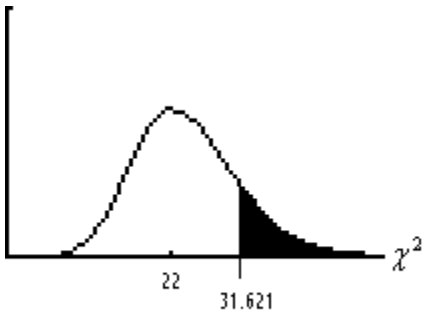
Say you want to find the area under a χ^2 pdf for a sample size of 10 to the left of 3.782. Here's the diagram:



I put 7 (d.f. = $n - 1 = 10 - 1 = 9$, and the hump is over d.f. $- 2 = 9 - 2 = 7$) under the hump so we could decide which side of the hump to put 3.782 on. Remember to label the axis χ^2. To find the shaded area, use χ^2 cdf, followed by the left-shaded number, the right-shaded number, and the number of degrees of freedom:

χ^2 cdf (0, 3.782, 9) ≈ 0.075

One more: What is the area under a χ^2 pdf for a sample size of 25 to the right of 31.621?



χ^2 cdf (31.621, 1000, 24) ≈ 0.137.

Notice how much more symmetrical and normal this curve looks than the one for d.f. = 9.

Now we're ready to test claims. The number that you'll be putting on the χ^2 axis, the test value, will be given by this formula:

TEST VALUE: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$.

Here are the directions for testing claims about σ:

A)State the hypotheses and identify the claim.
B)Find the test value, and determine the *p*-value. Make a diagram showing the distribution and the test value. Shade and label the *p*-value, which should be rounded to the nearest thousandth.
C)Decide whether or not to reject the null hypothesis.
D)Summarize the results.

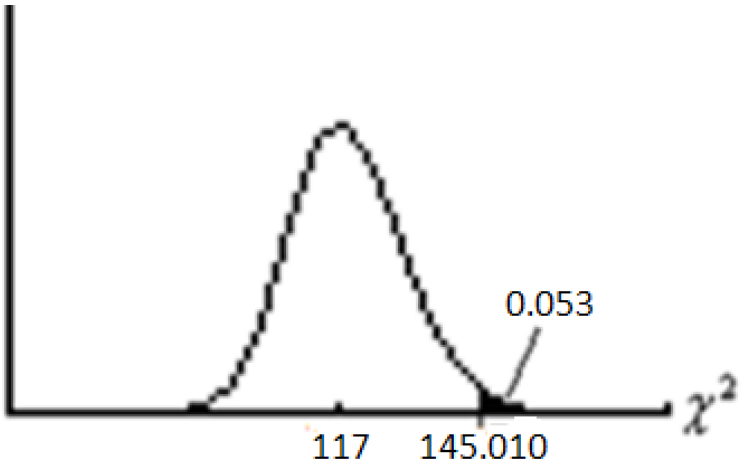
As you can see, the directions are identical to those for testing claims about μ and *p* except for the second step.

Testing the First Claim

The standard deviation of the shoe sizes of Mendocino College students exceeds 1.8 sizes. Use $\alpha = 0.10$.

A) $H_0: \sigma \leq 1.8$
$H_1: \sigma > 1.8$ (Claim)

B) So 1.8 is σ_0, and from the calculator we get 1.987 for *s*, so $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(120-1)\cdot 1.987^2}{1.8^2} \approx 145.010$. We're doing a right-tailed test, so the *p*-value is given by χ^2 cdf (145.010, 1000, 119)≈ 0.053.



C) 0.053< 0.10, or $p < \alpha$, so **reject** H_0.

D) Lower-left rectangle: There **is** sufficient evidence to **support** the claim that the standard deviation of the shoe sizes of Mendocino College students exceeds 1.8 sizes.

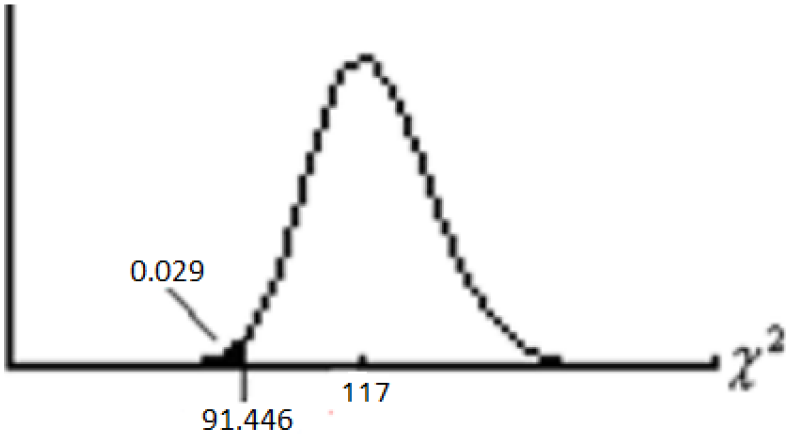
Testing the Second Claim

The variance of the ages of Mendocino College students is at least 110. Use the 10% significance level.

Notice that the parameter mentioned in this claim is the **variance**, σ^2, not the standard deviation, σ. This means that when you go to figure out the χ^2 test value, you'll use the 110 as is, without squaring it.

A) H_0: $\sigma^2 \geq 110$ (Claim)
H_1: $\sigma^2 < 110$

B) So 110 is σ_0^2, and the calculator gives 9.194 for *s*, so $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(120-1)\cdot 9.194^2}{110} \approx 91.446$. This is a left-tailed test, so the *p*-value is given by χ^2 cdf (0, 91.446, 119) ≈ 0.029.



C) 0.029 < 0.10, or $p < \alpha$, so **reject** H_0.

D) Upper-left rectangle: There **is** sufficient evidence to **reject** the claim that the variance of the ages of Mendocino College students is at least 110.

Testing the Third Claim

Remember the coffee machine from Lecture #11, the one that was way out of whack? Well, perhaps it's fixed now – we'll see. By 'fixed' we mean that the standard deviation of the number of fluid ounces it deposits in cups is at most 0.05 fluid ounces. If the standard deviation is less than or equal to 0.05 fluid ounces we say that the machine is working properly; otherwise we call the company that owns the machine to come and service it yet again.

So we take a random sample of eight such cups and measure their contents. These are the volumes in fluid ounces: 6.02 5.89 6.05 6.01 5.96 5.87 6.03 5.92

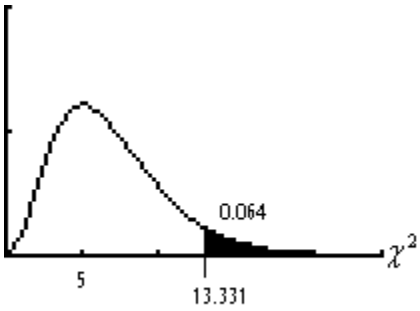
Let's claim that the machine is working properly, i.e. that the standard deviation of the volumes is at most 0.05 fluid ounces. Use $\alpha = 0.01$.

A) H_0: $\sigma \le 0.05$ fluid ounces (Claim)
H_1: $\sigma > 0.05$

We put the eight volumes in a list, and the calculator tells us that the sample standard deviation is 0.069.

B) So σ_0 is 0.05, and *s* is 0.069, so $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^{\,2}} = \dfrac{(8-1)0.069^2}{0.05^2} \approx 13.331.$

This is a right-tailed test, so the *p*-value is χ^2 cdf (13.331, 1000, 7) ≈ 0.064.



C) 0.064 > 0.01, or $p > \alpha$, so **do not** reject H_0.

D) Upper-right rectangle: There **isn't** sufficient evidence to **reject** the claim that the standard deviation of the volumes is at most 0.05 fluid ounces, or that the machine is working properly. No need to call for servicing.

Activity #18: Testing claims about the standard deviation

Perform hypothesis tests on these claims.

Include these steps in each test:

A)State the hypotheses and identify the claim.
B)Find the test value, and determine the *p*-value. Make a diagram showing the distribution and the test value. Shade and label the *p*-value, which should be rounded to the nearest thousandth.
C)Decide whether or not to reject the null hypothesis.
D)Summarize the results.

1) At the 5% significance level, test the claim that the standard deviation of the heights of Mendocino College students is at most 3.5 inches.

2) At the 1% level of significance, test the claim that the variance of the number of pets owned by students at Mendocino College is less than 6.

3) A machine is supposed to fill bags with exactly 16 ounces of pasta. The machine is said to be working properly if the standard deviation of the weights of the pasta in the bags is at most 0.1 ounce. Six randomly-selected bags are weighed, and the resulting weights in ounces are:

16.15	15.88	15.82
16.03	16.10	15.76

Test the claim that the machine is working properly at the 5% significance level.

TEST VALUE: $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^{\,2}}$

Lecture #19: Testing Claims about Two Populations

In practice, instead of making and testing a claim about one parameter, based on one sample, you might want to consider **two** samples, and test a claim about the difference in the value of the parameters of the populations from which the two samples are drawn. You might want to claim that Mendocino College men are taller than Mendocino College women, or that they are at least 4 inches taller on the average, instead of claiming that Mendocino College women average a certain height, or that Mendocino College students as a whole average more than a certain height. Or you might want to claim that Mendocino College women are more likely to be local high school graduates than Mendocino College men.

This lecture is about situations in which testing the claim involves using statistics from two samples to make inferences about the parameters of the populations the samples come from.

Claims about the Difference of Means

First, let me define the concept of **independent samples**. These are samples in which there's no way to pair the numbers in one sample with the ones in the other sample For instance, in the following claim, there are two groups, males and females, and they are **independent** of each other.

The other situation is **dependent samples**, in which the two samples have a natural pairing. For example, if you were measuring reaction time for people before they drink several beers and afterwards, you can directly compare the two numbers for each person. People naturally have different reactions time, and it makes more sense to look at the differences in individuals' reaction times before and after than to look at the average before and the average after. (We won't cover dependent samples in this course.)

We'll confine ourselves to claims that one mean is bigger than, smaller than, or different from another mean.

Here's a claim: The average Mendocino College male student wears shoes of a larger size than the average Mendocino College female student.

We need some notation. Let's call μ_M the population mean shoe size for the men, and μ_F the population mean shoe size for the women. Then our claim translates as $\mu_M > \mu_F$.

We'll alter the instructions about testing claims a little to eliminate the diagram – the situation has gotten somewhat more complicated.

STEPS IN TESTING CLAIMS

A) State the hypotheses and identify the claim.

B) Perform the test using the sample data or statistics, and state the *p*-value.
C) Decide whether or not to reject the null hypothesis.
D) Summarize the results.

This table shows the means, standard deviations, and sample sizes for shoe size for the males and the females in the Class Data Base:

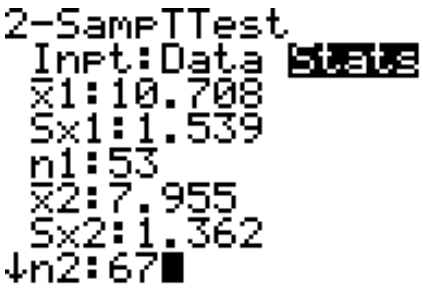
		Shoe Size
Males	**Mean**	10.708
	Standard Deviation	1.539
	Sample Size	53
Females	**Mean**	7.955
	Standard Deviation	1.362
	Sample Size	67

Step A should be familiar:

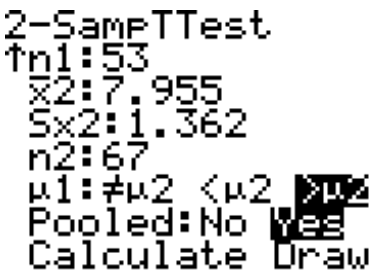
H_0: $\mu_M \leq \mu_F$

H_1: $\mu_M > \mu_F$ (Claim)

To perform Step B, we'll use a test called 2-SampTTest, calling the males Group 1 and the females Group 2:



There's more input than fits on one screen, so here's the rest. Ignore the overlap:



When we opt to calculate, the calculator gives us these results:

```
2-SampTTest
 μ1>μ2
 t=10.38057205
 p=1.250425E-18
 df=118
 x̄1=10.708
↓x̄2=7.955
```

There's more information below, but what we really want here is the *p*-value, which is very, very small – a decimal point followed by 17 zeros and a 1. This certainly qualifies as 0.000^{+} and is smaller than any imaginable α.

Step C: **Reject** H_0.

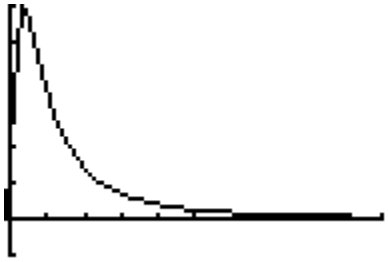
Step D: Lower-left rectangle: There **is** sufficient evidence to **support** the claim that the average Mendocino College male student wears shoes of a larger size than the average Mendocino College female student. It's pretty much a sure thing.

What does **pooled** mean, you might ask? Well, in the formula for computing the *p*-value for claims about the means of two populations, it makes a difference if you can say that the standard deviations of the two populations are the same or not. If you can say they're the same, you can **pool** them; if not, you can't.

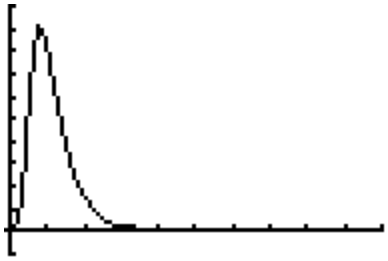
The *F*-Distribution

To pool or not to pool: How do you decide? To answer that question, we have to introduce the last of the four great distributions of statistics, joining the *z*, the *t*, and the χ^2. The *F*-distribution is the pdf that results when you take two χ^2-distributions and divide them. It's the ratio of two non-negative quantities and thus itself can never be negative. In that, its pdf looks rather like the χ^2 pdf, and it shares with the χ^2 also its asymmetry. **But** the *F* has **two** degrees of freedom – one for the numerator, and one for the denominator – because each part of the ratio, being a χ^2, has its own number of degrees of freedom.

Here's an *F* pdf with 5 numerator degrees of freedom and 3 denominator degrees of freedom:



Here's another, with 24 numerator degrees of freedom and 14 denominator degrees of freedom:



Unlike the χ^2, in which the hump keeps moving to the right as the number of degrees of freedom increases, the *F*'s hump gets closer and closer to an *F*-value of 1 (which expresses the situation in which the numerator and denominator are equal to each other).

If for any reason you wanted to test a claim about the standard deviations of two populations, you'd have to use the *F*-distribution and state the claim in terms of the **ratios** of the standard deviations instead of their differences. For instance, if you want to claim that the standard deviations are equal, you'd claim that $\frac{\sigma_1}{\sigma_2} = 1$.

We're going to use this test only to establish whether the standard deviations are equal or not so we know whether to pool them. Let's go back to our claim about shoe sizes and perform first the 2-SampFTest so we can decide about pooling:

```
2-SampFTest
 Inpt:Data Stats
 Sx1:1.539
 n1:53
 Sx2:1.362
 n2:67
 σ1:≠σ2 <σ2 >σ2
 Calculate Draw
```

Notice that we opted for $\sigma_1 \neq \sigma_2$, which is equivalent to the alternative hypothesis for $\frac{\sigma_1}{\sigma_2} = 1$, or $\sigma_1 = \sigma_2$. Here's the result of the test:

```
2-SampFTest
 σ1≠σ2
 F=1.276800442
 p=.3463859348
 Sx1=1.539
 Sx2=1.362
↓n1=53
```

The *p*-value of 0.346 means we should definitely pool, because we can't say that the population standard deviations aren't the same.

(If we had chosen not to pool the standard deviations, the *p*-value would have turned out to be 2×10^{-17}, hardly different from what we got when we pooled and not worth making a fuss about.)

Second Claim

So let's cut to the chase and use the *t*-distribution with pooling in testing a new claim: Mendocino College men are younger than Mendocino College women on average. Let $\alpha = 0.10$. As you can see from the table, the mean age of the men in the Class Data Base is smaller than the mean age of the women, but is this enough to infer that the difference would hold up for the population?

		Age
Males	**Mean**	21.623
	Standard Deviation	4.634
	Sample Size	42
Females	**Mean**	26.060
	Standard Deviation	11.254
	Sample Size	60

In Step A, I'm still calling the men Group 1 and the women Group 2 for the purposes of the calculator:

H_0: $\mu_M \geq \mu_F$

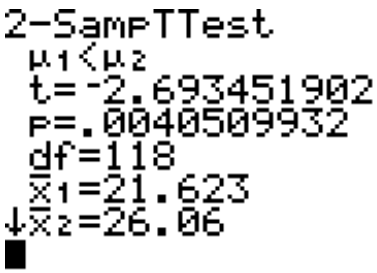
H_1: $\mu_M < \mu_F$ (Claim)

Here are the screens for Step B:

```
2-SampTTest
 Inpt:Data Stats
 x̄1:21.623
 Sx1:4.634
 n1:53
 x̄2:26.06
 Sx2:11.254
↓n2:67
```

```
2-SampTTest
↑n1:53
 x̄2:26.06
 Sx2:11.254
 n2:67
 μ1:≠μ2 <μ2 >μ2
 Pooled:No Yes
 Calculate Draw
```

with the result



The *p*-value is 0.004, which is less than α, so we **reject** the null hypothesis (Step 3), and we conclude that there **is** sufficient evidence to **support** the claim that Mendocino College men are younger than Mendocino College women on average.

Claims about the Difference of Proportions

What if you want to make a claim about the relative size of two population proportions, *p*, based on the sample proportions, $\hat{p}$, of samples from the populations? Here the math is much simpler, involving only the *z*-distribution.

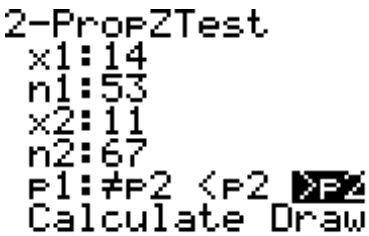
Here's a claim: Mendocino College men are more likely to pick blue as their favorite color than Mendocino College women. (The protocol will be that any color name containing the word "blue" will be considered as blue.) We'll let p_M represent the proportion of Mendocino College men who pick blue as their favorite color and p_F represent the proportion of Mendocino college women who pick blue as their favorite color. Let's use the 10% significance level. Here's Step A:

H_0: $p_M \le p_F$
H_1: $p_M > p_F$ (Claim)

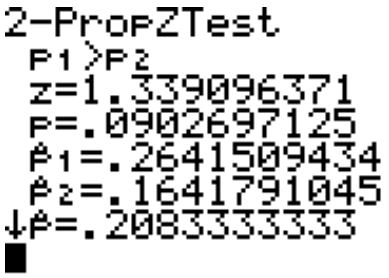
Note that we're using the parameter *p* instead of the parameter μ, and that when we refer to the **likelihood** of choosing blue, we're really talking about the **proportion** of the group that chooses blue.

Consulting the Class Data Base, we find that 14 men chose blue and 11 women chose blue.

We use 2-PropZTest, which is very simple:



Upon calculation, we get



It's easy to see that $\hat{p}_1$ is the sample proportion for the men: $\frac{14}{53} \approx 0.264$. And of course $\hat{p}_2$ is the sample proportion for the women: $\frac{11}{67} \approx 0.164$. And also the *p*-value is 0.090. But what is $\hat{p}$? It's what we call the **pooled *p*** (delightful name). It's the sample proportion if we lump the men and the women together and figure out the sample proportion for the whole group. The formula is $\frac{x_M + x_F}{n_M + n_F} = \frac{14+11}{53+67} = \frac{25}{120} \approx 0.208$. The pooled *p* is used in the formula for computing the *p*-value. It seems like we are simply drowning in *p*.

At any rate, the answer to Step B is that the *p*-value is 0.090.

Step C: Since 0.090 < 0.10 (which was α), **reject** H_0 .

Step D: There **is** sufficient evidence to **support** the claim that Mendocino College men are more likely to pick blue as their favorite color than Mendocino College women.

Here's a claim: Mendocino College men are more likely not to have pets than Mendocino College women. Let α be 0.10.

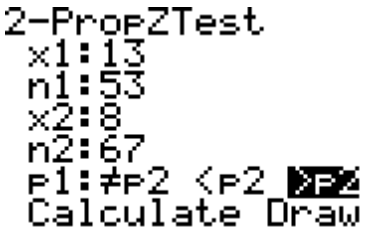
We'll let p_M stand for the population proportion of Mendocino College male students who have no pets and p_F stand for the population proportion of Mendocino College female students who have no pets.

Step A:

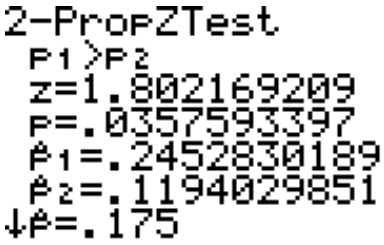
H_0: $p_M \leq p_F$
H_1: $p_M > p_F$ (Claim)

This is identical to Step A in the last claim about choosing blue. Just the interpretation is different.

There are 13 men in the Class Data Base whose number of pets is 0, and there are 8 women. 2-PropZTest looks like this:



with the result



(You might notice that here, as in the last test, $\hat{p}$, the pooled *p*, lies between $\hat{p}_M$ ($\hat{p}_1$) and $\hat{p}_F$ ($\hat{p}_2$), because it is the result of their merging, and also that it is closer to $\hat{p}_F$ because there are more women than men in the Class Data Base.) The *p*-value is 0.036.

Step C: Since 0.036 < 0.10, **reject** the null hypothesis.

Step D: There **is** sufficient evidence to **support** the claim that Mendocino College men are more likely to own pets than Mendocino College women.

Exam #3 – Confidence Intervals and Testing Claims – Sample

For Problems #1 and 2: A survey of 1200 California residents yielded 409 people who were born in other countries.

1) Find the point estimate of the proportion of all California residents who were born in other countries, to the nearest tenth of a percent.

2) Find the 95% confidence interval for this proportion to the nearest tenth of a percent.

For Problems 3 and 4: A random sample of 60 mature redwood trees yields a sample mean diameter of 135.6 cm and a sample standard deviation of 22.7 cm.

3) Find the 90% confidence interval for the population mean diameter, to the nearest tenth.

4) At the 90% confidence level, find the margin of error (also called the maximum error) of the estimate of this mean to the nearest tenth.

5) Suppose you wanted to estimate the proportion of Californians born in other countries within 3%, at a 99% level of confidence, and that you have no prior knowledge of the sample proportion. What is the minimum sample size you would have to use?

6) Using the standard deviation from Problems 3 and 4, find the minimum sample size you would have to use to estimate the average diameter of mature redwood trees within 5 cm at the 95% confidence level.

7) To the nearest hundredth, what is the point estimate for the population mean of heights of female college students, if a random sample of seven female college students is selected, and their heights are 162.5 cm, 158.3 cm, 182.1 cm, 173.2 cm, 160.9 cm, 154.5 cm, and 151.3 cm?

8) You're testing the claim that the average guinea pig weighs 1.21 kilograms. A sample of 23 guinea pigs has an average weight of 1.04 kg, with a standard deviation of 0.51 kg. The test has a *p*-value of 0.124. Make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value.

9) You're testing the claim that at least 54% of Californians are bilingual. In a sample of 300 randomly selected Californians, 147 were bilingual. The test has a *p*-value of 0.041. Make a diagram showing the distribution, the parameter and the statistic. Shade and label the *p*-value.

10) What would the null hypothesis be when testing the claim that the average miles per gallon of hybrid cars is under 50 miles per gallon?

11) What would the null hypothesis be when testing the claim that a majority of people watch the TV news?

12) What would the alternative hypothesis be when testing the claim that the standard deviation of scores on a standardized test is 25 points?

13) To the nearest thousandth, what is the *p*-value for a test of the claim that the average microwave costs at most $125, if a random sample of 40 microwaves yielded a mean cost of $137, with a standard deviation of $26?

14) To the nearest thousandth, what is the *p*-value of a test of the claim that the average child says his or her first word before the age of 15 months, if a random sample of 35 children shows that the average age when they said their first word was 13.5 months, and the population standard deviation is 3.7 months?

15) To the nearest thousandth, what is the *p*-value for a test of the claim that at least three-fifths of community college students receive some sort of financial aid, if a randomly selected sample of 1400 college students yielded 822 who receive financial aid?

16) To the nearest thousandth, what is the test value for a test of the claim that the population standard deviation of lengths of forearms is less than 5.2 cm, if a random sample of the lengths of 43 forearms had a standard deviation of 4.7 cm?

17) In testing the claim $\sigma > 10.4$, a sample of 12 produces a test value $\chi^2 = 17.234$. What is the *p*-value of the test, to the nearest thousandth?

18) If the claim "A majority of American children own at least one videogame system" is tested at the 1% significance level, and the test turns out to have a *p*-value of 0.027, what is the final conclusion? (In other words, summarize the results.)

19) If the claim "The average cat weighs less than 3.2 kg" is tested at the 5% significance level, and the *p*-value of the test turns out to be 0.042, what is the final conclusion? (In other words, summarize the results.)

20) If the claim "The standard deviation of the weights of cats is at most 1.5 kg" is tested at the 5% significance level, and the *p*-value of the test turns out to be 0.067, what is the final conclusion? (In other words, summarize the results.

Exam #3 – Confidence Intervals and Testing Claims – Sample

Answers

1) 34.1%

2) $31.4\% < p < 36.8\%$

3) $130.7\,cm < \mu < 140.5\,cm$ (t-interval)

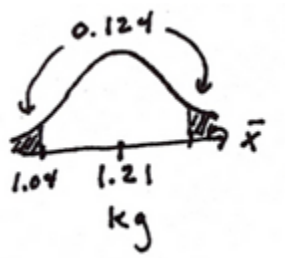
4) 4.8 cm

5) 1844

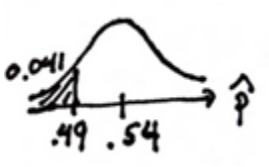
6) 80

7) 163.26 cm

8)



9)



10) $\mu \geq 50\,mpg$

11) $p \leq 0.5$

12) $\sigma \neq 25\,pts$

13) 0.003

14) 0.011

15) 0.163

16) 34.311

17) 0.101

18) There **is not** sufficient evidence to **support** the claim that a majority of American children own at least one videogame system.

19) There **is** sufficient evidence to **support** the claim that the average cat weighs less than 3.2 kg.

20) There **is not** sufficient evidence to **reject** the claim that the standard deviation of the weights of cats is at most 1.5 kg.

Lecture #20: Correlation and Regression, Part 1

This lecture and next, we'll be looking at scatter plots and inferences you can make about them. There are two topics here: **correlation**, which deals with describing the relationship between two variables and deciding about its strength, and **regression**, which involves using values of one variable to make **predictions** about the values of the other variable.

Today I'm going to talk about these concepts in general and for a specific example. Then you'll generate some data which we will use during the next lecture so you can perform correlation and regression analysis on your own.

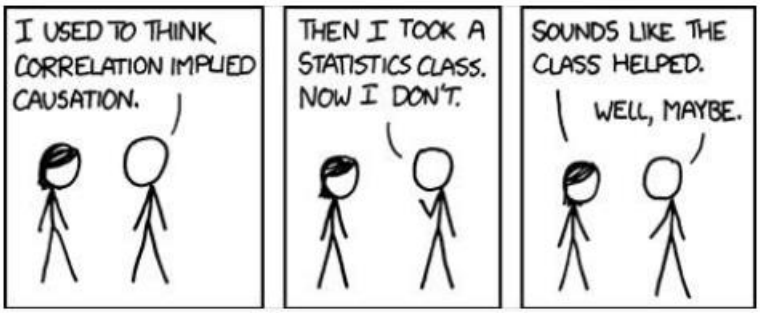
Let's say that you have ten people each taking two different medical tests, and you have the scores each person got on each of the tests. Here they are:

Person #	Test #1	Test #2
1	45.2	44.4
2	48.1	52.8
3	43.8	40.5
4	52.1	58.9
5	53.7	57.7
6	42.6	45.4
7	44.2	47.2
8	47.0	52.0
9	54.4	61.6
10	50.2	53.0

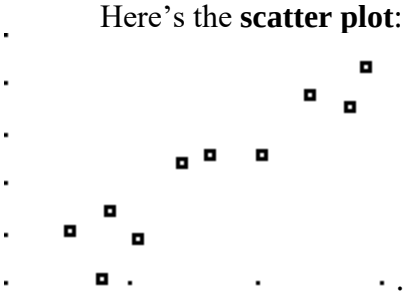
You're interested in how the scores for an individual relate to each other, how consistent or inconsistent they are, perhaps with a mind to eliminating one of the tests as redundant, or perhaps to use people's scores on one test to predict their scores on the other (maybe Test #1 is much cheaper to administer, but what you really want to know is how people do on Test #2 – can you get away with giving just Test #1?).

We make a scatter plot using the data, putting the score on Test #1 on the *x*-axis and the score on Test #2 on the *y*-axis and putting a dot (or some kind of mark) for each person. As before, the variable represented on the *x*-axis is called the **independent variable** but for the purposes of regression is also referred to as the **predictor variable**. The variable on the *y*-axis is called the **dependent variable** or the **response variable**.

It's important to remember that designating one variable to be the predictor and the other to be the response in no way implies that the value of the predictor variable **causes** the response variable to take on whatever value it has. That's pretty obvious in the medical test example, where the results of both tests are the consequence of some medical condition, and neither causes the other. The predictor variable **might** have a causal effect on the response variable, but you can't prove it by demonstrating correlation.



Here's the **scatter plot**:

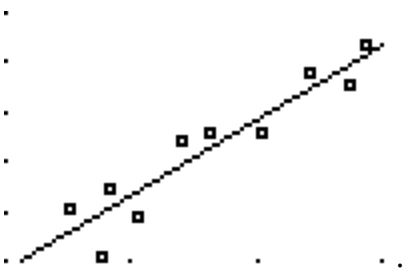


For instance, Person #1's marker and numbers are

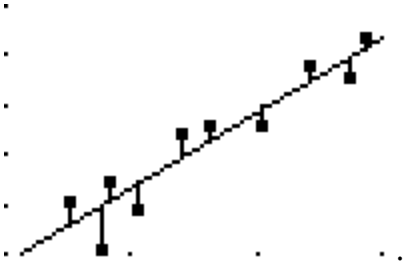
So what do we make of this picture? It's obvious that although the points are not on a single straight line, they are pretty close to one. They tend to rise from left to right. The higher the Test #1 score, in general the higher the Test #2 score. We call this a **positive relationship**. (If the points tended to fall going from left to right, meaning that higher values of one variable tend to go with lower values of the other, we'd have a **negative relationship.**) If you were to draw the straight line that comes as close as possible to the points, it would also go up from left to right, i.e. have a positive slope. The line we want is called the **least-squares best-fit regression line**, among other things, and its name reveals what it is that makes it the best fit.

First let's see its equation: $y = 1.508x - 21.216$. The slope, which we label *a*, is 1.508, and the *y*-intercept, referred to as *b*, is -21.216. So the line is of the form $y = ax + b$. You can see that *a* and *b* are statistics, because they describe a sample, in this case the ten pairs of numbers that generated the equation of the regression line. The actual computation of the slope and *y*-intercept of this line is very complicated, and we let

the calculator produce it for us, but you can understand the condition that it fulfills. Look at how the line fits in with the scatter diagram:



Some of the points are above the line, and some are below. In this next picture, I've had the calculator draw in the vertical lines between the points and the regression line:



The directed length of the vertical line connecting the point and the regression line is called the **residual** of the point. It's negative if the point is below the line, and it's positive if the point is above the line. If you find all the residuals for a scatter diagram and its regression line, square them, and add up the squares, then the least-squares regression line is exactly that: the line with the smallest such sum. Any other line, if you measured the vertical distances from the points to the line, squared them, and summed the squares, you would get a larger total than you do for this line.

This may not seem like such a big deal, but it is the accepted criterion for deciding which line is best.

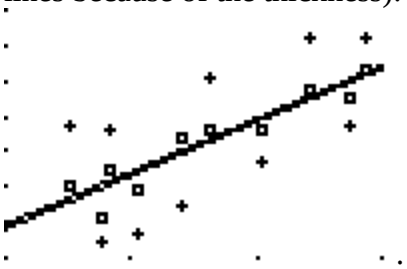
The line's equation is used for predicting how people will score on Test #2 given their score on Test #1. Going back to the person who got 45.2 on Test #1 and 44.4 on Test #2, we can see that this person got a lower than expected score on Test #2 because his or her point is below the line, thus having a negative residual. We would have predicted Person #1's score on Test #2 to be $y = 1.508(45.2) - 21.216 \approx 46.9$, not 44.4. (This person's residual would be $44.4 - 46.9 = -2.5$.)

The question remains whether we can use the Test #1 result as an accurate enough predictor of the Test #2 result that we can dispense altogether with performing Test #2. Of course, that depends partially on how closely we have to measure the result of Test #2 in order for it to be useful. That in turn would depend on the nature of Test #2 and what it is we're trying to measure. But another part of the question involves looking at how closely the line fits the points, because it would be possible to have a different set of ten

points with an almost identical regression line, but this set would contain points much further away from the line. Here's an example of such a set:

Person #	Test #1	Test #2
1	45.2	38.2
2	48.1	60.2
3	43.8	37.3
4	52.1	65.8
5	53.7	53.6
6	42.6	53.6
7	44.2	53.1
8	47.0	42.4
9	54.4	66.1
10	50.2	48.5

And here is that set graphed along with the first, with the points for the new set marked by plus signs. Also, both its own regression line and the regression line from the previous set are shown (they are almost indistinguishable, but you can tell there are two lines because of the thickness):



Here the points are still pretty close to the line but definitely not as close as before, yet we would make essentially the same predictions about the results of Test #2 based on the results of Test #1. And our predictions would be off by more.

We need a way of measuring how closely the line fits the points, and for this we use the **sample correlation coefficient**, symbolized by *r*. Like *a* and *b*, it's a statistic describing our sample. Its very complicated formula produces a number between -1 and $+1$. In essence, $r = -1$ if the points all lie on a straight line with a negative slope, and $r = +1$ if all the points lie on a straight line with a positive slope. And the closer the absolute value of *r*, $|r|$, is to 1 (in other words, the closer *r* is to either -1 or $+1$) the more closely the points hug the line, and the better the line is for purposes of prediction. We use LinReg to get both the equation and the correlation coefficient.

For our original table of results, $r = 0.938$, and for the altered set pictured above, $r = 0.615$. What does this tell us? The line is a better fit for the first set than for the second, because *r* is closer to 1 for the first set than for the second.

If *a* (the slope of the line) is positive, so is *r*, and if *a* is negative, so is *r*.

Now we're going to go back to our claims-testing mode.

Let's say that we want to be able to use the regression equation we got for our first set of test result to make predictions in general about someone's result on Test #2 on the basis of the person's result on Test #1, and we want to be fairly sure that we are justified in doing this.

If we had the entire population of pairs of results on Test #1 and Test #2, we could make a scatter diagram for the population and calculate the equation of the least-squares regression line and the correlation coefficient. The correlation coefficient would then be a parameter called the **population correlation coefficient**, labeled ρ, which is the Greek 'r' and is spelled 'rho' and pronounced 'row'.

What we do is to test the claim that the populations represented by scores on Test #1 and Test #2 are **positively correlated**, in other words that ρ is greater than 0. We could do this by performing what the calculator calls the LinRegTTest, but we'll take a different approach.

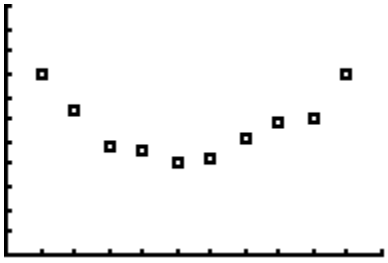
This approach involves what's known as a **critical value**. You find a number that, if your statistic is bigger than it, you can reject the null hypothesis. We could have used this technique in testing claims about other kinds of parameters, but we didn't because we used the *p*-value method.

A table of critical values for *r* reveals that for eight degrees of freedom (you have to deduct one for each variable from the sample size $(10-1-1=8)$ the correlation coefficient would have to be at least 0.621 for us to conclude at the 5% significance level that there **is** a positive relationship between the two variables. So our first set $(r=0.938)$ gives us the assurance we need that a positive relationships exists, and that we can use the regression equation for purposes of predicting a person's score on Test #2 from that person's score on Test #1.

Thus, if a person had a score of 51.3 on Test #1, we would be justified in predicting that the person would have a score of $y=1.508(51.3)-21.216\approx 56.1$ on Test #2. (Actually, 56.1, the value of *y* which the equation would yield, would be only a **point estimate** of the predicted value for a score of 51.3 on Test #1. In practice we would generate a **confidence interval** for the prediction. But we aren't going to cover how to do that in this course and will stick to a single predicted value.)

Using the second set, $(r=0.615)$, we see that *r* is slightly lower than the critical value, so if asked to predict someone's score on Test #2 from his or her Test #1 score, we would act cautiously and just say that our best guess would be the **average** of the Test #2 scores and leave it at that. That would be the mean, which is 51.9.

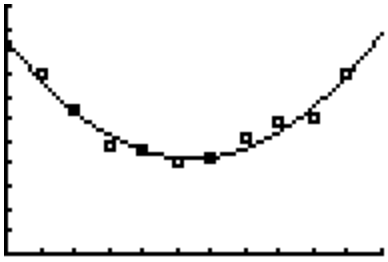
Everything we've done so far is called **linear regression**, because we've used a straight line. Sometimes a pattern of points is clearly not in the shape of a straight line, but there's a definite shape anyway. Look at this set of points:



The *x*-*y* pairs that formed this scatter diagram are

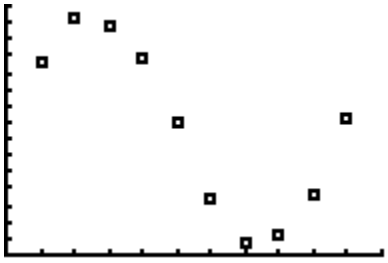
X	Y
1	8.0
2	6.4
3	4.8
4	4.6
5	4.0
6	4.3
7	5.2
8	5.8
9	6.1
10	7.9

No straight line could do it justice. The low and the high values of the *x*-variable appear to yield higher *y*-values than the middle *x*'s. How about fitting a parabola to the data? We call this **quadratic regression** for obvious reasons, and when we do it and graph the resulting parabola along with the points we get



Quite a good fit! This kind of relation could occur (though with different numbers than the ones I used to create the effect) for something like auto accident rates for drivers according to their ages, with the young and the old being involved in a greater proportion of accidents than the middle-aged.

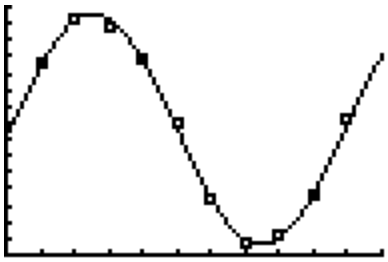
Or how about this one:



I got this using the following table:

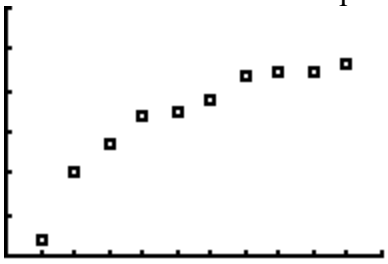
X	Y
1	57.6
2	71.1
3	68.8
4	58.7
5	40.2
6	16.4
7	3.7
8	5.5
9	17.6
10	41.5

The points lie near neither a straight line nor a parabola, but they have a suspiciously sinusoidal (the shape of a sine curve) pattern to them. How about **sinusoidal regression**? The scatter diagram and the sinusoidal regression line look like this:



What could possibly create such a pattern? Anything that depends on the seasons, like sales volume of ski equipment or camping gear.

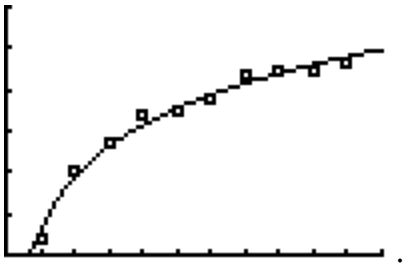
Here's another shape that a scatter diagram may take on:



It comes from this table:

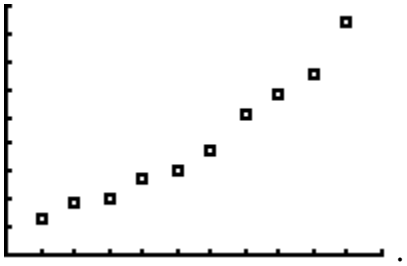
X	Y
1	3.8
2	19.9
3	27.1
4	34.1
5	35.2
6	37.8
7	43.7
8	44.1
9	44.3
10	46.2

Whereas a straight line isn't a terrible fit, there's a better one in a case like this where the growth rate is big at first but then slows down: **logarithmic regression**, in which the *y* is related to the natural log of the *x*, not to the *x* itself. Here's what it looks like



A relation like this might arise when looking at levels of sales for different amounts of money spent on advertising – additional advertising dollars have diminishing returns as far as promoting sales is concerned.

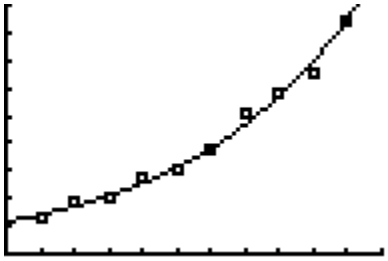
Or you could have the sort of relation in which growth accelerates rather than slows:



The table that produced this scatter diagram is

X	Y
1	13.2
2	18.2
3	20.3
4	27.4
5	31.0
6	37.2
7	50.6
8	58.2
9	65.7
10	84.6

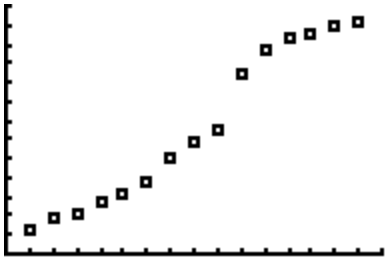
This is best served by the **exponential regression** model, in which *y* is related to *e* to some power containing *x* rather than to *x* itself. Here's the regression curve with its scatter diagram:



An illustration would be unrestrained growth in the size of a vegetable crop or in the value of a rare collectible.

But is growth ever really unrestrained, or unrestrained for very long? Sooner or later the crop has used up its space, its water, its nutrients, or whatever, and there's a limit even to the price that people will pay for a Ty Cobb baseball card. A more realistic growth model, and one that might be lurking just to the right of the scatter plot above, is one in which the growth rate levels off and finally approaches zero – the size of the crop or the price of the collectible reaches a peak.

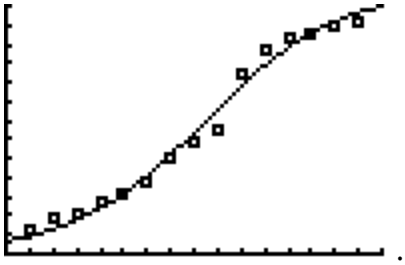
Look at what happens if we extend the table above to reflect this idea:



This results from adding the following rows to the table that made the scatter diagram:

11	107.1
12	113.7
13	116.3
14	119.9
15	120.7

This "growth with a cap" analysis is called **logistic regression**. Here it is with its scatter diagram:



The abrupt change in pattern from the exponential to the logistic growth example serves as a warning that applies to all use of regression analysis to make predictions. If you pick a value between the smallest x and the largest x that produced the regression equation in the first place and use it to predict y, that's called **interpolation**, and it's a legitimate thing to do, assuming the p-value as discussed above is small enough to justify claiming there's a real relationship between the variables. But if you use the equation to make a prediction for an x that is outside the range of the original x's (a process called **extrapolation**), you are on very shaky ground indeed and should watch your step!

Lecture #21: Correlation and Regression, Part 2

To give you some practice with the concepts of correlation and regression as developed in the last lecture, I've selected ten students and placed their heights, forearm lengths, and head circumferences in a table which you can copy into the lists on your calculator, putting Height in List 1, Forearm in List 2, and Head in List 3.

Person #	Height	Forearm	Head
1	174	43	56
2	166	44	57
3	150	39	56
4	176	48	60
5	160	42	56
6	167	45	56
7	181	47	56
8	170	42	57
9	172	45	55
10	155	40	52

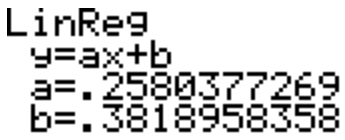
A good way to check that you've put in the correct numbers is to find the mean of each set and make sure it matches these means: Height: 167.1 cm, Forearm: 43.5 cm, and Head: 56.1 cm.

Finding the Regression Equation and the Sample Correlation Coefficient

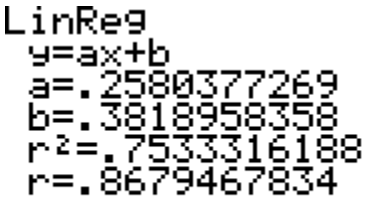
First, using height as the predictor variable and forearm length as the response variable, let's find the equation of the regression line and the value of the sample correlation coefficient. We'll round all numbers to the nearest thousandth.

Note that Height is in List 1 and Forearm in List 2.

Putting in LinReg (ax+b) L_1, L_2 – naming the predictor variable list first and the response variable list second – might yield the following screen:

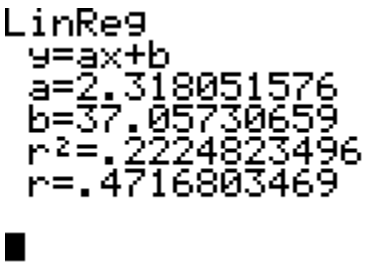


I say "might" because if this is what you got, though it gives the values of *a* and *b* in the regression equation, it doesn't tell you the sample correlation coefficient, and you'll have to change the settings on the your calculator. Push 2^{nd} 0 (CATALOG), cursor down to DiagnosticOn, and then press ENTER twice. That should do it, and now the same command – LinReg (ax+b) L_1, L_2 – should yield



So the regression equation, with *a* and *b* rounded to the nearest thousandth, becomes $y = 0.258x + 0.382$, and the sample correlation coefficient, *r*, is 0.868.

Let's try another one, this time using head size (List 3) as the predictor variable and height as the response variable. Using LinReg (ax+b) L_3, L_1, we get



So the regression equation is $y = 2.318x + 37.057$, and the sample correlation coefficient is $r = 0.472$ to the nearest thousandth.

Prediction

As I explained in the last lecture, the question of whether we're justified in using the regression equation to predict a value of the response variable for a certain value of the predictor variable can be settled by looking at the critical value for *r* at a given significance level and for a given number of degrees of freedom. The critical value for the 5% significance level with 8 degrees of freedom was 0.621.

So for the purpose of predicting forearm length from height, where *r* was 0.868, we **can** use the equation, and when we do, we give the *y* another name – $\hat{y}$, pronounced y-hat of course, and meaning 'the predicted value of *y*'. Let's use the equation to predict the forearm length of a person of height 179 cm:

$$\hat{y} = 0.258x + 0.382 = 0.258(179) + 0.382 = 46.564$$

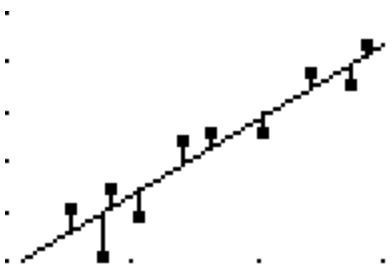
The forearm length we would predict for a person whose height is 179 cm is 46.6 cm, to the nearest tenth of a centimeter.

What if we were asked to predict the height of a person whose head size is 53 cm? Since $r = 0.472$ in this case, less than the critical value, we'll have to resort to predicting that the person, despite his or her head size, would be about average in height, and the

mean height of the sample, which we'll now call $\bar{y}$ because height is the *y*-variable, is 167.1 cm. So that's our prediction.

Residuals

Once you're established that two variables have a significant correlation, you can go on to see how accurate your prediction of the value of the response variable is in individual cases. Last lecture we defined the **residual** as the difference between the **actual** value of the response variable and the **predicted** value based on the regression line. Then I illustrated residuals graphically as the directed lengths of vertical lines on the graph, like this:



Now we'll use a formula for the residual for the nth individual. It's $y_n - \hat{y}_n$, where y_n is the **actual** value of the response variable for the nth individual, and $\hat{y}_n$ is the **predicted** value of the response variable for the nth individual.

In the regression analysis in which height is the predictor variable and forearm length is the response variable, what is the residual for Person #2 to the nearest tenth? We can ask this question, because we showed that height and forearm length are significantly correlated.

The *x*-value (height) of Person #2 is $x_2 = 166$ cm. The y-value (forearm length) is $y_2 = 44$ cm. Using the regression equation $\hat{y} = 0.258x + 0.382$, we get $\hat{y}_2 = 0.258(166) + 0.382 = 43.21$. The residual for Person #2 is thus $y_2 - \hat{y}_2 = 44 - 43.21 = 0.79$, or 0.8 cm to the nearest tenth. This means that Person #2's forearms are a little under a centimeter longer than we'd expect given his or her height.

What's the residual for Person #8? This person's height is $x_8 = 170$ cm, with forearm length $y_8 = 42$ cm. We get the predicted value $\hat{y}_8 = 0.258(170) + 0.382 \approx 44.24$. So $y_8 - \hat{y}_8 = 42 - 44.24 = -2.24$, and Person #8's residual to the nearest tenth is -2.2 cm. Person #8's forearms are a little more than 2 centimeters shorter than we'd expect given his or her height.

Multiple Regression

Finding and stating the linear regression equation and the sample correlation coefficient, making predictions using the equation (or not, if a significant correlation cannot be claimed to hold), and finding residuals – these are the tasks we've covered. Linear regression is just one way of attempting to make sense of the relationship between two variables. We saw some other patterns at the end of the last lecture.

But in all these cases, linear and otherwise, we're trying to express the value of the response variable in terms of one other variable, the predictor variable. We call comparing two variables in this way a **simple relationship**, and the prediction equation is an example of **simple regression**. It's 'simple' because there's only one predictor variable.

Sometimes, though, it's a lot better to use more than one predictor variable. Say you're trying to predict shoe size from height (as you'll be doing in Assignment #9). Height turns out to be a pretty good predictor, but, as you know, some people have large feet for their height and some have small feet. Perhaps the tendency to have large or small feet runs in families. Wouldn't it be helpful to know the shoe sizes a person's mother and father wear? You could label the person's height x_1, the mother's shoe size x_2, and the father's shoe size x_3, and then you could (or at least a computer could) construct a prediction equation which would look like this: $\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$. The relation between your height and your parents' shoe sizes on the one hand, and your own shoe size on the other is an example of a **multiple relationship**, and the prediction equation involves **multiple regression**. The job of the researcher is to find the best predictor variables to use, ones which do not overlap and depend upon each other, and which produce the closest fit to the data.

Activity #21: Correlation and regression

Person #	Height	Forearm	Head
1	174	43	56
2	166	44	57
3	150	39	56
4	176	48	60
5	160	42	56
6	167	45	56
7	181	47	56
8	170	42	57
9	172	45	55
10	155	40	52

1) Using forearm length as the independent variable and height as the dependent variable, state the equation of the regression line and the value of the sample correlation coefficient, to the nearest thousandth.

2) Using height as the predictor variable and head size as the response variable, state the equation of the regression line and the value of the sample correlation coefficient, to the nearest thousandth.

In #s 3 and 4, if r > 0.600, use the regression equation to make the prediction. If r < 0.600, predict the average for the response variable.

3) Predict to the nearest tenth the height of a person whose forearm length is 42 cm.

4) Predict to the nearest tenth the head size of a person whose height is 172 cm.

5) In the regression analysis using forearm length as the independent variable and height as the dependent variable, find the residual to the nearest tenth for Person # 4.

6) In the regression analysis using forearm length as the independent variable and height as the dependent variable, find the residual to the nearest tenth for Person #

Assignment # 9

In Problems 1-4, state the equation of the regression line and the value of the sample correlation coefficient, to the nearest thousandth.

1) Using age as the predictor variable and number of pets as the response variable.

2) Using height as the independent variable and shoe size as the dependent variable.

3) Using shoe size as the predictor variable and age as the response variable.

4) Using shoe size as the independent variable and height as the dependent variable.

Predict the following, rounding to the nearest tenth, using 0.600 as the cut-off for r:

5) The height of a person who wears a size 10 shoe.

6) The number of pets owned by a person who is 40 years old.

7) The age of a person with size 11 shoes.

8) The shoe size of a person who is 62 inches tall.

In the regression analysis involving height as the independent variable and shoe size as the dependent variable, find the residual to the nearest tenth for

9) Person #50

10) Person #57

11) Person #94

12) Person #111

In the regression analysis involving shoe size as the independent variable and height as the dependent variable, find the residual to the nearest tenth for

13) Person #20

14) Person #34

15) Person #54

16) Person #95

Lecture #22: Goodness of Fit

In earlier lectures we covered testing claims about the population proportion, *p*. We were dealing with categorical variables that were binomial, i.e. having only two categories. But what if a variable is categorical but **not** binomial, if it is **multinomial** – if there are more than two categories and we want to claim certain percentages for the proportions for the different categories, or we want to claim that the percentages for the proportions for the different categories do **not** follow a certain pattern?

In these cases, which are often very interesting situations, we use what's called a **goodness-of-fit test**. Here's an example.

Season of Birth

Are babies born with equal frequency during each of the four seasons, or are different fractions of people born in the different seasons? There are many popular views on the subject, and most people would maintain that the fractions of births **are** different for the different seasons. (People are more likely to be inside and in bed during the long, cold nights of winter, so nine months later....)

So let's claim that the fractions of births in the four seasons are not all the same. This would be the alternative hypothesis, H_1. To see this, look at what the opposite of the claim would be. Let p_{SP} be the fraction of people born in the spring, p_{SU} the summer, p_F the fall, and p_W the winter. To say that the fractions are the same would be $p_{SP} = p_{SU} = p_F = p_W$, and of course that fraction would have to be $\frac{1}{4}$. The complete statement would be $p_{SP} = p_{SU} = p_F = p_W = \frac{1}{4}$. It contains the condition of equality, several times over, and is thus the null hypothesis, H_0. So our claim, that not all the *p*'s are equal to $\frac{1}{4}$, has to be H_1.

We state the hypotheses and identify the claim:

H_0: The same fraction of people are born during each season.

($p_{SP} = p_{SU} = p_F = p_W = \frac{1}{4}$)

H_1: The fractions of birth during the four seasons are not all the same. (Claim)

There are many ways of stating these hypotheses. We don't want to say that each fraction is different from $\frac{1}{4}$, because it could be that two of the seasons each have one-

fourth of the births, but the other two don't. Don't worry too much about stating the hypotheses; just make sure that the claim winds up in the right place.

Now we collect some data. Spring starts with the vernal (spring) equinox on March 20; Summer begins on June 21, the summer solstice; Fall starts on the autumnal equinox, which is Sept. 22, and Winter begins on Dec. 21, the winter solstice.

For our analysis to be valid, we must expect at least five people to be born in each season if the null hypothesis is true. Since the null hypothesis states that the fraction of births in each season is $\frac{1}{4}$, that means we have to have at least $4 \cdot 5 = 20$ people in our survey.

Here are the results of a class survey:

SEASON OF BIRTH	FREQUENCY
Spring	8
Summer	6
Fall	12
Winter	7

As we suspected, Fall had the most births (the highest frequency). But are the numbers unbalanced enough to support our claim that for the population as a whole the fractions are different? After all, there were only 33 people in our sample.

To find out, we do a **goodness-of-fit** test. This is a lovely name, because it asks how good a fit the fractions in the null hypothesis are to the actual fractions we get from our sample.

First, we re-label "Frequency" as "**Observed Frequency**", or simply "O", because these are the frequencies which we **observed** when we did the survey. We want to use another term, "**Expected Frequency**", or "E" for the frequency we would **expect** to get if the null hypothesis (that all the fractions are $\frac{1}{4}$) were true. We find these expected frequencies by multiplying the *p* for each category ($\frac{1}{4}$ in this case) by the sample size *n* (33 in this case). Here's the augmented table:

SEASON OF BIRTH	OBSERVED FREQUENCY O	*p* from Null Hypothesis	EXPECTED FREQUENCY E (= p x n)
Spring	8	0.25	8.25
Summer	6	0.25	8.25
Fall	12	0.25	8.25
Winter	7	0.25	8.25
	n=33		

It's very true that you can't have 8.25 people born in a season, but don't round the expected frequencies to whole numbers, because that would throw off the calculations. If you have to round them, make it to the nearest thousandth unless told to do otherwise. The expected frequencies have to add up to *n*, except if rounding affects the sum slightly. This makes for a good check.

We're working up to a χ^2 here, and the next step is to calculate for each category what is called the **chi-squared contribution**. It has a formula with which you'll become very familiar: $\frac{(O-E)^2}{E}$.

Take the difference between the observed and the expected frequencies for a category, square that difference, and divide by the expected frequency. For Spring, it would be $\frac{(O-E)^2}{E} = \frac{(8-8.25)^2}{8.25} \approx 0.008$.

Here's the table with the $\frac{(O-E)^2}{E}$ column filled in:

SEASON OF BIRTH	OBSERVED FREQUENCY O	*p* from Null Hypothesis	EXPECTED FREQUENCY E (= p x n)	$(O - E)^2 / E$
Spring	8	0.25	8.25	0.008
Summer	6	0.25	8.25	0.614
Fall	12	0.25	8.25	1.705
Winter	7	0.25	8.25	0.189
	n=33			

There's one last step with the table – add up the χ^2 contributions and get $\Sigma\frac{(O-E)^2}{E}$:

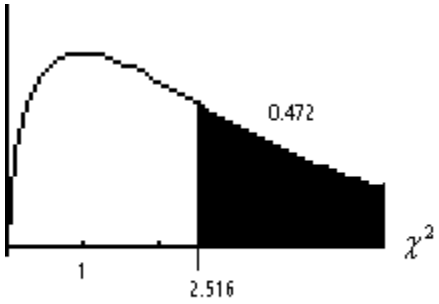
SEASON OF BIRTH	OBSERVED FREQUENCY O	*p* from Null Hypothesis	EXPECTED FREQUENCY E (= p x n)	$(O - E)^2 / E$
Spring	8	0.25	8.25	0.008
Summer	6	0.25	8.25	0.614
Fall	12	0.25	8.25	1.705
Winter	7	0.25	8.25	0.189
	n=33			2.516

The sum, $\chi^2 = 2.516$, is the **test value** for the goodness-of-fit test. We use it to find the likelihood that, if the null hypothesis were true, a group of 33 would produce frequencies as different from the expected, or even more different, as our group did.

Goodness-of-fit tests are almost always right-tailed (later you'll read an example of one that isn't). This is because if, say, the observed frequencies were exactly the same as the expected, $O - E$ would be zero, as would $(O - E)^2$ and $\frac{(O - E)^2}{E}$ and $\Sigma \frac{(O - E)^2}{E}$. The more different the observed frequencies are from the expected, the bigger the $\chi^2 = \Sigma \frac{(O - E)^2}{E}$.

But how many degrees of freedom are there for this χ^2? If you thought 32, then you made a smart mistake, because you concluded from previous χ^2 work that the degrees of freedom are one less than the sample size, which was 33. However, in goodness-of-fit tests, the degrees of freedom are one less than the **number of categories**, which we label *k*. In this case, with four seasons, $k = 4$. So there are three degrees of freedom.

To find the *p*-value, we use χ^2 cdf (2.516, 1000, 3), which is approximately 0.472. Here's a diagram of the distribution showing the χ^2 test value and the *p*-value:



As you can see, there's more than enough probability that seasons with equal births would produce a sample as unbalanced as ours or even more unbalanced. We **do not**

reject the null hypothesis. To summarize: There is **not** sufficient evidence to **support** the claim that the fractions of birth during the four seasons are not all the same. We failed to make our case. Let's ask for more research money so we can collect a larger sample!

Are the Dice Fair?

In our last example, the *p*'s of the null hypothesis were the same for all the seasons, so the expected frequencies were all equal. This doesn't have to be the case to perform a goodness-of-fit test. To demonstrate this, we go back to an unanswered question from the early part of the course: Are the dice you rolled fair?

You rolled two dice and recorded their sum. Here's the chart for how the rolls could turn out:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	1,1	1,2	1,3	1,4	1,5	1,6
	2	2,1	2,2	2,3	2,4	2,5	2,6
	3	3,1	3,2	3,3	3,4	3,5	3,6
	4	4,1	4,2	4,3	4,4	4,5	4,6
	5	5,1	5,2	5,3	5,4	5,5	5,6
	6	6,1	6,2	6,3	6,4	6,5	6,6

And here's the chart for the sum which would result from each roll:

		Die #2					
		1	**2**	**3**	**4**	**5**	**6**
Die #1	**1**	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

From these charts, we deduced that if the dice were fair, the probabilities of getting a certain sum (from 2 to 12) are those shown in this table:

SUM	p
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Let's claim that are dice **were** in fact fair dice, even though they came from the dollar store, and test it at the 10% significance level. This claim will be the null hypothesis, because it says that the probability of each sum **is equal to** the fraction given in the table above.

Here are the hypotheses with the claim labeled:

H_0: The dice are fair. (The population proportions for the different sums are equal to the ones listed in the table.) (Claim)
H_1: The dice aren't fair.

I totaled up your frequencies for the different sums, and we'll use these numbers for the observed frequencies in the table showing *O*, *E*, and $\frac{(O-E)^2}{E}$. Here's how the numbers look:

SUM	OBSERVED FREQUENCY O
2	162
3	395
4	621
5	816
6	1053
7	1363
8	1062
9	901
10	687
11	462
12	271

n = 7793

Now we'll put in the expected frequencies. For instance, the expected frequency for a sum of 2 is found by multiplying $\frac{1}{36}$ by 7793, which comes to 216.5, rounded to the nearest tenth.

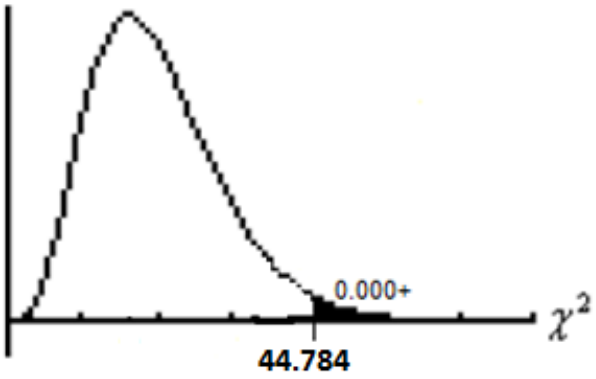
SUM	OBSERVED FREQUENCY O	*p* from Null Hypothesis	EXPECTED FREQUENCY E (= p x n)
2	162	1/36	216.5
3	395	2/36	432.9
4	621	3/36	649.4
5	816	4/36	865.9
6	1053	5/36	1082.4
7	1363	6/36	1298.8
8	1062	5/36	1082.4
9	901	4/36	865.9
10	687	3/36	649.4
11	462	2/36	432.9
12	271	1/36	216.5
	n = 7793		

The next step is to get the χ^2 contribution, $\frac{(O-E)^2}{E}$, for each sum, and then to add them:

SUM	OBSERVED FREQUENCY O	*p* from Null Hypothesis	EXPECTED FREQUENCY E (= p x n)	$(O - E)^2 / E$
2	162	1/36	216.5	13.707
3	395	2/36	432.9	3.326
4	621	3/36	649.4	1.243
5	816	4/36	865.9	2.874
6	1053	5/36	1082.4	0.796
7	1363	6/36	1298.8	3.170
8	1062	5/36	1082.4	0.383
9	901	4/36	865.9	1.424
10	687	3/36	649.4	2.175
11	462	2/36	432.9	1.950
12	271	1/36	216.5	13.735
	n = 7793			44.784

So the χ^2 test value is 44.784, and there are 10 degrees of freedom ($k-1=11-1=10$).

To find the *p*-value, we use χ^2 cdf (44.784, 1000, 10), which is 2.4×10^{-6}, or 0.000^{+}. Here's a diagram of the distribution showing the χ^2 test value and the *p*-value:



Since the *p*-value is less than α, we must **reject** the null hypothesis and say that there **is** sufficient evidence to **reject** the claim that the dice are fair. **The dice are not fair.**

After rejecting the distribution laid out in the null hypothesis, you can then theorize about the ways the actual distribution differs from it. Looking at the observed *versus* the expected frequencies, we notice that in general we got fewer small sums than expected and more large sums. So the dice appear to favor the sides with more dots on them landing up, and those with fewer dots landing on the bottom. (Remember, opposite sides add up to 7: 1 and 6, 2 and 5, 3 and 4.) Because these are cheap dice, the craters for the dots are left unfilled, so the sides with more craters are lighter, or less dense, whereas those with fewer craters are heavier and hence would tend to land facing down while their less dense opposite sides face up. In a casino, where fair dice are a must, the craters are filled with a substance of the same density as the rest of the die but a different color, to assure uniform density, and hence fairness.

Left-tailed Goodness-of-Fit Tests

Let's say that your instructor gives you a pair of dice and tells you to go home and roll them 7,234 times and record the sums. You decide that's too much work, and instead you'll make up the observed frequencies. For guidance, you look at the expected frequencies. Hmmm, for 2 it's 216.5, so you pick 217. For 3, it's 432.9, so you pick 433. And so on, until for 12 you subtract the sum of all the other expected frequencies from 7,793 and put down what's left. Your $\frac{(O-E)^2}{E}$'s are all going to be very small, and if you find χ^2 cdf (0, $\Sigma\frac{(O-E)^2}{E}$, 10), it will yield a very small *p*-value indeed. Your "observed" frequencies were just too close to the expected frequencies to be believable. It will be obvious that you made them up. So if you **are** going to cheat, make the "observeds" a little more different from the "expecteds."

A famous example of this occurred in the researches of geneticist Gregor Mendel, who studied the genetics of pea plants, and in doing so produced a sample with observed frequencies of the traits he was assessing which were far too close to his hypothesized 3-to-l ratio to be believable, in the same way that the made-up dice sums above were. However, Mendel's experiments have been repeatedly replicated and have supported his theories. One explanation is that he found the 3-to-1 ratio in an early, small-sample study, and that he kept going until he had a large sample that confirmed the ratio, a situation called **confirmation bias**. There are other theories as well, but nobody is accusing Mendel of cheating.

Activity #22: Goodness of fit

The manufacturers of M&Ms claim that the six colors M&Ms occur in the following proportions: Blue 24%, Orange 20%, Green 16%, Yellow 14%, Red 13%, and Brown 13%. Use your sample to perform a goodness-of-fit test on this claim. Let $\alpha = 0.05$.

a) **State the hypotheses (and identify the claim),**
b) **Fill in the chart for O, E, and $\frac{(O-E)^2}{E}$, and make a diagram showing the χ^2 test value and the *p*-value (to the nearest thousandth),**
c) **State your decision whether to reject H_0, and**
d) **Summarize the results.**

a) H_0: The color distribution of M&Ms is as described. (**Claim)** ($p_{blue} = 24\%$, $p_{orange} = 20\%$, $p_{green} = 16\%$, $p_{yellow} = 14\%$, $p_{red} = 13\%$, $p_{brown} = 13\%$)

H_1: The color distribution is M&Ms is **not** as described.

b)

Color	**Observed F O**	***p***	**Expected F E = *p* x *n***	$\frac{(O-E)^2}{E}$
Blue		0.24		
Orange		0.20		
Green		0.16		
Yellow		0.14		
Red		0.13		
Brown		0.13		

$\sum O = n =$ $\chi^2 = \sum \frac{(O-E)^2}{E} =$

c)

d)

Assignment #10

For each problem,

a) **State the hypotheses (and identify the claim),**

b) **Fill in the chart for O, p, E, and $\frac{(O-E)^2}{E}$ (round E's and $\frac{(O-E)^2}{E}$'s to the nearest thousandth if necessary, and make a diagram showing the χ^2 test value and the *p*-value (to the nearest thousandth),**

c) **State your decision whether to reject H_0, and**

d) **Summarize the results.**

1) A children's raincoat manufacturer claims that customers prefer certain colors over others. He selected a random sample of 50 raincoats sold and noted their colors. There were 21 yellow, 12 red, 9 green, and 8 blue. Test his claim at $\alpha = 0.10$.

2) According to a recent census report, 68% of families have two parents present, 23% have only a mother present, 5% have only a father present, and 4% have neither parent present. A random sample of families from a large school district revealed that 146 had two parents, 34 mother only, 7 father only, and 13 neither. Test the claim that the school district's distribution differs from the census at $\alpha = 0.01$.

3) The population distribution of federal prisons nationwide by serious offenses is the following: violent offenses, 12.6%; property offenses, 8.5%; drug offenses, 60.2%; weapons, 8.2%; immigration, 4.9%; other, 5.6%. A warden claims that his prison has the same percentages. A survey of 500 prisoners at his prison finds 64 violent offenders, 40 property offenders, 326 drug offenders, 42 weapons offenders, 25 immigration offenders, and 3 with other offenses. Test the warden's claim at $\alpha = 0.05$.

4) A quality control engineer claims that defective items are manufactured with the same frequency each day of the work week. A week is randomly selected at her factory, and the number of defective items produced each day is recorded. There were 32 on Monday, 16 on Tuesday, 23 on Wednesday, 19 on Thursday, and 40 on Friday. Test her claim at $\alpha = 0.05$.

Lecture #23: Tests of Independence

In Lecture #8, while studying probability, we developed a formula for P(A **or** B), the probability that one or both of the two events A and B happens. The formula was P(A **or** B) = P(A) + P(B) – P(A **and** B). Perhaps you wondered then why we didn't do the same for P(A **and** B).

In fact, there **is** a formula for P(A **and** B), but it's true only in certain cases, and when it **is** true, we call events A and B **independent**. The formula is P(A **and** B) = P(A) x P(B). If they're not independent, if P(A **and** B) $\neq$ P(A) x P(B), we say that they are **dependent**, or **related**. Informally, if two events are independent, they have no influence over each other. Whether one takes place or not has no effect on whether the other one occurs. Their probabilities do not interact.

I'm going to give two examples, one where the events **are** independent and one where they are dependent.

Here's the first probability experiment: I have four pieces of paper in my pocket. They're labeled 1, 2, 3, and 4. I take one out and look at it, note its number, put it back, pick one out again and look at it, and note its number. I'm as likely to pick any number as any other on any turn – i.e., this is classical probability. (When I put the first number back before selecting the second number, so in fact one number could be chosen twice, we call that **with replacement**.) So the outcome 31 means that first I chose the 3 and then I chose the 1. Here's the sample space:

		Second Pick			
		1	**2**	**3**	**4**
First Pick	**1**	11	12	13	14
	2	21	22	23	24
	3	31	32	33	34
	4	41	42	43	44

There are 16 outcomes in the sample space, so $n(S) = 16$. Let A be the event of picking a 3 on the first pick, and let B be the event of picking a 4 on the second pick.

What is P(A **and** B)? Clearly it's $\frac{1}{16}$, because there is precisely one outcome in the event A **and** B:

		Second Pick			
		1	**2**	**3**	**4**
First Pick	**1**	11	12	13	14
	2	21	22	23	24
	3	31	32	33	**34**
	4	41	42	43	44

What is P(A)? It's $\frac{4}{16} = \frac{1}{4}$, since the whole row beginning 31 has 3 picked first:

		Second Pick			
		1	**2**	**3**	**4**
First Pick	**1**	11	12	13	14
	2	21	22	23	24
	3	**31**	**32**	**33**	**34**
	4	41	42	43	44

How about P(B)? It's also $\frac{4}{16} = \frac{1}{4}$, since the whole column beginning 14 has 4 picked second:

		Second Pick			
		1	**2**	**3**	**4**
First Pick	**1**	11	12	13	**14**
	2	21	22	23	**24**
	3	31	32	33	**34**
	4	41	42	43	**44**

Putting it all together, P(A **and** B) = $\frac{1}{16} = \frac{1}{4} \times \frac{1}{4}$ = P(A) x P(B), and the events A and B are independent because they satisfy the equation that defines independence. Also, you can see that they have no influence on each other because the number chosen first is eligible to be chosen again.

Now let's change the probability experiment slightly. This time I **won't** put the first number back before picking the second. We call this arrangement **without replacement** for obvious reasons. This means that outcomes like 11, 22, etc., are impossible. The sample space this time is

		Second Pick			
		1	**2**	**3**	**4**
First Pick	**1**		12	13	14
	2	21		23	24
	3	31	32		34
	4	41	42	43	

and $n(S) = 12$, because the outcomes along the **main diagonal** (upper left to lower right) have been eliminated.

Using A and B as we did in the first version of the experiment, P(A **and** B) = $\frac{1}{12}$, because there's still only one outcome in the event A **and** B, but this time it's out of a total of 12 possible outcomes:

		Second Pick			
		1	**2**	**3**	**4**
First Pick	**1**		12	13	14
	2	21		23	24
	3	31	32		**34**
	4	41	42	43	

Meanwhile, P(A) and P(B) remain $\frac{1}{4}$. P(A) = $\frac{3}{12} = \frac{1}{4}$, as illustrated below, and so is P(B).

		Second Pick			
		1	**2**	**3**	**4**
First Pick	**1**		12	13	14
	2	21		23	24
	3	**31**	**32**		**34**
	4	41	42	43	

So P(A **and** B) = $\frac{1}{12}$, but P(A) x P(B) = $\frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$. Now A and B are **not** independent, because the equation for independence isn't true any more.

In the experiment without replacement, what I pick the first time has an effect on what I pick the second time. Another way to say this is that my second pick is **dependent** on my first pick. An extreme example is that if I pick a 3, say, on the first pick, the probability that I pick a 3 on the second pick is 0, not $\frac{1}{4}$, as it would be if I used replacement.

A Test of Independence

In real life, things aren't as cut-and-dried as they are in classical probability. Consider the empirical situation we used earlier in the semester, the contingency table for Sex and Number of Pets:

		Number of Pets			
		No Pets	**1 or 2 Pets**	**3 or 4 Pets**	**At Least 5 Pets**
Sex	**Male**	13	24	12	4
	Female	8	34	12	13

Is the pet-owning behavior of men different from that of women, or are sex and pet ownership independent? The frequency patterns **look** different, but are they different enough to say that sex and pet ownership are **related**, or, to put it another way, that pet ownership is **dependent** upon the sex of the owner? In answering this question, we'll encounter a lot of familiar concepts like observed and expected frequencies, the χ^2 contribution $\frac{(O-E)^2}{E}$, and the usual parts of claims testing, but we'll also need some new concepts.

Let's claim that the level of pet ownership **is dependent upon** the sex of the pet owner, and let's test that claim at the 10% significance level.

First we'll state the hypotheses and identify the claim. If our claim is that pet ownership is dependent upon the sex of the owner, which hypothesis is it? Since the definition of the independence of two events is that P(A **and** B) = P(A) x P(B), and the definition of dependence is that P(A **and** B) $\neq$ P(A) x P(B), clearly a claim of dependence has to be the alternative hypothesis, H_1.

H_0: Sex and pet ownership are independent.

H_1: Sex and pet ownership are related, or pet ownership is dependent upon the sex of the pet owner. (Claim)

(Note that to say that one attribute depends on another is not necessarily to say that one **causes** the other. It might or it might not. But certainly if there is any causal relationship here, it's the sex of the pet owner that causes the level of pet ownership, not *vice versa*.)

Now to make a table showing *O*, *E*, and $\frac{(O-E)^2}{E}$. This is a familiar process from the last lecture, and it turns out to be applicable to tests of independence as well. Each cell of the contingency table produces one row of this table. We could start like this:

CELL	OBSERVED FREQUENCY O
Male None	13
Male 1 or 2	24
Male 3 or 4	12
Male at Least 5	4
Female None	8
Female 1 or 2	34
Female 3 or 4	12
Female at Least 5	13

It doesn't matter what order we list the cells in.

Now we hit a snag. If the null hypothesis is true, that sex and pet ownership are independent, how do we find the expected frequencies? This is going to be a long story, but at the end there's a simple little pattern that you'll follow to get the expected frequencies.

Let's concentrate on one cell of the contingency table, the females with 3 or 4 pets. Call being female event A, and having 3 or 4 pets event B. If they're independent, then P(A **and** B) = P(A) x P(B), or P(Female **and** 3 or 4 Pets) = P(Female) x P(3 or 4 Pets). We covered how to find P(Female) and P(3 or 4 Pets) in Lecture #8. First we find the row, column and grand totals for the contingency table:

		Number of Pets				
		None	**1 or 2**	**3 or 4**	**At least 5**	**Sum**
Sex	**Male**	13	24	12	4	**53**
	Female	8	34	12	13	**67**
	Sum	**21**	**58**	**24**	**17**	**120**

So P(Female) = $\frac{67}{120}$, and P(3 or 4 Pets) = $\frac{24}{120}$, and if the events are independent, their joint probability (the **and** option) is the product of their separate probabilities, so P(Female **and** 3 or 4 Pets) = $\frac{67}{120} \times \frac{24}{120}$. I won't multiply this out, but I hope you can see that if $\frac{67}{120} \times \frac{24}{120}$ is the probability for that category, then the expected frequency for the category is $p \times n = \frac{67}{120} \times \frac{24}{120} \times 120$. (If this seems unclear, imagine that the p for the cell is $\frac{1}{2}$. Then the expected frequency would be $p \times n = \frac{1}{2} \times 120$.)

As you can see, in $\frac{67}{120} \times \frac{24}{120} \times 120$, one of the 120's on the bottom would cancel with the 120 at the end, and the expression we'd wind up with would be $\frac{67 \times 24}{120}$. Take a close look at this. It is the **product of the row total for the category described in the rows and the column total for the category described in the columns, divided by the grand total**. So for any cell, to compute its expected frequency, go all the way to the right to get the row total, go all the way down to get the column total, multiply those two numbers, and then go diagonally down and to the right to the grand total, and, finally, divide the product you got by that grand total. This may sound complicated, but if you do it enough it will become second nature. Here's the completed *O*, *E*, and $\frac{(O-E)^2}{E}$ table, including the sum of the χ^2 contributions:

CELL	OBSERVED FREQUENCY O	How to Find E	EXPECTED FREQUENCY E	$(O - E)^2 / E$
Male None	13	53 x 21 / 120	9.3	1.5
Male 1 or 2	24	53 x 58 / 120	25.6	0.1
Male 3 or 4	12	53 x 24 / 120	10.6	0.2
Male at Least 5	4	53 x 17 / 120	7.5	1.6
Female None	8	67 x 21 / 120	11.7	1.2
Female 1 or 2	34	67 x 58 / 120	32.4	0.1
Female 3 or 4	12	67 x 24 / 120	13.4	0.1
Female at Least 5	13	67 x 17 / 120	9.5	1.3
				6.1

This is a right-tailed test, for the same reasons explained in the last lecture, but before we use χ^2 cdf we have to determine the number of degrees of freedom. What's that you say? Seven? In other words $k-1$, or one less than the number of categories? Another smart mistake! You'd be right if this were simply a situation in which eight categories are presented, but because of the contingency-table arrangement we are free to choose considerably fewer than seven of the eight numbers.

Let's start with the contingency table empty but with its various sums as fixed numbers:

		Number of Pets				
		None	**1 or 2**	**3 or 4**	**At least 5**	**Sum**
Sex	**Male**					**53**
	Female					**67**
	Sum	**21**	**58**	**24**	**17**	**120**

Surely I can pick 13 for 'Male None' (I can pick any number from 0 to 21, for that matter).

		Number of Pets				
		None	**1 or 2**	**3 or 4**	**At least 5**	**Sum**
Sex	**Male**	**free**:13				**53**
	Female					**67**
	Sum	**21**	**58**	**24**	**17**	**120**

"Free" means I was able to choose it. But I'm not free to pick the 8 for 'Female None,' because it has to be the number that brings the 13 up to 21.

		Number of Pets				
		None	**1 or 2**	**3 or 4**	**At least 5**	**Sum**
Sex	**Male**	**free**:13				**53**
	Female	forced: 8				**67**
	Sum	**21**	**58**	**24**	**17**	**120**

I can pick 24 for 'Male 1 or 2' and 12 for 'Male 3 or 4,' but then my hands are tied for 'Female 1 or 2' and 'Female 3 or 4.'

		Number of Pets				
		None	**1 or 2**	**3 or 4**	**At least 5**	**Sum**
Sex	**Male**	**free**:13	**free**:24	**free**:12		**53**
	Female	forced:8				**67**
	Sum	**21**	**58**	**24**	**17**	**120**

When I say that I am free to pick 'Male 1 or 2' and 'Male 3 or 4,' of course I'm not entirely free, because I have to work with the column totals so as not to exceed them, and I also have to keep an eye on the row totals so I don't violate them either. But I do have **some** choice.

But now I'm finished. 'Male At Least 5' has to make 13, 24, and 12 add up to 53, so it must be 4, and 'Female At Least 5' is then determined as all the Female cells were after we chose or calculated the Male numbers.

		Number of Pets				
		None	**1 or 2**	**3 or 4**	**At least 5**	**Sum**
Sex	**Male**	**free**:13	**free**:24	**free**: 12	forced: 4	**53**
	Female	forced:8	forced:34	forced:12	forced: 13	**67**
	Sum	**21**	**58**	**24**	**17**	**120**

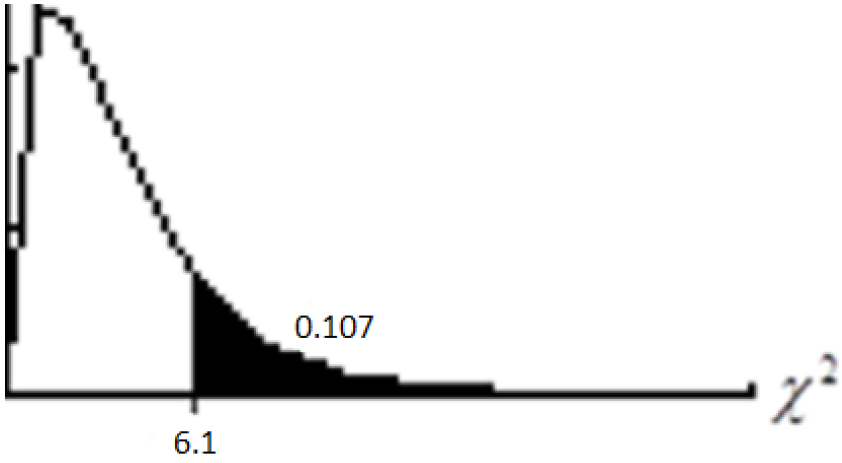
So there are exactly 3 degrees of freedom in this χ^2 test. Before we finish the test, let's continue studying the degrees of freedom, because you can see that our method of determining their number is a bit cumbersome, to say the least. There must be a better way, and there is. Let *r* be the number of rows in a contingency table and *c* the number of columns. I can fill in all but the last cell in the first row, and I can do that in all but the final row, which is a complete wash-out. So I can fill in all the cells in a contingency table that has $r-1$ rows and $c-1$ columns. This makes a total of $(r-1)\times(c-1)$ cells. In other words, d.f. = $(r-1)\times(c-1)$. In our case this comes to $(2-1)\times(4-1)=1\times 3=3$ degrees of freedom, as we found earlier.

Just for practice, what if you had a contingency table with 5 rows and 3 columns?

You'd have $(r-1)\times(c-1)=(5-1)\times(3-1)=4\times 2=8$ degrees of freedom. Here's confirmation:

	A	B	C	SUM
1	**Free**	**Free**	Forced	**Given**
2	**Free**	**Free**	Forced	**Given**
3	**Free**	**Free**	Forced	**Given**
4	**Free**	**Free**	Forced	**Given**
5	Forced	Forced	Forced	**Given**
SUM	**Given**	**Given**	**Given**	**Given**

Now we can get back to testing our claim that level of pet ownership and sex of the pet owner are related. To find the *p*-value, we use χ^2 cdf (6.1, 1000, 3) ≈ 0.107. Here's a diagram showing the χ^2 test value and the *p*-value:



Since 0.107 > 0.10, or $p < \alpha$, we **don't reject** the null hypothesis. To summarize: There **isn't** sufficient evidence to **support** the claim that sex and pet ownership are related, or that pet ownership is dependent upon the sex of the pet owner.

Activity #23: Tests of Independence

Test this claim: The type of pet owned is dependent on annual household income. Significance level: 1%.

a) **State the hypotheses (and identify the claim),**
b) **Fill in the chart for O, E, and $\frac{(O-E)^2}{E}$, and find $\Sigma\frac{(O-E)^2}{E}$ and the *p*-value.**
c) **State your decision whether to reject H_0, and**
d) **Summarize the results.**

Here's the table:

	Type of Pet		
Income ($)	**Dog**	**Bird**	**Horse**
Under 25,000	318	382	298
25,000-59,999	431	395	449
60,000 and Over	254	223	254

Lecture #24: Analysis of Variance

In previous lectures we've covered tests of claims about the population mean μ and about the differences in the population means of two populations. But what if you want to test whether there are differences in the population means of more than two populations?

Why don't we just pair off the various populations and test for differences in means for all the pairs? There's a problem there. Say you have four populations you're interested in. Four populations make ${}_4C_2 = 6$ pairs. So you have to do six tests. What if you're conducting the tests at the 5% significance level? Then there's a 95% chance that you **won't** make a Type I error on any given test. But the chance that you won't make a Type I error on **any** of the tests is $(0.95)^6 \approx 0.74$, because you have to multiply the chance that you don't make a Type I error on the first test by the chance that you don't make a Type I error on the second test by the ... you get the idea. The significance level for these tests as a group will be $1-0.74=0.26$, or 26%. Not what we want – the risk of having made a Type I error somewhere is too great.

To counter this difficulty, a technique called **analysis of variance** was invented. It appeared first in 1918 in a paper by R. A. Fisher, a British statistician and geneticist. It has the nickname **ANOVA**.

Here's an example to give you an idea of the concepts involved. Let's say you have three different populations, A, B, and C, and you take a sample of size 3 from each population. You're interested in the population means of these populations. You're claiming that these population means are **not** all the same.

Compare two scenarios, which I call Set 1 and Set 2:

Set 1			Set 2		
A	**B**	**C**	**A**	**B**	**C**
5	10	15	9	14	19
10	15	20	10	15	20
15	20	25	11	16	21

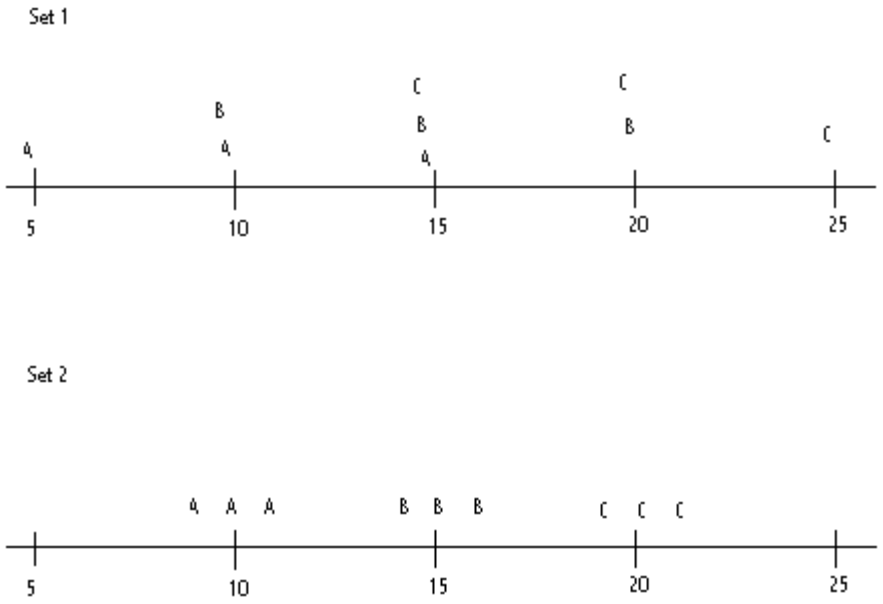
In which case would you be more likely to conclude that the population means **aren't** the same? If you said Set 2, you have the right idea.

As you can see, the sample means are the same for Sets 1 and 2:

	Set 1			Set 2		
	A	**B**	**C**	**A**	**B**	**C**
	5	10	15	9	14	19
	10	15	20	10	15	20
	15	20	25	11	16	21
Sample Mean	10	15	20	10	15	20

However, the numbers in Set 1 are all spread out and overlapping, whereas those in Set 2 are tightly grouped around the means and make you believe that they might actually come from populations with different population means.

Here are graphs of the two sets:



The basic idea of analysis of variance is to compare the variability of the sample means (which we call the variance **between** groups) to the variability of the samples themselves (which we call the variance **within** groups). If the former is large compared to the latter, as in Set 2, we feel that there really **are** differences among the population means, but if the variability between groups is **not** large compared to the variability within groups, we're **not** going to conclude that there are differences among the population means. In the second case we say that there is too much "noise" to draw a conclusion about the differences.

Let's use the **range** (the difference between the largest number and the smallest number) as a measure of variability. For both Set 1 and Set 2, the sample means have a range of 10. That is a measure of the variability **between** groups. But for Set 1, the variability **within** the groups, if measured by the range, is 10 (e.g. $15-5$), whereas for Set 2, the variability within the groups measured this way is 2 (e.g. $11-9$). So compared to the variability within groups, the variability between the groups is much larger (five times larger) for Set 2 than it is for Set 1 (where the two are the same).

The comparison of variance between groups and variance within groups is done by using a ratio, so ANOVA winds up using the *F*-distribution, the last of the four great pdf's we use in statistics.

Here's the formula for finding the *F* test value:

$$F = \frac{s_B^{\,2}}{s_W^{\,2}} = \frac{\dfrac{\Sigma n_i \left(\bar{x}_i - \bar{x}_{GM}\right)^2}{k-1}}{\dfrac{\Sigma \left(n_i - 1\right) s_i^{\,2}}{\Sigma \left(n_i - 1\right)}}$$

The *k* you've seen before – it's the number of groups or categories. If you take the mean of all the data, it's called the **grand mean** and written $\bar{x}_{GM}$. The variance **between** groups is $s_B^{\,2}$, and the variance **within** groups is $s_W^{\,2}$.

Can you see an easier way to calculate $\Sigma \left(n_i - 1\right)$? You're adding up one less than the sample size for each sample. Since there are *k* samples, you're adding up one less *k* times, so $\Sigma \left(n_i - 1\right) = n - k$, where *n* is the total of all the sample sizes.

The *F* distribution has two degrees of freedom – one for the numerator and one for the denominator. In the ANOVA formula, the number of degrees of freedom for the numerator is $k - 1$ (the denominator of the numerator), and the number of degrees of freedom for the denominator is $n - k$ or $\Sigma \left(n_i - 1\right)$ (the denominator of the denominator).

Glossary of Symbols

Number in parentheses refers to lecture the symbol first appears in.

f Frequency (3)

$\sum$ Sum of; the Greek letter capital sigma (3)

n Sample size (3)

N Population size (3)

$\frac{f}{n}$ Relative frequency (3)

$\bar{x}$ Sample mean, $\frac{\sum x}{n}$ (4)

μ Population mean, $\frac{\sum x}{N}$; Greek letter mu (4)

s Sample standard deviation, $\sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$ (5)

s^2 Sample variance, $\frac{\sum (x-\bar{x})^2}{n-1}$ (5)

σ Population standard deviation, $\sqrt{\frac{\sum (x-\mu)^2}{N}}$; Greek letter lower-case sigma (5)

σ^2 Population variance, $\frac{\sum (x-\mu)^2}{N}$ (5)

C_{var} Coefficient of variation, $\frac{s}{\bar{x}}$ (5)

Q_1 First quartile (6)

Q_3 Third quartile (6)

IQR Interquartile range, $Q_3 - Q_1$ (6)

P_1 First percentile (6)

z Z-score, $z = \frac{x-\bar{x}}{s}$ (6)

S Sample space (7)

$n(S)$ Size of sample space S (7)

E Event (7)

ϕ Empty set (7)

$n(E)$ Size of event E (7)

${}_nP_r$ Number of permutations of n things taken r at a time (9)

! Factorial (9)

${}_nC_r$ Number of combinations of n things taken r at a time (9)

$P(E)$ Probability of event E, $\frac{n(E)}{n(S)}$ (9)

CL Confidence level (13)

p Population proportion (14)

$\hat{p}$ Sample proportion, $\frac{x}{n}$ (14)

α Significance level, Greek lowercase letter alpha (15)

t t distribution (16)

df Degrees of freedom (16)

χ^2 Chi-square (18)

F F distribution (19)

r Sample correlation coefficient (20)

ρ Population correlation coefficient; Greek letter rho (20)

b Sample slope of regression line (20)

Student Survey

1) What sex are you (M or F)?

2) Are you a native of Mendocino or Lake County?

3) Are you a graduate of (or currently attending) a Mendocino or Lake County high school?

4) What is your favorite color?

5) What is your zip code?

6) Which statement best characterizes your attitude towards taking math classes?
   1) If I never have to take another math class it will be too soon.
   2) I would prefer not to take a math class.
   3) I can take them or leave them.
   4) I enjoy math classes.
   5) I adore taking math; it's my favorite thing to study.

7) Which statement best characterizes your attitude towards social media? (Facebook, Snapchat, Instagram, Twitter, etc.)
   1) I never use social media and can't stand them.
   2) I seldom use social media.
   3) I can take them or leave them.
   4) I enjoy social media.
   5) I love using social media; I can't get through the day without them.

8) If you own a car, what year is it?

9) What size shoes do you wear?

10) How old did you turn on your last birthday?

11) What is your height to the nearest inch?

12) How many pets do you have?

MATH 220 -- SPRING 2018 CLASS DATA BASE

#	Sex	Native	Grad	Color	Zip	Math	Social Media	Car	Shoes	Age	Height	Pets
1	M	N	N	Brown	95490	3	2		13	22	72	0
2	M	Y	Y	Black	95482	4	1	1995	9.5	22	70	4
3	M	N	Y		95482	5	3		9	16	68	1
4	M	Y	Y	Green	95458	4	4	2015	10	26	70	3
5	M	N	N	Blue	95482	2	4	2011	13	20	71	6
6	M	N	N	Burgundy	95482	4	3	2017	10	18	69	2
7	M	Y	Y	Black	95482	3	3	1998	10.5	18	71	2
8	M	Y	N	Purple	95453	1	5		11.5	19	69	3
9	M	N	N	Grey	95482	3	4	2018	10.5	36	71	0
10	M	Y	N	White	95422	3	2	2003	12	22	75	3
11	M	Y	Y	Blue	95482	2	5	2005	10	21	66	2
12	M	N	Y	Blue	95481	4	3		8	18	66	0
13	M	Y	Y	Burnt Orange	95482	4	3	2006	11	20	69	2
14	M	Y	Y	Purple	95469	3	4	2002	11.5	19	71	4
15	M	Y	N	Red	95482	4	4	2016	11.5	19	71	2
16	M	Y	Y	Black	95482	4	5	2014	12	19	70	1
17	M	Y	Y	Blue	95482	3	4	2015	8.5	19	64	0
18	M	Y	Y	Black	95470	4	3	2014	13	19	74	5
19	M	Y	Y	Wine Red	95453	4	2	2003	9.5	20	66	5
20	M	N	Y	Metallic Blue	95482	2	5	2004	8	20	70	3
21	M	Y	Y	Black	95482	2	2	2006	12	26	71	1
22	M	Y	Y	Grey	95482	4	1	1998	11.5	22	76	1
23	M	Y	Y		95482	3	2	2000	10	26	71	1
24	M	Y	Y	Black	95482	2	3	2007	8.5	26	63	0
25	M	N	Y	Brass	95482	3	3	1997	10	24	72	2
26	M	N	Y	Dark Blue	95482	4	4	2003	9	19	71	0
27	M	Y	Y	Waluigi Purple	95482	4	4		9.5	19	70	3
28	M	Y	Y	Blue	95482	3	4		12	18	74	3
29	M	Y	N	Dark Emerald	95482	4	3	2016	16	24	77	1
30	M	Y	Y	Teal	95482	3	4	2002	10.5	19	72	1
31	M	Y	Y	Red	95482	3	3	2003	11	19	72	1
32	M	N	Y	Blue	95482	3	3	2013	10.5	26	71	1
33	M	N	Y	Red	95482	4	3	2001	10.5	31	71	0
34	M	Y	Y	Midnight Blue	95482	2	2	2003	11	18	70	0
35	M	N	N	Black	95482	1	3	1998	11.5	26	72	0
36	M	N	Y	Olive Drab	95490	2	2		12	19	72	2
37	M	N	N	Blue	95470	3	2	2004	12	39	73	1
38	M	Y	Y	Blue	95482	4	3	2014	10	22	69	2

39	M	N	Y	Mahogany Red	95490	3	3	2012	12	18	70	3
40	M	N	Y	Black	95490	3	4	2004	8.5	19	69	1
41	M	N	N	Orange	95490	4	3		10.5	31	70	2
42	M	N	N	Turquoise	95482	2	4	2001	11	20	69	1
43	M	N	Y	Green	95482	4	3		11	19	68	2
44	M	Y	Y	Black	95482	3	4	2005	10	18	70	0
45	M	Y	Y	Black	95482	4	4	2015	11.5	18	74	2
46	M	Y	Y	Blue	95482	3	3	2008	9.5	24	68	4
47	M	Y	Y	Red	95449	2	4	2001	12.5	19	71	3
48	M	Y	Y	Blue	95482	4	4	2000	12.5	20	71	3
49	M	Y	Y	Black	95482	5	4	2014	8	19	65	2
50	M	N	Y	Black	95482	3	4	2004	9.5	18	61	0
51	M	Y	Y	Blue	95482	3	2		12	18	72	6
52	M	N	N	Red	95482	2	4		10	24	72	0
53	M	N	N	Green	95444	2	2	2009	9.5	25	70	0
#	Sex	Native	Grad	Color	Zip	Math	Social Media	Car	Shoes	Age	Height	Pets
54	F	N	N	Purple	95490	4	3	2004	6.5	51	60	4
55	F	N	N	Red	95470	4	4	2012	5.5	20	57	0
56	F	Y	Y	Purple	95482	4	4	2016	7	56	65	3
57	F	N	N	Purple	95437	4	4	2002	6	39	60	2
58	F	Y	Y	Turquoise	95482	3	4	2008	7.5	31	67	4
59	F	Y	Y	Pink	95482	3	4	2016	11	27	70	3
60	F	Y	Y	Champagne	95470	4	4	2015	7	37	64	2
61	F	Y	Y	Black	95490	2	5	2000	8	46	63	1
62	F	Y	Y	Black	95482	2	4		4.5	19	62	0
63	F	Y	Y	Pink	95453	3	5	2015	8.5	20	70	2
64	F	Y	Y	Pink	95482	4	4	2015	7.5	22	62	1
65	F	N	N	Teal	95482	4	4	2003	9.5	29	66	1
66	F	N	N	Teal	95468	4	4	2016	8.5	33	65	10
67	F	N	N	Purple	95451	3	4	2003	10	41	71	2
68	F	Y	Y	Green	95485	2	5	2015	9	15	67	7
69	F	Y	Y	Green	95482	4	4		11	15	71	3
70	F	Y	Y	Blue	95428	2	4	2007	9	57	61	2
71	F	Y	N	Purple	95482	4	4	2005	8	47	67	4
72	F	N	Y	Red	95482	4	4	2003	7.5	19	64	2
73	F	Y	N	Green	95485	4	1	2001	10	31	69	7
74	F	N	Y		95482	2	2	2014	8	47	65	5
75	F	Y	Y	Purple	95470	3	2	2017	7	19	63	7
76	F	Y	Y	Black	95470	2	4	2015	7.5	22	63	2
77	F	Y	Y	Pink	95470	4	4	2015	7.5	19	63	1

78	F	Y	Y	Midnight	95482	2	4		11	18	67	2
79	F	Y	Y	Light Blue	95482	4	5		7	18	64	2
80	F	Y	Y	Blue	95482	2	3	2015	7	21	62	3
81	F	N	N	Teal	95482	4	3	2011	9	27	67	4
82	F	Y	Y	Blue	95482	3	4	2001	10.5	19	68	1
83	F	Y	Y	Green	95453	4	4	2013	9.5	19	64	0
84	F	Y	Y	Teal	95482	3	4	2017	6.5	20	61	2
85	F	N	Y	Purple	95437	2	4	2014	8.5	44	68	5
86	F	Y	Y		95482	4	2	2009	8.5	19	68	1
87	F	Y	Y	Blue	95470	3	4		8	17	66	4
88	F	Y	Y	Purple	95482	2	3	2008	7.5	27	65	0
89	F	N	Y	Turquoise	95482	4	4	2000	8.5	18	68	6
90	F	N	Y	Forest Green	95454	1	3	2003	6.5	20	62	5
91	F	Y	N	Turquoise	95428	3	4	2012	9	39	63	2
92	F	Y	Y	Purple	95490	2	4		6	18	63	1
93	F	Y	Y	Purple	95470	2	4	1995	9	19	68	2
94	F	Y	Y	Gray	95470	3	4	2014	7	21	66	1
95	F	Y	N	Matcha Green	95490	2	3		8	21	66	1
96	F	N	Y	Blue	95482	2	4		5	19	60	1
97	F	N	Y		95481	2	4	2015	7	19	64	1
98	F	Y	Y	Pink	95453	2	4	2006	7	20	60	0
99	F	Y	Y	Purple	95428	3	3	2007	8	20	64	1
100	F	N	Y	Green	95482	3	3	2005	8	21	62	2
101	F	Y	Y	Yellow	95482	4	4	2018	7.5	18	66	10
102	F	N	N	Pickled Beet	95482	4	3	2002	7.5	26	66	0
103	F	N	Y	Blue	95482	3	3	2017	8.5	18	67	9
104	F	Y	Y	Emerald Green	95490	3	3	2007	8	26	72	1
105	F	Y	Y	Periwinkle Blue	95490	2	3		8.5	18	66	5
106	F	Y	Y	Pink	95490	3	4	2000	9	20	68	2
107	F	Y	Y		95490	3	4	2006	9	22	68	3
108	F	Y	Y	Purple	95428	2	3	2004	8.5	26	63	1
109	F	N	N	Blood Orange	95490	3	4	2000	9.5	50	66	6
110	F	N	N	Blue	95482	4	2	2014	8	54	66	2
111	F	N	N	Teal	95441	3	3	2014	8	30	63	1
112	F	N	Y		95482	2	4	2010	7	18	62	2
113	F	Y	Y	Purple	95482	3	4	2015	8.5	20	63	6
114	F	N	Y	Burgundy	95482	3	5	2012	7	19	64	2
115	F	Y	Y	Red	95482	3	4	2011	9	18	65	0
116	F	Y	Y	Red	95482	2	4	2009	7	18	64	1
117	F	Y	Y	Blue	95482	4	4	2013	8	21	66	3
118	F	Y	Y	Purple	95482	4	4		5	19	60	0
119	F	Y	Y	Pink	95453	2	5	2002	8	20	65	3
120	F	N	Y	Blue	95453	4	5		8	19	68	2

Index