Sampling Distribution and 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 tools we have been using are valid only for normal distributions.

There are various sophisticated techniques for making this determination, one of them is called "Normal Quartile Plot", which is included in GeoGebra for you to explore.

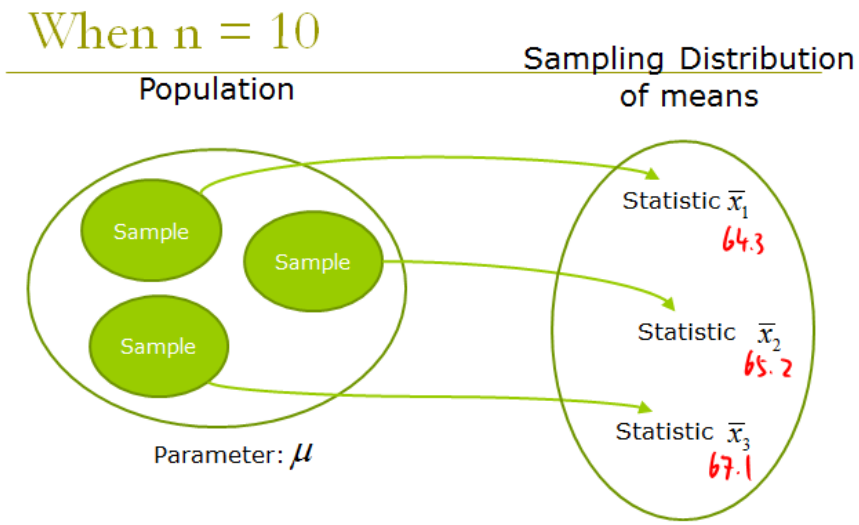
But more importantly, there's a way in which **every** distribution can be turned into a normal one, allowing us to find probabilities and cut-offs, and this way is part of what we call the **Central Limit Theorem**, a result from advanced calculus (don't worry, though), which we will use throughout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

Sampling Distribution of Sample Means

One cannot discuss the Central Limit Theorem without the concept of a **sampling distribution**, which explains why inferential statistics is not just a blind guess. 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 (65 inches) and a certain standard deviation (3 inches). We called the random variable for height *X*. Instead of saying $\mu = 65$ inches, we could write more precisely by attaching the name of the random variable as a subscript, and $\mu_X = 65$ inches, $\sigma_X = 3$ inches. As we will see in a minute, using subscripts helps us clarify which distribution we are talking about at the moment.

Now imagine that we form groups of ten women, many, many such samples, the members of which are randomly selected (say, using random sampling) from the population of 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, or to be more exact, **sampling distribution of sample means**. I know this sounds redundant, but it is necessary to use all fo these words to fully convey what we are doing. A picture may help you see it better:



I've included some sample means just to emphasize that they are not all equal to the original population mean $\mu_X = 65$ inches. Although only 3 samples are shown, the sampling distribution actually contains infinitely many means, since the original population is infinite.

Here is what you need to create a sampling distribution:

1. Pick a sample size and a statistic (say the mean)
2. Randomly draw a sample from the population with the same size
3. Calculate the statistic from the sample and record it
4. Repeat from Step #2

Central Limit Theorem for Sample Means

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 sample means**, a kind of a second-level mean (a "second-level supervisor" if you found the analogy helpful). How do we name this mean mathematically? Using our naming convention, we shall call this $\mu_{\bar{x}}$, which reads exactly as "mean of sample mean".

Now we are talking about three kinds of means: there is the original population mean μ_X, which is a fixed number; the sample mean $\bar{x}$, which varies with each sample, and this new "mean of sample means", what we call $\mu_{\bar{x}}$. The first and third mean are both parameters, while the second one is a statistic.

It shouldn't come as a surprise that μ_X and $\mu_{\bar{x}}$ should be the same. 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 $\bar{x}$'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, $\mu_{\bar{x}} = \mu_X$. This is the first piece of the Central Limit Theorem.

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 our parameters, $P(X > 72) = 0.00982$, 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 nearly none, unless you are watching a basketball team.

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. When you have n approaches infinity, then you are almost drawing the entire population at once, and the sample mean will always land at the population mean, leading to a standard deviation of zero.

But if the standard deviation of women's height is 3 inches, the standard deviation of the sample mean of 10 women is:

$$\sigma_{\bar{x}} = \frac{3}{\sqrt{10}} = 0.95$$

In other words, $\bar{x} \sim N(65, 0.95)$ And the probability that the average height of 10 women is over six feet is:

$$P(\bar{x} > 70) \approx 0$$

The probability is so small that GeoGebra simply displays zero, since we are talking about 5 standard deviation away from the mean here.

Some books give the standard deviation of the sampling distribution, $\sigma_{\bar{x}}$, a special name. It's called the **standard error of the mean**, because of its use in statistics. But when you first start, you should probably just stick with **standard deviation of the sample mean**.

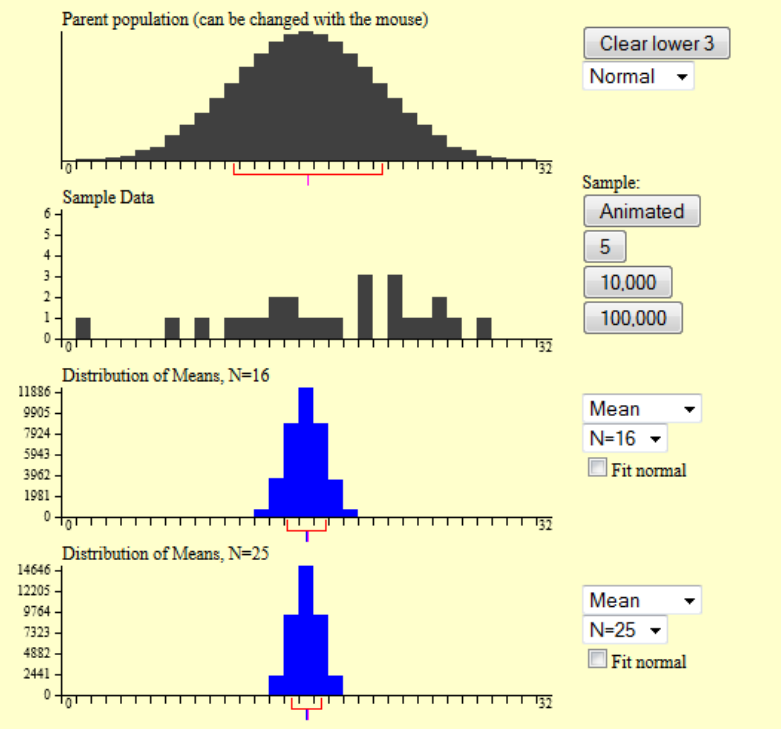
So a sampling distribution, while not changing the mean of the parent distribution, tightens it up and draws it together, and the larger the sample size the greater this effect. But that's not all it does. Remember how I said that every distribution could in some sense become a normal one? That's the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

Demonstration of Central Limit Theorem

Since sampling distribution and Central Limit Theorem are probably two of the most abstract topics in the text, it helps to be able to visualize them with the help of some technology.

Open the applet from the following link: http://onlinestatbook.com/stat_sim/sampling_dist/. After reading the instructions, click "Begin" on the left to launch the applet.



1. What happens when you click "Animated"?
   (A random sample is drawn from the parent population, and the sample mean is computed)

2. How are the 2nd and 3rd figures related to the first figure in the applet?
   (2nd figure is the sample, 3rd figure is the sampling distribution of sample means)

3. Click the buttons named “5 samples”, and “10,000 samples”. What happened each time?
   (More and more sample means are generated by repeating the random sampling process)

4. Keep drawing more samples until the statistics of sample means stop changing. Record the mean and standard deviation of Sample Means (shown in the 3rd table on the right side) below in the appropriate row.

Sample size	Mean of sample means	Std dev of sample means
n=1 (population)		
n=5		
n=16		
n=25		

   (Note: the standard deviation of sample means should decrease as n increases)

5. Reset the applet by click "Clear lower 3". Repeat Steps 1 – 4 until you have completed the table with your experimental results. What is the effect of changing the sample size when you examine these statistics?
   (as the sample size increases, the mean of sample means approaches the original population mean, while the standard deviation decreases according to $1/\sqrt{n}$)

The web applet also allows you to change the parent distribution from normal to something else (e.g. uniform), and you can still see the Central Limit Theorem at work.

So here are the three pieces of the Central Limit Theorem for sample means:

1. The mean of the sample means is the same as population mean, i.e. $\mu_{\bar{x}} = \mu_X$
2. The standard deviation of the sample means decreases as the sample size increases, i.e.
   $\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$
3. The distribution of the sample means approaches a normal distribution, under certain conditions, i.e. $\bar{x} \sim N(\mu_X, \sigma_X / \sqrt{n})$

Assumptions for the Application of CLT

The first part of the Central Limit Theorem regarding the normality of the sample means is not an obvious fact. Specifically, either one of these two conditions must be met before we can treat the sampling distributions of sample means ($\overline{X}$) as a normal distribution). The first condition states that:

- The original (parent) distribution (X) itself is normal.

But this may be too much to ask, since as we saw in the class data: many of the quantitative variables do not even look symmetrical (such as AGE0, let alone normal. What makes Central Limit Theorem special is the second condition:

- The sample mean is taken over a somewhat large sample size (typically n > 30 is used).

As seen in the simulation in the last section, when the distribution of X is not normal to begin with (say uniform or skewed), the sampling distribution of means does not look normal for small n. But as n increases, the familiar bell-shaped curve start to emerge. Although this is another place where we wish we had calculus to be able to back up our claims, computer-based simulations provide an excellent window into how abstract mathematical truths such as CLT are able to predict the outcomes of random experiments.

Problem Solving with the CLT

Have you noticed the sign posted inside the elevator? It usually says the capacity in terms of pounds, and the maximum number of passengers. Here is a picture from a Japanese elevator:

Although you probably don't read Japanese, you can probably guess that 1150 kg (2535 lbs) is the weight limit, and the elevator can fit at most 12 people. Suppose the population of people who use this elevator has normally distributed weights with a mean of 180 lb and a standard deviation of 40 lb. If this elevator is used by the same population of men, how often does the elevator exceed capacity when it's full (n=12)?

Suppose 12 men enter the elevator and their total weight is exactly 2535 lbs. Then their average weight is 211 lbs. If we look at the probability of one person exceeding 211 lbs, it is:

$$P(X > 211) = 0.219$$

About 1 in 5 times, which will be quite frustrating if you have to take it every day. But this is not what we are looking for. Since in a group of 12, the heavier people are going to be balanced by the light-weight people, we should be looking at the probability that the average weight exceeds 211 lbs, i.e. $P(\bar{x} > 211)$.

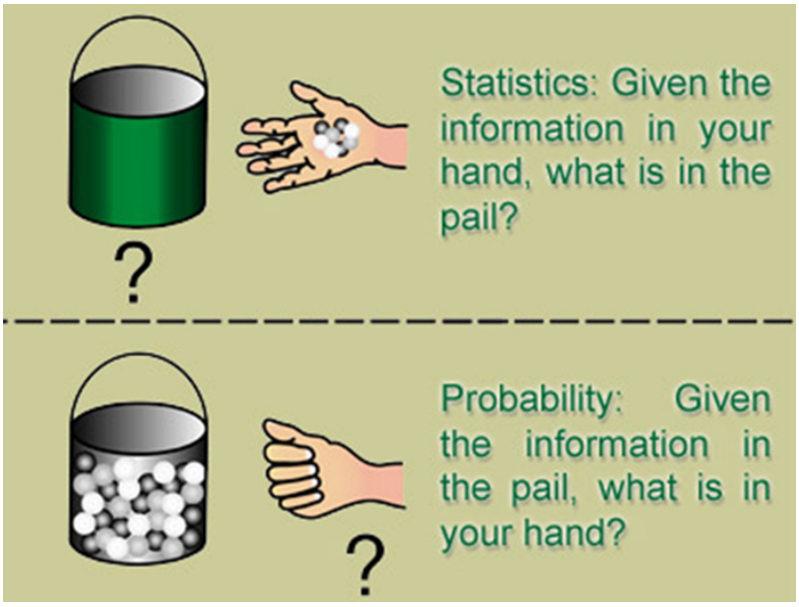
Since CLT tells us that $\bar{x} \sim N(180, 40/\sqrt{12})$, we arrive at something much more reasonable:

$$P(\bar{x} > 211) = 0.004$$

Sampling Distribution and CLT of Sample Proportions

(This section is not included in the book, but I suggest that you read it in order to better understand the following chapter. You can skip it for now, and revisit after you have done the reading for Chapter 8.)

Sampling distribution and Central Limit Theorem not only apply to the means, but to other statistics as well. Since we are about to start inferential statistics in Chapter 8, let's revisit a picture that we saw a few weeks ago, when we first started the chapter on probability:

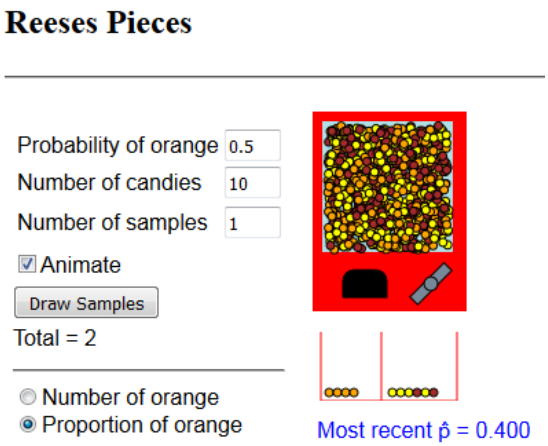


Instead of having marbles and pail, let's replace them with something more interesting -- Reese's pieces, which I'm sure most people have had a taste of. For those of you who haven't had Reese's pieces lately, let me remind you that there are three colors in Reese's pieces, orange, yellow, and brown (the colors that made the brand's logo).

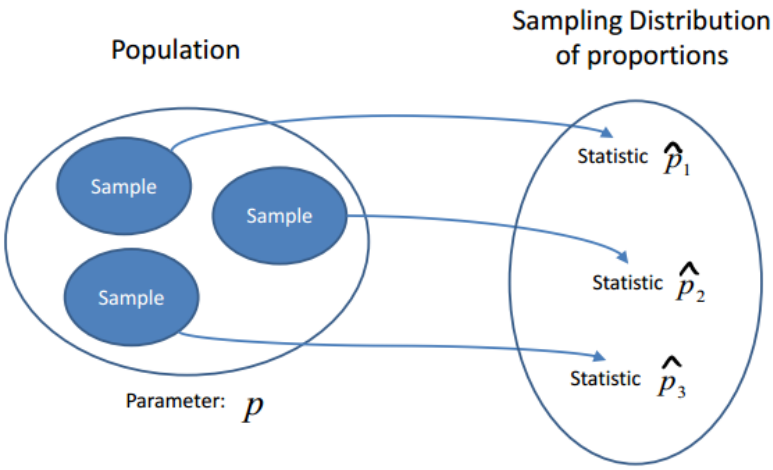
Imagine we have a huge inventory of Reese's pieces. We start by having each group draw 10 Reese's pieces, which represents a random sample with sample size $n = 10$. Based on their respective sample, if someone needs to give an estimate of the real proportion of orange pieces among all Reese's pices, the best guess one can give is simply based on counting the number of orange pieces, then divided by the size of the sample. In Chapter 8, we call this a "point estimate" for the population proportion (which is unknown):

$$\hat{p} = \frac{x}{n}$$

If you want some visual demonstration, this applet can be quite helpful: http://www.rossmanchance.com/applets/OneProp/OneProp.htm?candy=1



For example, if you counted 6 orange pieces in a batch of 10, then $\hat{p} = 6/10 = 0.6$. Because of the randomness in choosing the samples, the $\hat{p}$ value will vary depending on the sample. If we look at $\hat{p}$ as a random variable, and consider all the possible values of $\hat{p}$, these values (ranging from $0, 0.1, \cdots, 0.9, 1.0$) form a "**sampling distribution of sample proportions**". This may look like a lot of words for a single symbol, but I haven't found a way to use fewer words to convey the same idea. Perhaps the following picture, similar to the one we used for sample means, will help you understand sampling distribution a little further:



The blue bubbles on the left represent individual samples (they might overlap since it was taken with replacement). Each sample produces a statistic $\hat{p}$. So the sampling distribution is a distribution of "statistic-s", which is another way to think about this concept.

Because in our example, there are so few pieces in the sample (a total of 10), it is actually possible to calculate the probability of each $\hat{p}$ using the binomial probability from Chapter 5, if we know the population proportion p. Recall that binomial probability requires we known the total sample size n and the probability of success p (in this case, the proportion of orange pieces). Assuming that we are taking 10 pieces at a time, and there are exactly 50% orange pieces ($p = 0.5$), by using the binomial calculator in GeoGebra, we can calculate the probability distribution as follows:

$\hat{p}$ Probability

0 P(x=0) = 0.001

0.1 P(x=1) = 0.01

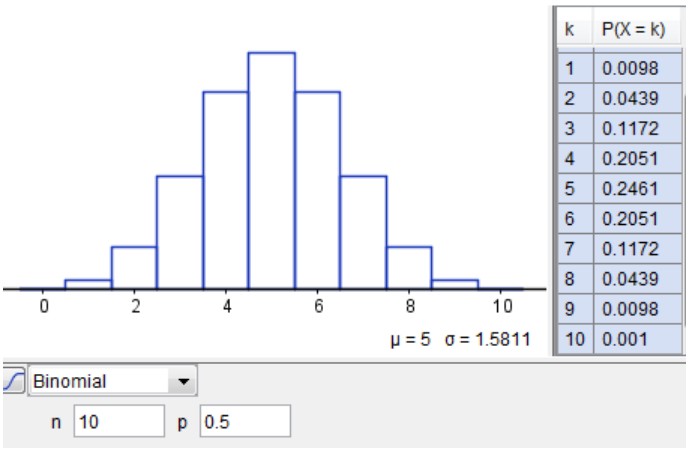
0.2 P(x=2) = 0.04

.....

0.9 P(x=9) = 0.01

1.0 P(x=10) = 0.001

You can verify these probabilities by using the binomial calculator. The calculation for $\hat{p} = 0$ (which corresponds to $x = 0$) is shown in the following screen shot:



k	P(X = k)
1	0.0098
2	0.0439
3	0.1172
4	0.2051
5	0.2461
6	0.2051
7	0.1172
8	0.0439
9	0.0098
10	0.001

Notice the shape of the binomial distribution is very similar to that of the normal distribution, except that binomial is discrete, and normal is continuous. In fact, there is a version of the Central Limit Theorem (not included in the book) that addresses exactly this issue:

- **Central Limit Theorem for Sample Proportions**:
    1. The sampling distribution for the sample proportion $\hat{p}$ is approximately normal
    2. The mean of $\hat{p}$ is equal to p, i.e. $\mu_{\hat{p}} = p$
    3. The standard deviation of $\hat{p}$ is equal to $\sqrt{\frac{pq}{n}}$, i.e. $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

To see why this is true, recall that in a previous lecture, we showed that x follows a binomial distribution with mean $n \cdot p$, and standard deviation $\sqrt{npq}$. Since $\hat{p} = \frac{x}{n}$, divide the mean and standard deviation each by n, we have the mean of $\hat{p}$:

$\mu_{\hat{p}} = np / n = p$,

and the standard deviation of $\hat{p}$:

$$\sigma_{\hat{p}} = \sqrt{npq} / n = \sqrt{\frac{pq}{n}}$$

One useful example for thinking about the standard deviation of $\hat{p} = \sqrt{\frac{pq}{n}}$ (which we will represent using $\sigma_{\hat{p}}$) is by varying the sample size n: if you take a small hand of orange pieces (say 4), then compared to $n = 10$ pieces, it's much more likely you will get some extreme values for $\hat{p}$, such as all orange ($\hat{p} = 1$), or no orange ($\hat{p} = 1$). Although the mean $\mu_{\hat{p}} = 0.5$ stays the same (50% orange), the

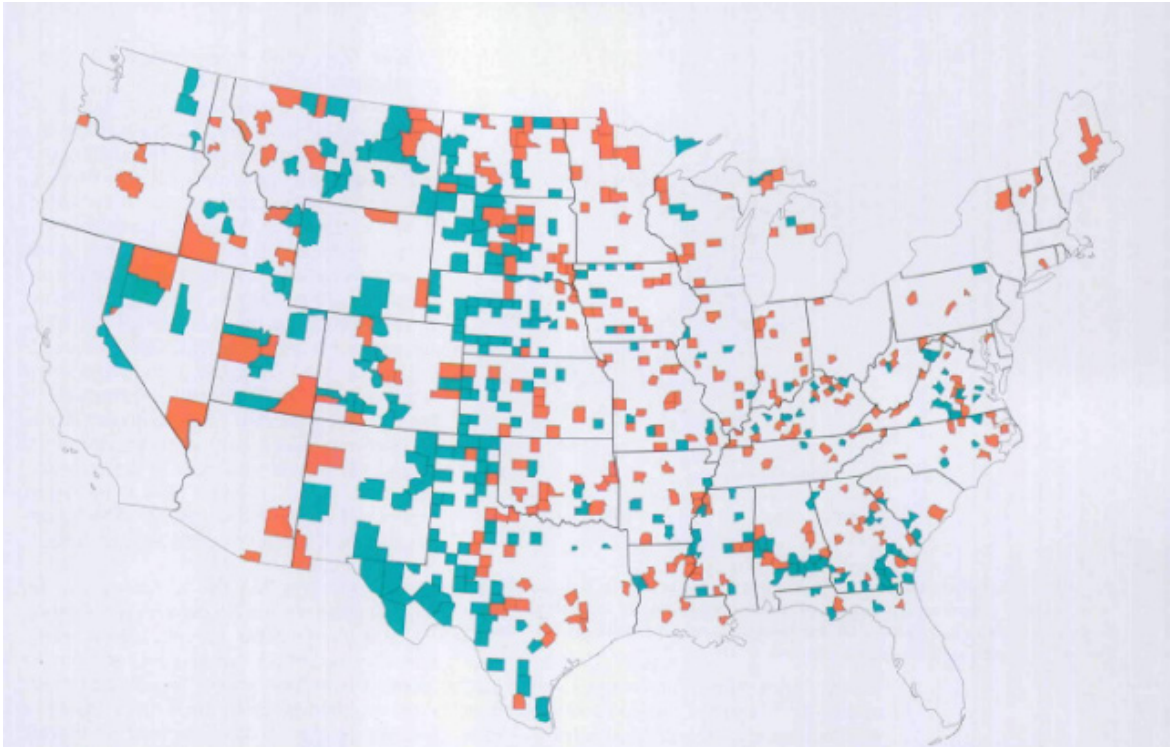
standard deviation is larger for $n = 4$, since there are fewer pieces in the sample. As we learned in algebra, the value of $\sqrt{\frac{pq}{n}}$ decreases as n increases.

Central Limit Theorem and the Small-Sample Illusion

The Central Limit Theorem has some fairly profound implications that may contradict our everyday intuition. For example, if I tell you that if you look at the rate of kidney cancer in different counties across the U.S., many of them are located in rural areas (which is true based on the public health data). What is going through your mind?

(Think about this question before you read the next paragraph.)

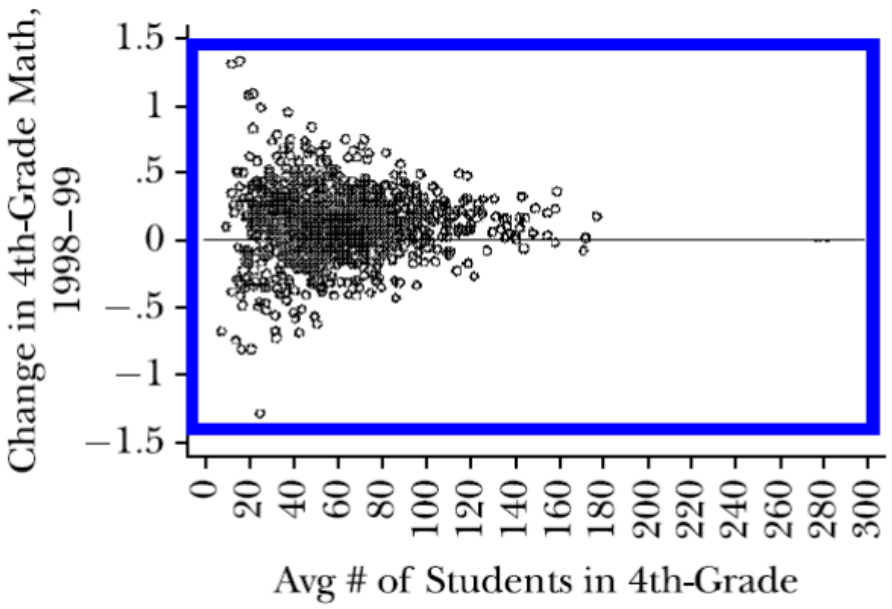
Within the space between these two paragraphs, you probably thought of a dozen possible explanations, many of them making perfect sense: rural areas have fresher air, rural people exercise more, healthier diets, well water without chlorine, etc. Some of you may be thinking that perhaps this is another example of correlation, but not causation. But what if I give you another piece of information: if you look at the counties with the highest rates of kidney cancer, a large proportion of them are also rural counties. In other words, when you look at the data, not only is there no causation, there isn't any correlation either. In fact, many of the counties with the lowest (in green) and highest (in orange) rates of kidney cancer are adjacent to each other. But they have one thing in common: location in a rural area, which translates to small populations.



Assuming that nothing really strange is going on in Rural America, what could be going on here?

It turns out that the rate of kidney cancer is just a massive example of the Reese's pieces. Instead of counting the proportion of orange pieces, we are counting the proportion of kidney cancer. What distinguishes rural counties from their urban counterparts is their size: rural counties have fewer people living in them compared to the cities. So the rural county is like the example of $n = 4$, while the urban county is like $n = 10$. As we saw above, the standard deviation is much larger for the smaller sample, therefore explaining the extreme values that you obtain from the rural counties.

Failure to understand the Central Limit Theorem can lead to some costly mistakes. One is example of the "small school" movement that was backed by several foundations (notably, Gates Foundation and Annenberg Foundation). The idea originated in the observation that among the nation's best performing schools, many of them are small schools with few students. Several million dollars then went into breaking larger public schools into smaller, "specialized" schools that hopefully boost the success rate of students. Unfortunately, these efforts did not produce the expected results. It turns out that the difference between small and large schools is yet another example of the effect of changing the sample size. In the following graph, the change in the average 4^{th} grade math assessment score is plotted again the size of the school. A positive change thus indicates an improvement; while a negative change indicates perhaps a school is performing “poorly”:



If you look at the schools with the worst performance (a negative change in 4-th grade Math scores), you will find many of them are small schools of 60 or fewer students. In fact, there are about as many "failing" schools as the "successful" schools. What we have learned from CLT told us that perhaps "failing" and "successful" are just both illusory labels that we put on the data – they are the reflection of the fact that sample means from small samples tend to have more variation.

The small-sample illusion illustrates one of the innate limitations of human cognition: we are not naturally inclined to think statistically about the information we receive. This is one of the good reasons why it's good for everyone to take this course!

Team Homework #6: Sampling Distributions and Central Limit Theorems

Part 1: Constructing a sampling distribution from the class height data.

1. Open the class data and copy the height column to GeoGebra (do not include the header row).
2. Randomly select a pair of values using a random number generator. You can use the random number generator provided in Canvas, or your graphing calculator: MATH → PRB → randInt(1, n). For example, if the random number is 8 and 34, you will average the values in the 8th row and 34th row in your sample, and record the value below:

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

3. Based on the means generated above, calculate the following statistics from GeoGebra:

   Mean = _____ *Std Dev* = _____

Collecting Averages of Groups of Five. Repeat steps above with one exception. Instead of recording the height of 30 pairs, record the average height of 30 groups of five. A GeoGebra applet is provided in Canvas to help you automate this task. You can draw a new sample by clicking the button on the right.

1. Randomly select 30 **groups of five**, and record the values of their **sample mean**. You can press the F9 key to refresh the random sample as well as the mean.

   _____ _____ _____ _____ _____

   _____ _____ _____ _____ _____

   _____ _____ _____ _____ _____

   _____ _____ _____ _____ _____

   _____ _____ _____ _____ _____

   _____ _____ _____ _____ _____

2. Based on the means generated above, calculate the following statistics from GeoGebra:

   Mean = _____ *Std Dev* = _____

Discussion:

1. How do the mean and standard deviation of the distribution of the sample mean change, as sample size changed?

2. How do your results compare with predictions of the Central Limit Theorem (show your calculations)?

Part 2: Sampling Distribution of Proportions

1. Open the "Reese's Pieces web applet" link in Canvas. What happens when you click "Draw Samples"?

2. Now change Number of Samples to 100. What happens when you click "Draw Samples" a few times?

3. Draw more samples until the total exceeds 3000 samples. Now check the "Summary Stats" box, and change "Number of orange" to "Proportion of orange". What do you see in the bottom graph?

4. Repeat your experiment by changing the "Probability of Orange" to 0.3, and "Number of Candies" to 100. (Your number of samples should still be fairly large.) Report your findings in Step #3.

5. Compare your results in #3-#4 with predictions of the Central Limit Theorem for proportions (show your calculations).

Part 3: Use the Central Limit Theorem to answer the following questions. Identify the sampling distribution in each question.

1. A bottling company uses a filling machine to fill plastic bottles with cola. The contents in the bottles vary according to a normal distribution with mean 300 ml and standard deviation 3 ml.

    a. What is the probability that an individual bottle contains between 298 and 302 ml?

    b. What is the probability that the average content of a 6-pack of soda is between 298 and 302 ml?

2. The total cholesterol levels for healthy adult males in the United States without a cardiovascular condition follows a normal distribution, with mean 201 mg/dl, and a standard deviation of 46 mg/dl.

    a) Find the 95-th percentile for the average cholesterol level of 25 randomly selected healthy adult males from the U.S.

    b) 10 men from Jackie's company took a physical exam and their average cholesterol level turns out to be 230. What is the probability that the average cholesterol level of 10 randomly selected healthy males from the U.S. population is higher than theirs? What does this mean for her co-workers?

3. Assume the passengers who use a particular elevator have normally distributed weights with a mean of 150 lb and a standard deviation of 35 lb. The elevator has a weight limit of 1600 lb and maximum carrying capacity of 10 passengers. What is the probability that the elevator exceeds its weight limit when it is filled to the capacity? (Hint: consider the average weight when the elevator has exceeded the capacity)

4. The wait time at the popular Russian River Brewing company follows a uniform distribution between 0 and 45 minutes. If 30 customers are randomly surveyed, what is the probability their average wait time is less than half an hour?

5. Suppose 60% students bring their textbooks to class on any given day. What is the probability that in a randomly selected group of 80 students, less than 30% bring their textbook on a certain day? Use the Central Limit Theorem for proportions to answer the question.

Introduction to Hypothesis Testing

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 the Hypotheses

The first, and perhaps the most importance piece of hypothesis testing, is stating hypotheses and labeling the claim.

Let me describe three different situations that someone might be interested in studying:

- The life expectancy of professional athletes
- How many miles a brand of tires last
- The percentage of Americans who think our president is doing a good job

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 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. As for the presidential job approval rating, sometimes you see headlines like "president's job approval falls below 50% for the first time since 2009". So it's obviously a bit deal for people who write the news.

I'm going to pick a position for each of these situations and make a claim:

- Athlete: Professional athletes don't have the same life expectancy as other people.
- Tire: This brand of tires is worse than average.
- President: the majority of American think our president is doing a good job.

Now each claim has to be translated into a mathematical sentence. The sentences will begin with a parameter: p if it's about a proportion, and μ if it's about a mean. It will be followed by one of three symbols, just like the verbs in a sentence: >, <, and $\neq$. A common error is writing the hypothesis without the parameter, so if you notice yourself writing down hypotheses like ">5", then the subject of the sentence is missing.

Sometimes it's tricky to see which of the three symbols the claim implies. But once you have chosen the correct subject (parameter), the verb (>, <, or $\neq$) and the object (some value the parameter is compared to), then you'll have the sentence!

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 "majority" simply means anything over 50%. Then we will have the following claims, written as mathematical sentences:

- Athlete: $\mu \neq 79$ years
- Tire: $\mu < 30000$ miles
- President: $p > 0.50$

Notice that in the last case, we've deliberately used the proportions in the claim, instead of using percentages, to be consistent with the confidence interval for proportions.

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_a, or sometimes H_1 in other books).

To differentiate the null from the alternative hypothesis is a relatively simple matter: the statement containing the idea of 'equal' becomes the null hypothesis, H_0; while the remaining statement, the one containing either < (less than), > (greater than) or $\neq$ (different from), becomes the alternative hypothesis, H_a.

Returning to our three situations, let's consider what the H_0 and H_a should be in each case:

- Athlete: $H_0 : \mu = 79$, $H_a : \mu \neq 79$
- Tire: $H_0 : \mu \geq 30000$, $H_a : \mu < 30000$
- President: $H_0 : p \leq 0.50$, $H_a : p > 0.50$

The alternative hypothesis gives each test one of three names, as follows: If H_a contains the symbol "<", the test is left-tailed; 'Less than' means 'to the left of.' If H_a contains the symbol ">", the test is right-tailed. 'Greater than' means 'to the right of.' If H_a 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 ">". The following imply $\neq$: 'has changed from,' 'is not the same as,' 'is different from,' and 'deviate from.'

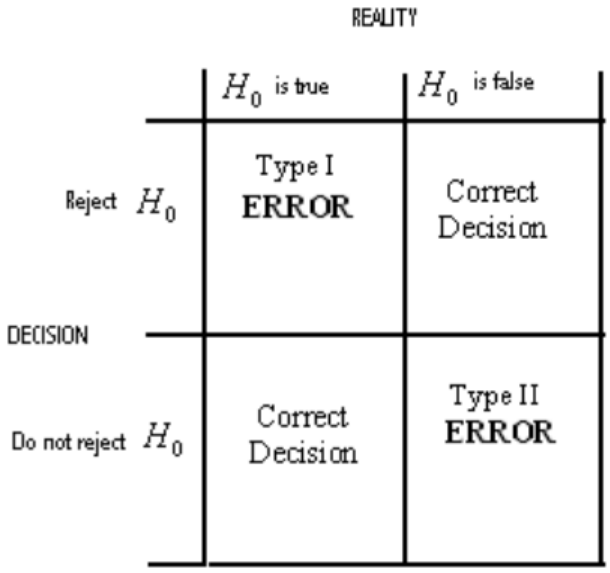
Translating a claim properly and situating it and what's left over from it correctly into H_0 / H_a structure is essential to performing the test of the claim correctly, which is why I've placed so much emphasis on it. The next step will be to look at the statistic (either proportion or mean) of a random sample from the population, and compare it with the value of the parameter that is included in the null hypothesis to see whether this could happen by chance. The likelihood that a statistic occurred given that H_0 is true is called **P-value**. We will spend a great deal of time on p-value in the next lesson, but we shall just focus on the decision regarding the hypotheses just stated.

Before leaving this section on hypotheses, in case you have looked at other texts on statistics (or the problems in the homework), you may also see that the alternative hypothesis is sometimes represented by H_1 instead of H_a. In addition, some books only use "=" in the null hypothesis, instead of the three different scenarios ($\geq$, $\leq$, and =) we listed above. This is mostly done out of convenience, since as we shall see, it is the equal sign that allows us to use the null hypothesis to compute the P-value.

Decision and Error in Hypothesis Testing

For now, we'll investigate what happens when we make a decision. The decision is about the null hypothesis. It's always about the null hypothesis. It's not about the claim, which is the statement that contains <, >, or $\neq$.

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 criterion for making 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 that also appeared in the book:



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 more than half of the students who take the hybrid Math 15 are female. Because we're talking about a proportion, the claim translates as $p > 0.50$; The claim, lacking the equals component, is thus H_a. The statement that the equal component is thus H_0. Thus stating the hypotheses and identifying the claims, we have the initial setup:

- $H_0 : p \leq 0.50$, $H_a : p > 0.50$

Written in words, this becomes

- H_0: Half of the students in hybrid Math 15 are female.
- H_a: More than half of the students in hybrid Math 15 are female (Original Claim).

A Type I error will mean deciding that over half students are female, when in fact the true proportion of female students is 50%.

A Type II error will mean deciding that there is no gender disparity, when the true proportion of female students is over 50%.

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.

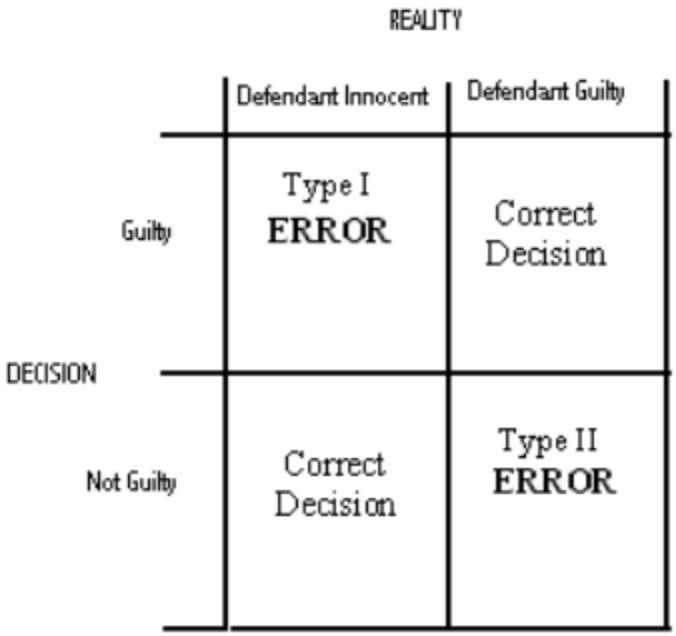
The probability that the data could have occurred, if the null hypothesis were true, is what we called the P-value (we'll go into P-value in the next section in more detail), and we compare it to a quantity called α (the same old α, 1 minus the confidence level, that we encountered in creating confidence intervals), which in this context is called the level of significance, 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. There is a rather intricate relationship between the two errors: but there is a crucial difference from a practical point of view: once the data has been collected, the likelihood of Type I error is entirely up to us! Just choose a small enough α will ensure the Type I error is small. While to keep Type II error low, you will need more data. You can take entire course on these two types of errors, and some of the exercises in our team homework are designed to give you some visual examples of what each one represents.

What if we already obtained our data, and are eager to just test our claim against it? 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, 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.

Analogy: "Guilty" and "Not Guilty"

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 rather 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.'

Is our legal system confusing? For someone who just came from another country (especially one without a criminal justice system), it probably is. Is our legal system perfect? Probably not. We have Type I errors (innocent people locked away), and Type II errors (criminals who get away), and it's hard to see how you are able to eliminate both errors, if you have only limited data. But we are all glad that we have such a system, don't we?

If you would like to read a little more about the analogy of criminal justice, you can also visit this website for some helpful discussions.

A Historical Note on Hypothesis Testing

There is another surprising connection between hypothesis testing and our criminal justice system -- they were both British in origin! (remember the term "null" in H_0?) The modern theory of hypothesis testing was invented by a British statistician named Karl Pearson, and to this day, the top journals in statistics are still published in England (such as Biometrika and Journal of the Royal Statistical Society).

So here is my (completely unsubstantiated) theory of how the convoluted language in hypothesis testing was invented: as a student, Karl Pearson studied Roman Law. So naturally, he saw the parallel between statistical decisions and criminal justice. Moreover, we know that the British people love to make sentences with multiple negatives! One need to look no further than the popular British TV Series: Downton Abbey . Coincidently, the story line spanned the years from 1912 to 1921, which was the time Karl Pearson established Hypothesis Testing (I would like to believe this one was statistically significant:)

I don't know if it was a popular style of speech at the time, but the dialogues in Downton Abbey frequently use double-negatives. For example, when Lady Mary is asked how she thinks of an attractive suitor, her response is often "I don't dislike him." Apparently, it is not lady-like to say you like someone! If she puts on a statistician's hat, she might as well say "I reject the null hypothesis (that this guy is boring)!"

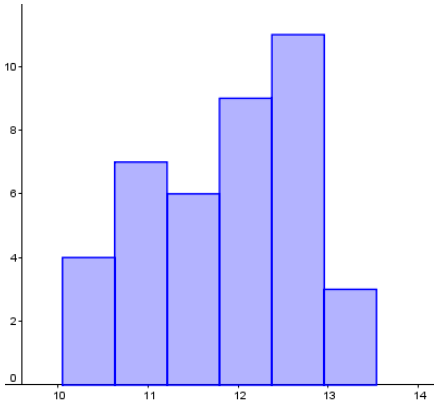
My favorite character in Downton Abbey is the grandma, the Dowager Countess of Grantham. She even uses triple negatives! Take an example from Season 2, when she was asked how she thought of Sir Richard Carlisle, a potential suitor of her granddaughter: "It is not the case that I don't dislike him. I just don't like him." So here goes our aristocratic rendition of "failure to reject the null hypothesis!"

Hypothesis Testing of a Mean

In the last section, we introduced the idea of translating your question into two statements – the null hypothesis and the alternative hypothesis. We also discussed the distinction between the reality and the outcome of your decision: there are two possible types of errors that we named "Type I" and "Type II". To continue with the rest of the hypothesis test, we'll examine how the procedure can be applied to testing claims about the mean.

A Simple Example: a Left-Tailed Test and P-value

Let's start with a hypothetical scenario: suppose we have a normal distribution that describes how many ounces of coffee that I'm supposed to get out of a coffee vending machine when I press the button for "12 oz". In particular, we assume that the machine is a bit old, and this normal distribution has mean of 12 oz and standard deviation of 1 oz. Suppose we've run an experiment to sample 40 values from the normal distribution X ~ N(12, 1), obtained from measuring the amount of coffee dispensed by the machine each time we press "12 oz". Here the histogram of these 40 values:



The mean of these 40 values is $\overline{x} = 11.8$. If we trust the standard deviation σ is indeed 1 oz, does it mean the old coffee machine is defective? Your intuition probably tells you that 11.8 oz is "close enough" to the advertised 12 oz for this to be something that just happens by chance. But how close is "close enough"? Hypothesis test will help us answer these questions.

Let's start with a quick review of the Central Limit Theorem:

1. The mean of the sample means is the same as population mean, i.e. $\mu_{\overline{x}} = \mu_X$
2. The standard deviation of the sample means decreases as the sample size increases, i.e.

$$\sigma_{\overline{x}} = \frac{\sigma_X}{\sqrt{n}}$$

3. The distribution of the sample means approaches a normal distribution, under certain conditions, i.e. $\bar{x} \sim N(\mu_X, \sigma_X / \sqrt{n})$

So if our old vending machine is indeed dispending coffee according to the normal distribution N(12, 1), then we'll be expecting the following:

$$\bar{x} \sim N(12, \frac{1}{\sqrt{40}}) = N(12, 0.16)$$

Now a key question is "how likely is this scenario ($\bar{x} = 11.8$)?" On one hand, if it happens somewhat frequently (say 1 out of 5 times), then there is probably nothing wrong with the vending machine. But if it is only supposed to happen 1 out of 100 times, then something is probably not right – the vending machine is probably no longer dispending coffee at the average of 12 oz. To associate the $\bar{x} = 11.8$ with a probability, we need to identify two things:

1. We need to set up the null and alternative hypotheses. Imagine we are the customers who were worried that we were shortchanged on coffee ("the machine is giving us than 12oz!"), then a sensible pair of hypotheses will be:
   $H_0 : \mu \geq 12, \ H_a : \mu < 12$

   It's also possible to use slightly different notations for this left-tailed test, such as what's used in the free online homework system:

   $H_0 : \mu = 12, \ H_1 : \mu < 12$

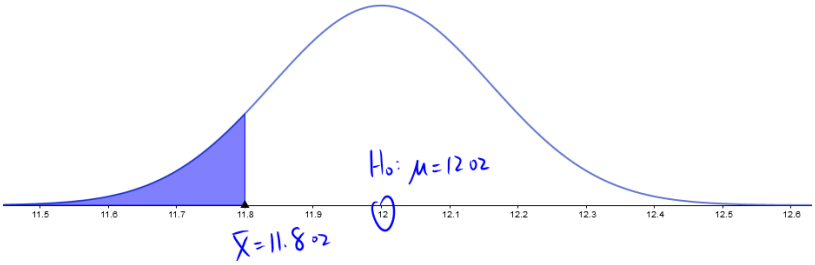
   The key is to associate the null hypothesis with the equal sign, since this is what provides an mean for the sampling distribution, as required for the CLT. So our question has been re-phrased as:

   "how likely is this scenario ($\bar{x} = 11.8$), if the null hypothesis was true ($\mu = 12$)?"

2. Although it's tempting to assign a probability to $P(\bar{x} = 11.8)$, recall that this will result in zero, since any continuous distribution will have zero probability when you are looking at the area under just one point. Instead, a better way to quantify whether $\bar{x} = 11.8$ is "close enough" is to look at the area that's further to the left, since our alternative hypothesis believes that the mean was moving to the left of 12 oz. In other words, to answer "how likely is this scenario ($\bar{x} = 11.8$), if the null hypothesis was true", we use the following conditional probability:

   $P(\bar{x} < 11.8 \mid \mu = 12)$

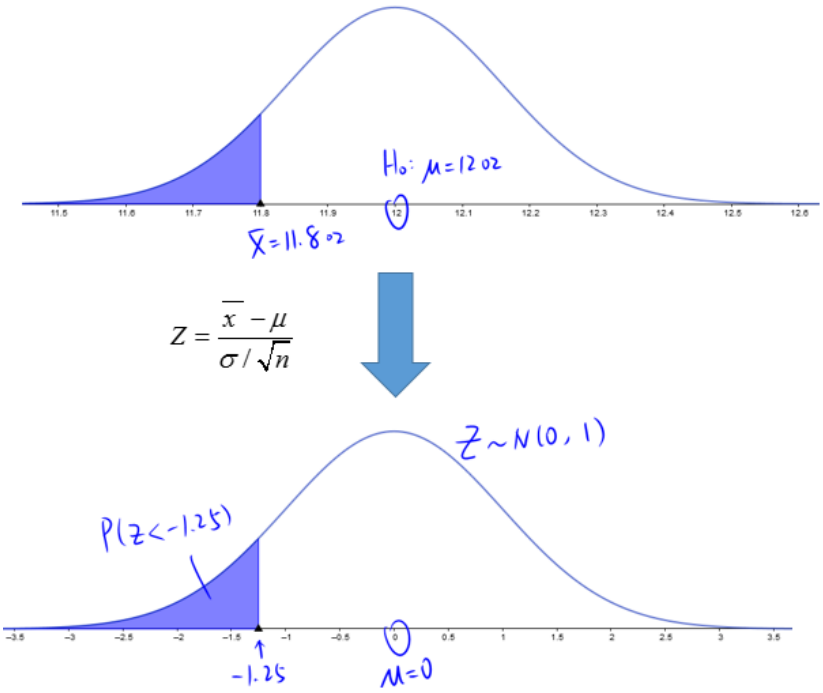
Now this probability is no longer zero! This left-tailed probability will be referred to as P-value.



To compute the P-value, remember we'll need to convert this normal distribution $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$ to the standard normal distribution Z via the formula introduced in the section on normal distributions: $Z = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{n}}$. In the context of hypothesis testing, we'll call this particular statistic (since it's still derived from the data) a **test statistic** that depends on both the data and the null hypothesis. In our example, the test statistic shows:

$$Z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{11.8 - 12}{1/\sqrt{40}} = -1.25$$

It's helpful to place the original P-value associated with sampling distribution of $\bar{x}$ side-by-side with the distribution of the test statistic, since they remind us of two ways to look at the same P-value: while the first graph is most helpful in helping us understand what conclusion to draw, the second one shows us how to compute the P-value.



According to the graph of the test statistic, the P-value is equal to P(Z < -1.25)=0.106, or approximately 1 out of 9 times you are collecting a sample of 40 from the vending machine. What do we make of this number? There is a school of thought among statisticians that regards this as a subjective matter: if you are super cautious, and are always looking for things that are broken, this may indicate to you that the vending machine is indeed broken. (Keep in mind that you may be making a Type I error though, since the machine may be working just fine.) On the other hand, if you think 1 out of 9 times is a bit too frequent to be taken seriously as a sign of trouble, then you can just choose to ignore it.

But the most of people who use statistics do not like such ambiguity, since we as humans prefer clear-cut answers: is the machine broken, or is it not? To go from P-value to this binary decision, we'll need a "significance level" (or the probability of a Type I Error) mentioned in the introduction, represented by the symbol α. For example, if we consider 1 out of 20 experiments an acceptable level of occurrence for Type I Error , then we'll stick to the significance level $\alpha = 0.05$. Going back to the example, we noticed that:

P-value > α (since 0.106 > 0.05)

In other words, 1 out of 9 times is not your daily occurrence, but it occurs often enough so that we cannot rule out the possibility that vending machine is performing correctly (with mean = 12 oz). The decision can be phrased as:

We fail to reject the null hypothesis.

In other words, the intuition we had earlier was confirmed: $\bar{x} = 11.8$ was indeed too close to 12 in order to be consider a red flag. So at the end of the hypothesis test, we say:

There is insufficient evidence to show that the vending machine is dispensing an average of less than 12 oz.

Example of a Right-Tailed Test: Perceived Age

Remember the example of estimating the average perceived age? By using a random sample of 31 guesses, we were able to derive two confidence intervals: one using the assumption that we know the population standard deviation σ, one using just the same (hence the sample standard deviation s). Of course, I did not tell you my actual age at the time of the experiment: I just turned 34 when the experiment was conducted. So the natural question is: "on average, did I look older than my age at the time of the experiment?"

Here is the catch about statistical inference that many people do not find intuitive: we can never be completely certain about our answers. Instead of a resounding “Yes” or “No”, what we will provide is a more nuanced answer: based on pre-determined criteria of statistical significance, we will say there is “enough” evidence to support my claim, or “not enough” evidence to support my claim that I look older than my actual age.

Since I have been doing this activity for many years in my stats class, I have noticed that although the mean of people's guesses keeps going up year after year (I wonder why this is the case :), the standard deviation is remarkably stable (approximately 2.6 years). If our class is no exception (unless many of you possess supernatural powers or have seen my passport), I can safely assume that the population standard deviation for my perceived age is known to be 2.6 years. Take, for example, one of my classes last semester provided these responses:

35 33 32 46 33 32 38 38 35 32 32 35 32 37 35 42 32 33 38 38 33 35
34 39 38 35 36 32 34 34 30

(I do not know who said I looked like I was 46, but I assume this person was not joking, since that will constitute a sampling error)

The sample mean of these 31 guesses was 35.1 years. Recall that in the chapter on Central Limit Theorem, if we know the mean μ and the standard deviation σ of the population, then the sample mean $\bar{x}$ will follow a normal distribution, with mean given by:

$$\mu_{\bar{x}} = \mu$$

(In English: the mean of the sample means is the same as the population mean.)

Note we previously used μ_X instead of μ, but here we try to stick with μ to emphasize it is the parameter we are trying to estimate.

And the standard deviation of the sample means (also known as the **standard error** of $\bar{x}$) is given by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(In English: the standard deviation of the sample means decreases inversely proportionally with the square root of sample size)

What now follows us parallel to the previous lesson: the CLT for sample means leads to the following result that is similar to what we saw with sample proportions: if we convert $\bar{x}$ to a standard normal distribution, we will obtain:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Let's start by writing down the hypotheses. Since the claim is about the mean (average perceived age), this will involve the parameter μ :

$$H_0 : \mu \leq 34$$

$$H_a : \mu > 34$$

Notice that we have translated the original question to the alternative hypothesis. Although the sample mean 35.1 is a bit higher than my actual age 34, the question is whether this is an accident due to the normal variation in the data, or a significant result due to the fact that I do look older (at least in the classroom). To quantify what counts as a significant result, first we'll have to pick a level of significance: we'll use $\alpha = 0.05$, which corresponds to 1 Type I error in 20 similar experiments.

The next step in hypothesis test is translating the null hypothesis and the data to a test statistic. We will discuss two scenarios: the standard normal Z, if we assume that all of the guesses were from the population with the same standard deviation with $\sigma = 2.6$ years. When we let go of this assumption, and let data speak for itself in terms of the standard deviation, the test statistic will be Student's t. So let's discuss each scenario.

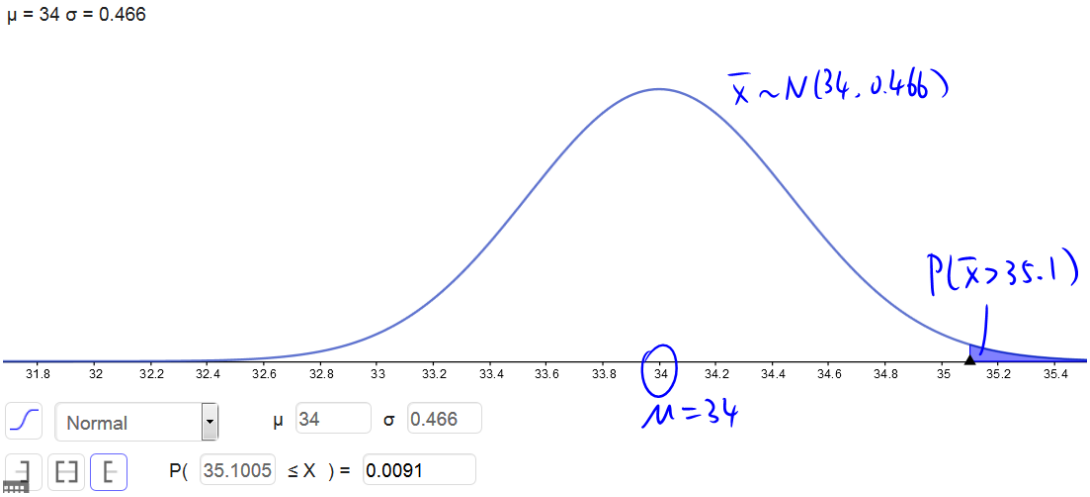
Testing A Claim about the Mean with Given Population Standard Deviation

Scenario 1: using the standard normal Z as the test statistic, assuming that the population standard deviation σ is known to us. (In our earlier example on confidence intervals, we had used $\sigma = 2.6$ years).

As we mentioned previously, if the null hypothesis were true ($\mu = 34$), by the Central Limit Theorem, the sample mean $\bar{x}$ is supposed to follow a normal distribution $N(\mu, \sigma/\sqrt{n}) = N(34, 0.47)$. Hence we can measure the "extreme-ness" of the sample mean (i.e. the P-value) by looking at the right-tailed probability (since our alternative hypothesis is right-tailed):

P-value = $P(\bar{x} > 35.1)$

We'll illustrate this P-value with the following graph:



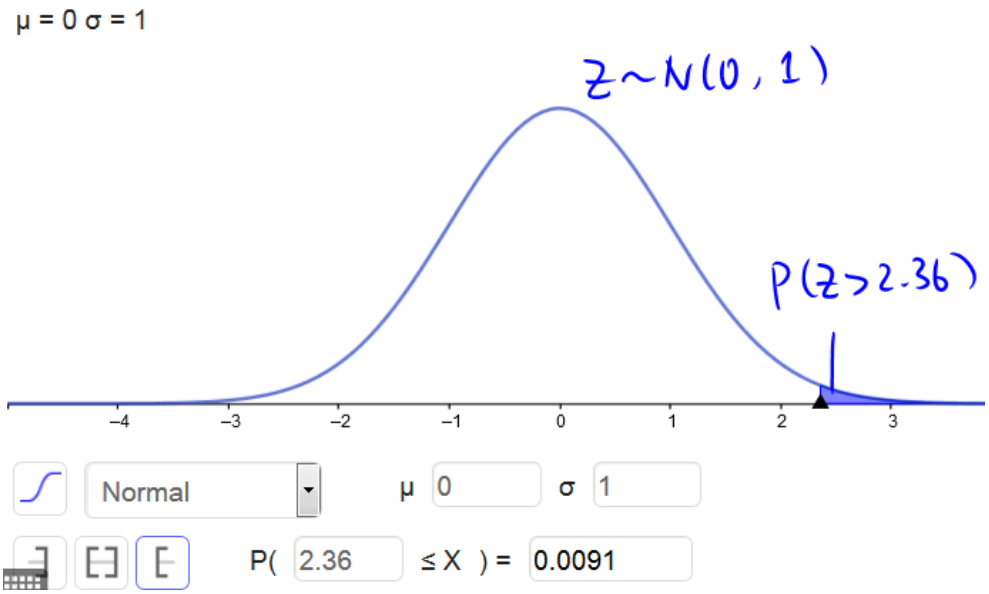
To evaluate this probability, we could either use the normal calculator with 34 as the mean, and $2.6/\sqrt{31}$ as the standard deviation of the sampling distribution. Alternatively, we could also convert the normal distribution to Z, and compute the P-value based on Z:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{35.1 - 34}{2.6/\sqrt{31}} = 2.36$$

The P-value will then be the area to the right of the test statistic:

P-value= $P(\bar{x} > 35.1) = P(Z > 2.36) = 0.009$

For the sake of comparison, we'll also illustrate the P-value in the graph showing the test statistic, which has a different mean and standard deviation, but showing the same tail as the graph for $\bar{x}$ above.



These two pictures are both important in terms of helping understand what P-value represents. The second graph (showing Z) is what allows us compute the P-value by using tables related to standard normal distributions (which is the way all normal probabilities are calculated, regardless which software/calculator you used).

But it is the first graph showing the distribution of $\bar{x}$ that reveals the meaning of the P-value:

- *"If my average perceived age were 34 years, then assuming that the population standard deviation is 3.0 years, there is a 0.9% chance that the average guess of my age by a randomly selected group of 31 students is over 35.1 years."*

Why is this statement so long? Recall that the general format of the P-value is a conditional probability:

P(data more extreme than yours **GIVEN** null hypothesis is true)

In this case, the null hypothesis states that $\mu = 34$ years, which provides a mean for the sample means $\bar{x}$. So here is P-value interpretation translated symbolically:

$$P(\bar{x} > 35.1 \,|\, \mu = 34) = 0.009$$

The language referring to *"a randomly selected group of 31 students"* is simply to ensure the same sampling distribution will be used – if the sample size has changed from 31, then the entire sampling distribution will change with it too, something we don't want to see.

Needless to say, this probability is quite small (less than $\alpha = 0.05$): P=0.009 means if I am convinced that I do look like 34 years to all my students and do this experiment over and over again with groups of 31 students, I am only going to get sample means over 35.1 about 1 in 110 times. That's very unlikely! So our decision is pretty clear: the null hypothesis is quite implausible in face of the evidence, and we should reject the H_0: in fact, I do look a little older than my actual age. Let's state the conclusion a bit more officially:

- *Based on the significant level of 0.05, there is enough evidence to show that on average, I look significantly older than my actual age (34).*

Testing A Claim about the Mean by Only Using the Data

Introducing the Student's t Distribution

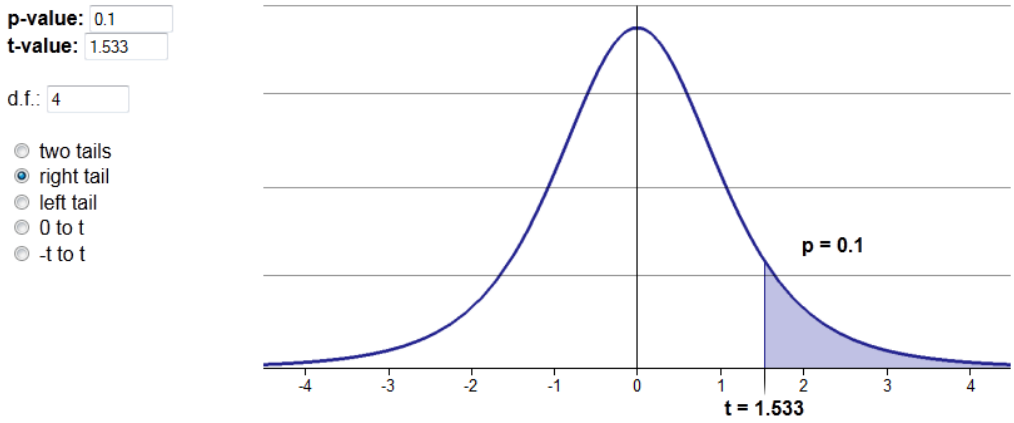
What we did with testing the mean looks quite reasonable, right? What could go wrong here? Now imagine that I wanted to do this experiment on Halloween, and came to class dressed in a costume. Now this may have some consequences with regard to the standard deviation, since I am not wearing my usual semi-professional outfit anymore, and people's guess can differ wildly. If we look at our original assumption that the population standard deviation is given, this may not be true anymore, and Central Limit Theorem does not apply, since σ is hidden from us.

What can we do instead? It turns out, this was a problem solved by a guy who worked at the Guinness Brewing company that makes the famous beer from barley. What he discovered that was if we replace σ in the formula $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ by the standard deviation calculated from the sample, the result is a random variable that follows a different distribution, named Student's t-distribution ("student" was the pseudonym used by the author to keep the trade secret for the Guinness Company):

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

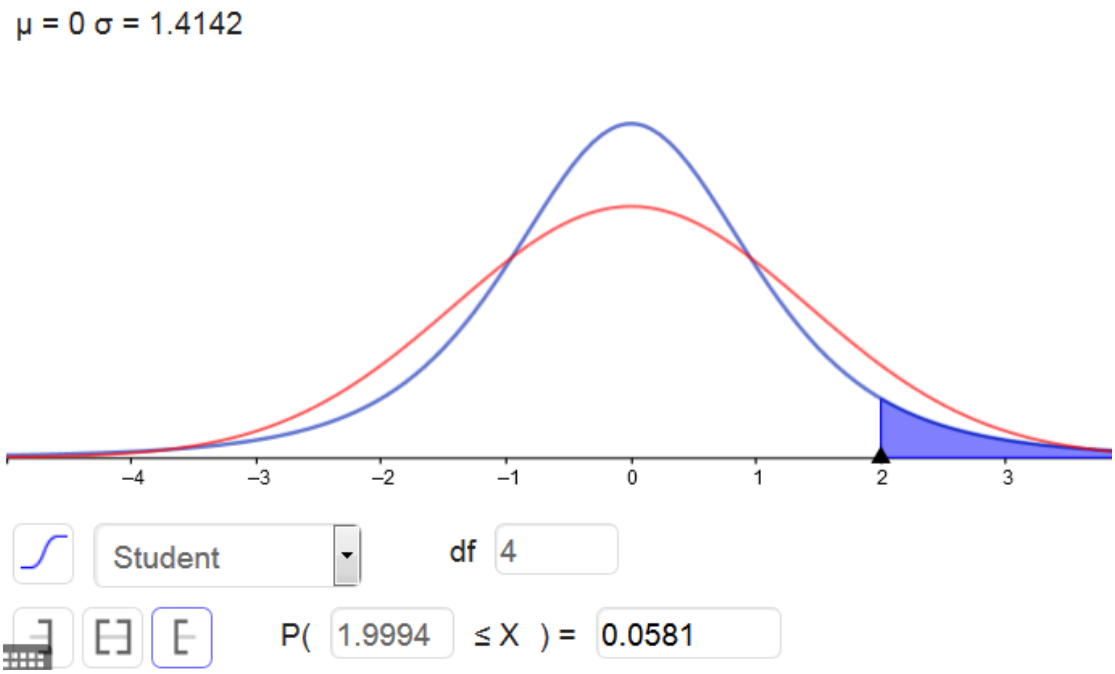
(notation review: although both were referred to as "standard deviations", s refers to the statistic, while σ refers to the parameter)

A picture of the t-distribution is shown here, together with the right-tail probability associated with $P(t > 1.533)$



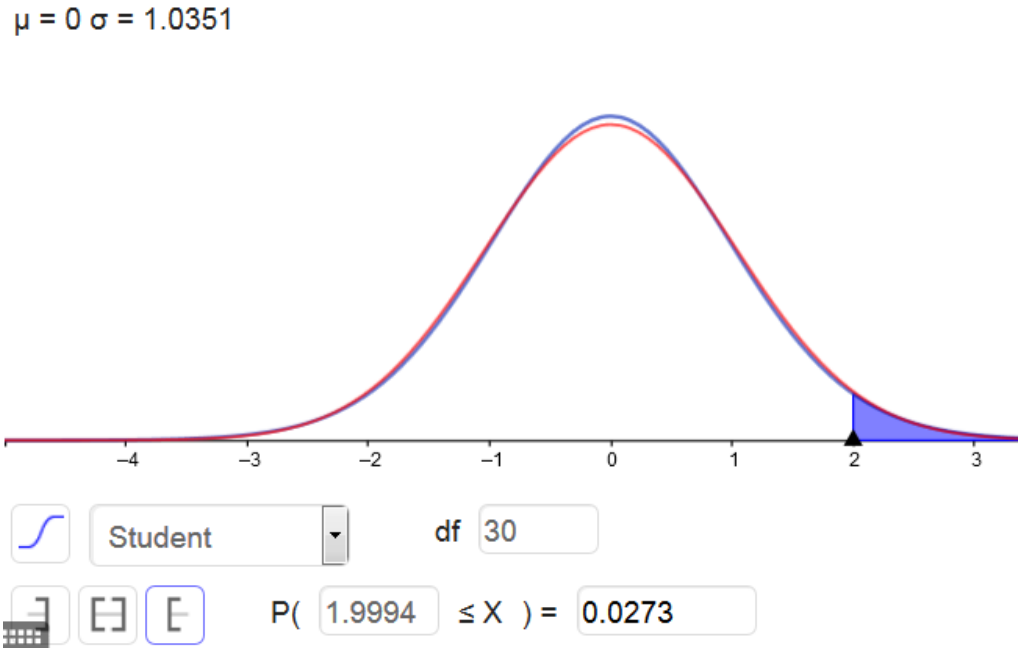
Although the graph of the t-distribution does look like its poor cousin, the standard normal (Z), the main thing that distinguishes the t-distribution from the normal is the so-called "degree of freedom", which is related to the sample size through $df = n - 1$ (remember that the sample standard deviation from Chapter 3: $s = \sqrt{\frac{\sum (x-\mu)^2}{n-1}}$)?

It is useful to compare the t-distribution and the normal distribution side-by-side.



The red curve is the standard Normal Z ~ N(0, 1). The blue curve is showing the t distribution with degree of freedom (df) equal to 4. The t-distribution has a standard deviation of 1.414, which is quite bit a larger than the standard normal ($\sigma = 1$). You can tell this by examining the tails of the two distributions: the t-distribution has much larger tails.

Now look at the following pictures, which shows what happens to t when df = 100:



Now the two distributions look practically identical, as you can tell from the standard deviation as well.

Going back to our original problem of testing my original hypothesis (whether I look older than 34), the process looks very similar to Scenario #1, except that we will need to substitute Z with the new Student's t statistic.

Scenario 2: assuming that the population standard deviation σ is unknown to us, use the Student's t to conduct the test.

Recall that if $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$, the t distribution is obtained from the sample mean and standard deviation as follows:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

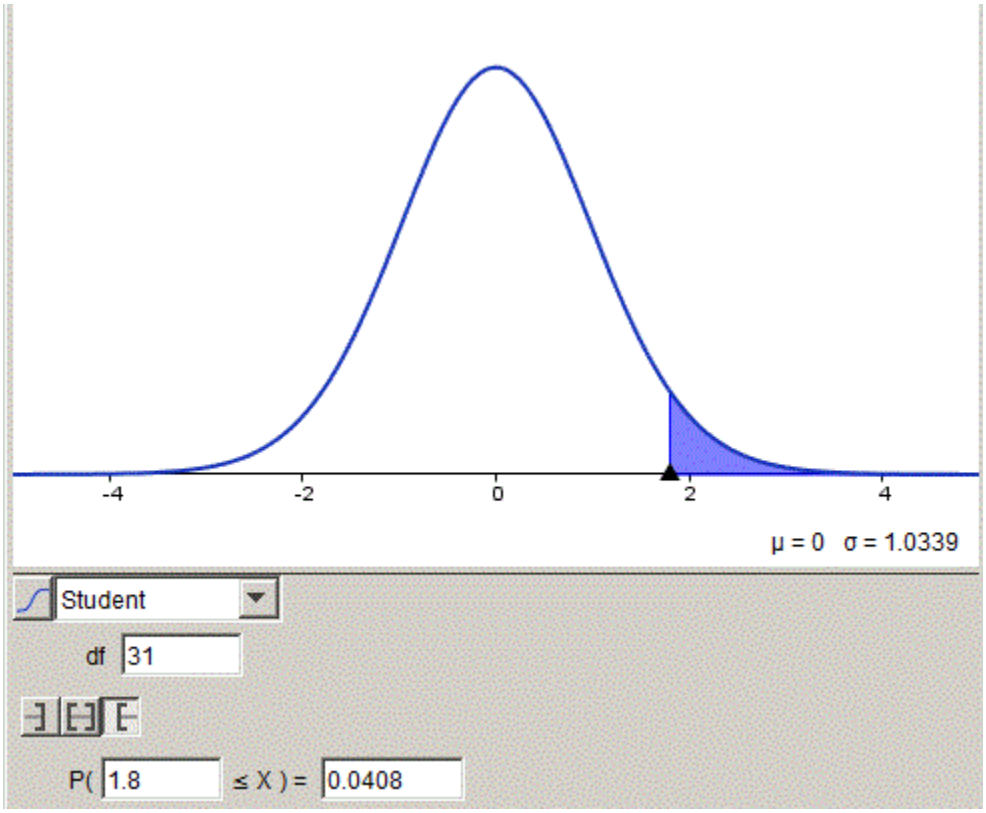
If σ is unknown (which is most likely the case for testing claims about the mean), then Student's t is the preferred test statistic for the hypothesis test. In our example earlier, using $s = 3.4$ years (you can find this from the original data by using your calculator or GeoGebra), and the $\mu = 34$ according to the same H_0, we can calculate the t statistic as follows:

$$t = \frac{35.1 - 34}{3.4/\sqrt{31}} = 1.8$$

The degree of freedom for the t statistic is $n - 1 = 30$. Hence the P-value (again being right-tailed) will be:

P-value= $P(t > 1.8) = 0.04$

This P-value was found through the t-distribution calculator in GeoGebra:



Although our decision did not change (the significance level is still 0.05), our P-value is a lot closer to what's required for a significant result. In fact, if we had chosen 0.01 to be our significance level, our entire conclusion will change. Again, it's useful to look at the interpretation of the P-value according to what we just did:

- *"If my average perceived age were 34 years, then assuming the population standard deviation is unknown, there is a 4% chance that the average guess of my age over 31 students is over 35.1 years."*

Importance of P-value in Hypothesis Testing

Comparing the two different approaches discussed above, you will notice the simplicity of the P-value approach: once we have identified a suitable test statistic, how much the data deviate from the null hypothesis will be quantified by the P-value: the smaller the P-value, the larger the discrepancy. Hence with a suitable tool for computing the P-value, anyone who correctly understands this logic will be able to apply the hypothesis test, without necessarily understanding the mathematics behind the sampling distribution of the test statistic. We shall see that the interpretation of P-value allows us to apply hypothesis testing to a much wider scope of problems beyond testing claims about a single proportion or mean.

Because of its important, it's worth repeating the general meaning of P-value as follows:

- General Interpretation of p-value: "*the chance of seeing the evidence more extreme than the data (as per the appropriate sampling distribution), given that* H_0 *is true.*"

When you are asked to state the "interpretation of P-value", please use the proper context (i.e. the null hypothesis and what "more extreme" means) in your statement.

Summary: Hypothesis Tests about One Mean

It's also helpful to summarize our two different approaches to testing claims about one mean. When you are trying to show your work in conducting a hypothesis test in the exam, be sure to include these components.

- **Testing a mean with a given population standard deviation (based on α=0.05)**

a) State the null and alternative hypotheses.
$H_0 : \mu \le 34 \qquad H_a : \mu > 34$

b) Show the test statistic.
$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{35.1 - 34}{2.6 / \sqrt{31}} = 2.36$$

c) Show the P-value
$$P(\bar{x} > 35.1 \mid \mu = 34) = P(Z > 2.36) = 0.009$$

d) State the decision of your hypothesis test
Reject the null hypothesis (since P-value < α)

e) State your conclusions in words a non-statistician would understand.
There is significance evidence to show that my average perceived age is more than 34.

- **Testing a mean without population standard deviation (using data only, α=0.05)**

a) State the null and alternative hypotheses.

$H_0 : \mu \leq 34 \qquad H_a : \mu > 34$

b) Show the test statistic.

$t = \dfrac{35.1 - 34}{3.4 / \sqrt{31}} = 1.8$ (degree of freedom = n-1 = 29)

c) Show the P-value

$P(\bar{x} > 35.1 \mid \mu = 34) = P(t > 1.80) = 0.04$

d) State the decision of your hypothesis test

Reject the null hypothesis (since P-value < α)

e) State your conclusions in words a non-statistician would understand.

There is significance evidence to show that my average perceived age is more than 34.

Example of a Two-Tailed Test

As a final example, we'll also look at a two-tailed test that produced a null result.

- *Nationally, patients who go to the emergency room wait an average of 6 hours to be admitted into the hospital. Do patients at rural hospitals have a different average waiting time? The 15 randomly selected patients who went to the emergency room at rural hospitals waited an average of 5 hours to be admitted into the hospital. The standard deviation for these 15 patients was 2.3 hours. What can be concluded at the the α = 0.01 level of significance level of significance?*

The key piece of information in this problem is "*The standard deviation for these 15 patients was 2.3 hours*", which indicates that we ought to use a t-test in this case, and the 2.3 hours should be regarded as the sample standard deviation: s = 2.3. In addition, the 6 hours is what appears in the null hypothesis, and the question "*Do patients at rural hospitals have a different average waiting time*" translates to a two-tailed test, where the null hypothesis states that the rural patients have the same average waiting time as national average (6 hours).

The main difference between two-tailed and one-tailed test are what is accepted as valid evidence against the null hypothesis. For one-tailed test of means, any sample mean smaller than the population mean in H_0 will be consider relevant evidence: the further the sample mean away from μ, the small the P-value, thus the stronger the evidence. In comparison, the two-tailed test admits both very large and

very small sample means as evidence against H_0, and the tails were supposed to be chosen before any experiment was ever conducted. In our example, although the sample mean (5 hours) was smaller than the population mean (6 hours), we are interested in the chance of having both very short waiting time (less than 5 hours) and very long waiting time (more than 7 hours), since they are supposed to be symmetrical across μ = 6.

As a result, the two-tailed test requires that we are looking at the left side of the test statistic (since the sample mean 5 hours is less than 6 hours, the population mean according to the null hypothesis), and multiply by 2 to account for the same area in the right tail.

a) State the null and alternative hypotheses.

$H_0 : \mu = 6 \qquad H_a : \mu \neq 6$

b) Show the test statistic.

$t = \dfrac{\bar{x} - \mu}{s/\sqrt{d}} = \dfrac{5-6}{2.3/\sqrt{15}} = -1.684$ (degree of freedom = n-1 = 14)

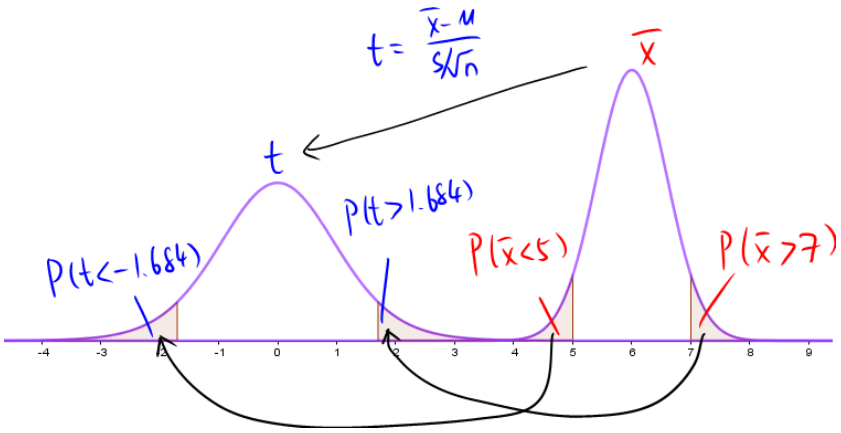
c) Show the P-value (also see the graph below)

$P(\bar{x} < 5 \text{ or } \bar{x} > 7 \mid \mu = 6)$

$= P(t < -1.684 \text{ or } t > 1.684)$

$= 2 \cdot 0.057$

$= 0.114$



d) State the decision of your hypothesis test

Fail to reject the null hypothesis (since P-value > α)

e) State your conclusions in words a non-statistician would understand.

There is insufficient evidence to show that the average waiting time for rural patients to be different from 6 hours.

Hypothesis Testing for A Proportion

Intuition: Comparing Hypotheses with Data

In the last section, we discussed the underlying logic of hypothesis testing. Let us now turn to a more concrete example of conducting hypothesis test on a specific type of parameter: population proportions. Recall in the previous chapter, we also started with the question of estimating proportions. So we will follow the same order, first looking at testing claims about proportions, then testing claims about means. This may be a bit different from the textbook, but you can choose to follow either order.

From the class data, I have found something that sets the hybrid Math 15 classes from the same traditional sections that I taught in the past: there seem to be more women in hybrid sections than men. In the class data, we see that there are 45 responses from women, and 29 responses from men. Using the symbols introduced earlier, we will represent our qualitative data as follows:

$$\hat{p} = \frac{x}{n} = \frac{45}{45+29} = 0.608$$

Does this mean that hybrid sections are for some reason more "popular" among female students? If the answer is yes, then it will be very valuable for the college to have this information, since it may help us schedule hybrid sections in the future (for example, on Fridays when parking is more readily available).

However, we want to be sure that the alleged gender disparity is not just a fluke: if we toss a fair coin 74 times, then it is certainly possible to get 45 heads, although such an outcome is quite unlikely, because you would expect close to half of the tosses turn up heads. So we'd better have a way to say that 45 (sample proportion of 60.8%) is a little too far from 37 (50%) for this to occur by chance, but we will need also be clear what we mean by "occur by chance." When this unlikely event does happen in the 50/50 split scenario, and we mistakenly concluded that more than 50% students are women, a Type I error occurs. Based on our discussion in the last section, we want to control this type of error by choosing a small α. How small? Just like in confidence interval, where a variety of confidence levels are in use, we can choose α to be as small as we like. For now, let's use the common value of $\alpha = 0.05$.

Intuitively, we reasoned that "more than half of the hybrid Math 15 students are female", because "60.8% (the sample proportion) is too high above the 50%." Let us make this argument more official by using hypothesis testing.

As we emphasized last time, the most important step in hypothesis testing is setting up the null and alternative hypothesis. In our case, because our suspicion is that there are more women, so if you use p for the proportion of women, this leads to a right-tailed test (“>”) :

$H_0 : p \leq 0.50$

$H_a : p > 0.50$

In plain English, H_0, always containing the equal sign, says exactly half are women, and H_a is our original claim: more than half are women.

Test Statistic

The second step in hypothesis testing is to convert our data and null hypothesis into a single numerical summary, also called the **test statistic**. Intuitively, the test statistic measures how "far away" our sample proportion is from the alleged 50% population proportion according to H_0. The calculation of test statistic is familiar from the chapter on confidence interval of proportions:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

The various symbols appearing in the test statistic formula are as follows:

1. The sample proportion of female studnets: $\hat{p} = \frac{x}{n} = \frac{45}{45+29} = 0.608$
2. The population proportion of female students according to the null hypothesis: p = 0.50
3. The population proportion of male students according to the null hypothesis: q = 0.50
4. The sample size: n = 45 + 29 = 74

Similar to our discussion in testing a claim about means, this test statistic is also based on **Central Limit Theorem for Sample Proportions**:

1. The sampling distribution for the sample proportion $\hat{p}$ is approximately normal
2. The mean of $\hat{p}$ is equal to p, i.e. $\mu_{\hat{p}} = p$
3. The standard deviation of $\hat{p}$ is equal to $\sqrt{\frac{pq}{n}}$, i.e. $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

If we recognized that p and $\sqrt{\frac{pq}{n}}$ are the mean and standard deviation of the sample proportion $\hat{p}$, then the test statistic formula is another way to convert a non-standard normal distribution to the standard normal, hence the letter “Z”.

Notice in confidence interval (see the previous notes), we used this test statistic to construct the margin of error (we also discussed it as finding the cut-off values from the area in the middle, i.e. confidence level). The main difference is that in Confidence Intervals, we are using $\hat{p}$ to predict a range for p; while in Hypothesis Testing, we are comparing $\hat{p}$ with p to see whether the two are significantly different. The way we interpret the test statistic is exactly the same as how we look at the Z-scores: if the test statistic is more than 2 standard deviations away from zero, then it has gone "too far".

In our scenario, the test statistic comes out to be:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.608 - 0.500}{\sqrt{\frac{0.5 \cdot 0.5}{74}}} = 1.86$$

This is somewhat marginal according to the criteria of 2 standard deviations. So to decide whether this could "occur by chance", we need to do one of two things that help us arrive at the final decision.

Making Decision with P-value

Option 1 is called the "P-value" method (be careful of the multiple "P" used in different contexts), which is the default method used in the book. It is named because we need to put a probability on how likely it is to see data like ours, when H_0 is true. Because the Z (standard normal) distribution is continuous, to find the probability, you must look at an area. So the natural thing to do to characterize "data like ours" is to look at the area to the right of our test statistic, since any test statistic landing in this tails will support our claim (larger than 0.50). Using the normal calculator in GeoGebra, we found that:

$$P(z > 1.86) = 0.031$$

As we mentioned in the introduction to hypothesis testing, this probability gives us an idea of how the evidence measures against H_0:

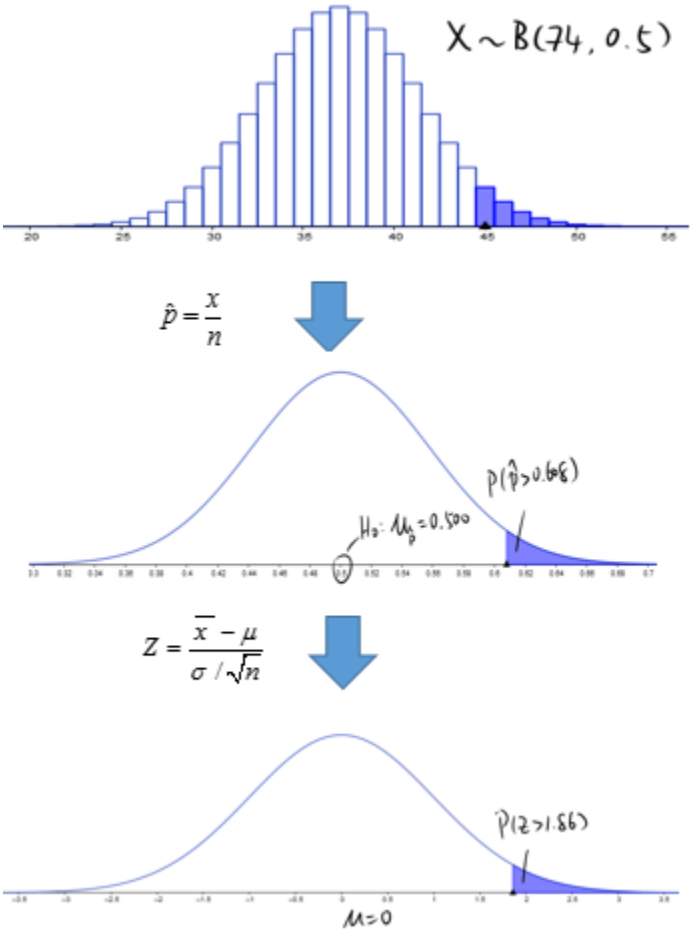
- Interpretation of p-value*: "If 50% of the students are female, then there is a 3.1% chance that in a sample of 74 students, 45 or more of them are female."*

Alternatively, you can also refer to the sample proportion in your P-value interpretation:

- I*"If 50% of the students are female, then there is a 3.1% chance that in a sample of 74 students, 60.8% or more of them are female."*

I know this is a rather long sentence, but trust me, every word in the definition of P-value is necessary to convey the full meaning -- it's a rather complex idea that cannot be expressed with fewer words. If you have trouble reading the above sentence, I would suggest that you take a pause, and read the sentence out loud to yourself.

Here is the graph that illustrates the different views of the P-value:



Comparing the P-value to our significance level α, we basically concluded that getting a sample proportion of 60.8% or more is very unlikely to occur by chance. So following the logic of hypothesis testing, we will decide to reject H_0, i.e. supporting our original suspicion that more than half of the students are female (remember the double-negatives?). In statistical language, we often state that "there is significant evidence to support that the majority of students in hybrid Math 15 are female," and we are done!

Let's pause for a second and take a look at why the fuss about null/alternative hypothesis is necessary. In order to calculate the p-value, we must have some type of ground truth so that we can use to calculate the probability of our evidence. H_0, the null hypothesis, provides a ground truth. In fact, p-value can be roughly stated as the following statement:

- General Interpretation of p-value: "*the chance of seeing the evidence more extreme than the data, given that* H_0 *is true.*"

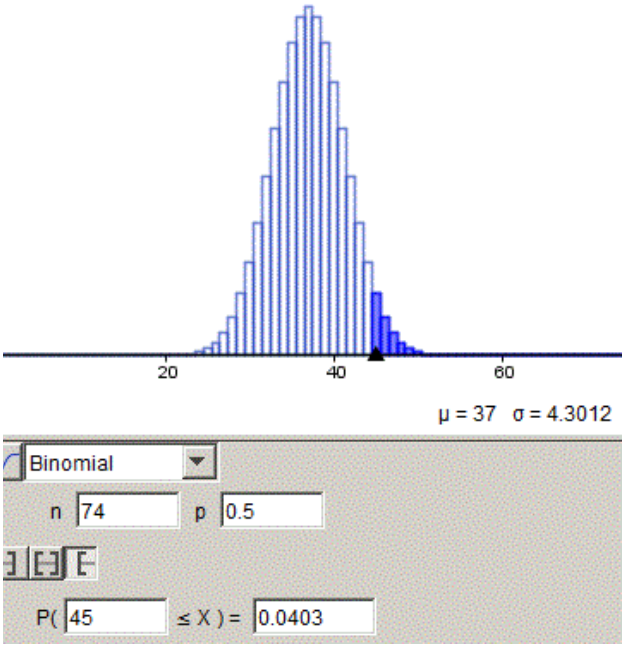
Although our interest is always on H_a, we arrive at our conclusion by negating its opposite H_0, and p-value is a way for us to point to the data and say "that's ridiculous."

P-value is the preferred method in most applications of hypothesis testing because of the simplicity in decision rule: if P-value < α, you have a significant result (either <, >, or $\neq$, depending on H_a). However, historically, it was invented after the more traditional method of using critical values to arrive at the same decision.

Binomial Test: An Alternative to Using the Z Test Statistic

A final note on the p-value method: using the definition of p-value as a conditional probability ("the probability of getting data more extreme than yours, given that the null is true"), there may be multiple methods of computing the p-value. In fact, the method of converting the proportion data to standard normal was used because of limited computational resources, since computing the binomial probabilities with $n = 74$ is a bit tricky in the old days of sliding rules. By using the Binomial calculator that we introduced in the chapter on discrete probability distributions, we can also use a binomial distribution with $n = 74$, and $p = 0.50$ to evaluate the p-value:

$$P(X \geq 45) = 0.040$$



When the sample size n is fairly large, the p-value obtained from using the standard normal Z serves as a fairly good approximation of the p-value obtained from the binomial distribution, and you can use either method in testing one proportion. If a binomial distribution is used to calculate p-value, then the hypothesis test is also called "binomial test" (as opposed to "Z-test for a proportion"), if it is included in different statistics software.

Making Decision with Critical Values (optional)

Option 2: instead of comparing p-value with α (both are areas under the Z curve), we could also directly compare the test statistic with the critical value(s), since they are both Z-scores. In our earlier example, since H_a is right-tailed, we will be looking for the critical value (C.V.) that marks the 0.05 area to the right of the Z-score. In other words, we are looking for:

$$P(Z > C.V.) = 0.05 = \alpha$$

Using our normal calculator to solve for Z-score again, we obtain C.V. = 1.64. We call the area to the right of critical value "critical region". When you look at where the Z-test statistic stands, it is certainly inside the critical region. So we have arrived at the same decision as Option 1: we will reject H_0, the proportion of female students is indeed over 50%.

Comparing the p-value method with the traditional method, you can see that they are two different ways to make the same decision with regard to H_0: the p-value method uses area for comparison, and the traditional method uses the Z-scores. Although one-tailed tests are relatively straightforward, the two-tailed tests can be a little tricky: the p-value should include the area in BOTH the left and right tail, and the traditional method will rely on TWO critical values, one associated with a critical region in each tail.

Hypothesis Testing and Proof by Contradiction

So in a nutshell, here is how we confirmed our initial suspicion that more women than men are taking the hybrid Math 15:

1. Assume $p = 0.50$ according to H_0.
2. Compare our sample proportion $\hat{p}$ with 0.50 and summarize the comparison with the test statistic.
3. Use either the p-value (compared with α) or the test statistic (compared with critical value) to conclude that $\hat{p}$ is simply too far from 0.50 to occur by chance.

4. Revisit the assumption in Step 1: H_0 must be false, hence we have proved our point that $p > 0.50$.

If you have taken a philosophy course that deals with logic, you may have recognized that this type of argument is also known as "proof by contradiction". To give a simple example, imagine that if you want to show I am not a vegetarian, after I have told you that I like steaks. So your reasoning goes something like the following:

1. Assume Dr. Lin is vegetarian.
2. Vegetarians don't eat steaks.
3. But Dr. Lin likes steaks.
4. So there is a contradiction. The assumption that he is vegetarian must be false.

Of course, hypothesis testing does not deal with black-and-white situations like this. The real-world scenarios have many shades of gray. Suppose you are trying to test the same claim "Dr. Lin is not a vegetarian", and I told you that I like sushi. Is there a contradiction if you assume I am vegetarian? Well, apparently there are many sushi restaurants in California that serve vegetarian sushi, so the probability that a vegetarian likes sushi, although fairly small, is certainly not zero! (a probability of zero will be equivalent to a contradiction). Should you reject the null hypothesis that I am vegetarian? This will depend on a number of things: first, you will need some type of mathematical model that predicts the probability of a vegetarian liking sushi, which gives us the p-value; then you will need to compare the p-value with a pre-determined significance level. If the p-value is smaller than α, then this means that it must be a miracle that you just met a sushi-loving vegetarian, so you have effectively found a contradiction and proved that I am not vegetarian. On the other hand, if the p-value turns out be quite large, then you will "fail to reject null hypothesis." In simple terms, you just don't have enough evidence to prove I am not vegetarian, which is quite different from proving I AM vegetarian! (remember the "guilty" and "not guilty"?)

Equivalent Statements in Stating the Conclusion

So here are the two conclusions again:

Fail to reject H0 $\Leftrightarrow$ not enough evidence to support Ha $\Leftrightarrow$ p-value is large $\Leftrightarrow$ test statistic is close to zero (outside the critical region) $\Leftrightarrow$ if H0 were true, the data still could occur by chance

Reject H0 $\Leftrightarrow$ enough evidence to support Ha $\Leftrightarrow$ p-value is small $\Leftrightarrow$ test statistic is far away from zero (in the critical region) $\Leftrightarrow$ if H0 were true, the data is not likely to occur by chance

It's useful to become familiar with all the equivalent statements so that you can see hypothesis testing from all the angles. Perhaps the most important piece is the understanding of p-value, which we will revisit when we look at the hypothesis testing of means.

Summary: Hypothesis Test about One Proportion

First let's summarize the example above in terms of the five components of a hypothesis test. To test the claim that more than half of the Math 15 students are female, we proceed as follows:

a) State the null and alternative hypotheses.

$H_0 : p \le 0.50$ $\quad\quad$ $H_a : p > 0.50$

b) Show the test statistic.

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.608 - 0.500}{\sqrt{\frac{0.5 \cdot 0.5}{74}}} = 1.86$$

c) Show the P-value

$$P(\hat{p} > 0.608 \mid p = 0.500)$$
$$= P(z > 1.86)$$
$$= 0.031$$

d) State the decision of your hypothesis test

Reject the null hypothesis (since P-value < α)

e) State your conclusions in words a non-statistician would understand.

There is significance evidence to show that more than 50% of the students taking Math 15 are female.

Hypothesis Testing of One Parameter: A Simplified Outline

After reviewing some examples of conducting hypothesis tests of a mean or a proportion, let's summarize some of the main components of a hypothesis test here. Please keep in mind that our presentation is a bit simplified for the purpose of an introduction, and you may notice some differences when you consult other books on statistics. Two major simplifications were made:

- We assume that the research hypothesis is always the alternative hypothesis H_a. Although some textbooks allow both the null and alternative hypothesis to be the original hypothesis, testing a claim of equality is rarely done in actual application of statistics in research, because of the weakness in the null result: recall that we cannot state that the evidence supports the null. Instead, the best we can say is we cannot reject the null.

- We only discuss testing claims about a mean or a proportion. This is done to simplify the pictures we are showing for the P-value. As we shall see later in the semester, some test statistics have non-symmetric distributions and only one tail (such as testing a claim about standard deviation).

However, the most essential idea: using the null hypothesis and the data together to construct a conditional probability that we call "P-value", is shared across all different presentations of statistical inference. Once you reach some clarity about what a P-value is, then you can apply this understanding to any modern hypothesis test.

---

1. **State the null and alternative hypotheses**: identify the appropriate sampling distribution, parameter, and the tail of the test, and align the research statement with the alternative hypothesis.

Type of Data	Quantitative	Categorical
Sampling Distribution	Sample Mean: $\bar{x}$ (x-bar)	Sample Proportion: $\hat{p}$ (p-hat)
Parameter	Population Mean: μ	Population Proportion: p

Tail of the Test	Left-tailed	Right-tailed	Two-tailed
Symbol in H_a	$<$	$>$	$\neq$

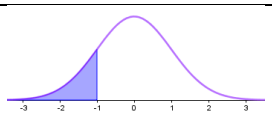
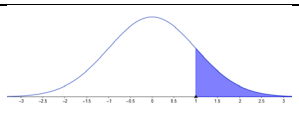
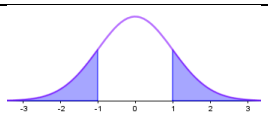
---

2. **Compute the test statistic based on the parameter and the assumption** (usually 2 decimals)

Proportion (p)	Mean (μ)	
	σ is given	σ is unknown
$z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{pq}{n}}}$	$z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$	$t = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}$

---

3. **Find the P-value based on the test statistic and the tail of the test** (usually 3 decimals)

Tail of the Test	Left-tailed	Right-tailed	Two-tailed
P-value as area			

---

4. **Interpret the P-value as a conditional probability.**

General interpretation: *Assuming H_0 is true, the chance of getting a mean / proportion from another sample of the same size that is more extreme than the current sample mean / proportion.*

Specific interpretation: *replacing "extreme" in the general statement with the appropriate tail.*

---

5. State the decision of your hypothesis test by comparing P-value with the significance level (α).

P-value < α : *reject the null hypothesis.*

P-value $\geq$ α : *do not reject / fail to reject the null hypothesis.*

---

6. State your conclusions in general terms.

Significant result (rejecting the null): *there is sufficient evidence to support the claim that … (H_a).*

Null result (not rejecting the null*): there is not enough evidence to support the claim that … (H_a).*

What is the meaning of P-value?

Importance of P-value in Hypothesis Testing

As discussed in previous notes, P-value gives us a way to quantify how much the data deviate from the null hypothesis: the smaller the P-value, the larger the discrepancy. Let's review the general meaning of P-value as follows:

"*Given that* H_0 *is true, the chance of seeing the evidence more extreme than the data (as per the appropriate sampling distribution and the alternative hypothesis).*"

In some places (such as the chapter quizzes), you see that "Given that H0 is true" is placed at the end of the sentence rather at the beginning. They are just two variations of the English syntax.

Now there are two places that you need to replace with the appropriate context of the problem. The first is what H_0 specifies: this is the easy part, since H_0 is the claim that contains the equal sign, which we use as the value of the parameter in computing the test statistic.

The second part, which corresponds to "*the evidence more extreme than the data"*, is a bit more complex. This event for which the P-value is a probability will depend on your data as well as the tail of the alternative hypothesis, as seen through these examples:

- Example 1 (from supplemental notes on testing the proportion): suppose there are 45 women in a class of 74. We are testing whether this means that the proportion of women is more than 50%.

We have seen that the appropriate hypotheses to use are: $H_0 : p \le 0.5; H_a : p > 0.5$

In addition, we also discussed how to compute the test statistic and P-value: P(Z > 1.86) = 0.031. Since the test is right-tailed (which means only sample proportions larger than 50% are consider evidence for H_a), the interpretation of p-value is as follows*:*

"If 50% of the students are female, then there is a 3.1% chance that in a sample of 74 students, 45 or more of them are female."

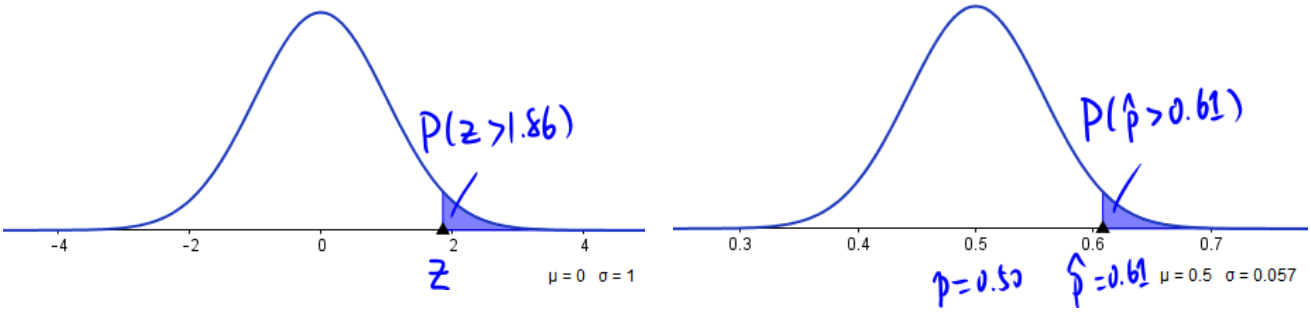
One way to see this is by using a conditional probability notation seen in Chapter 4:

$$P(X \ge 45 \mid p = 0.50) = 0.031$$

if you prefer using the sample proportion as evidence:

$$P(\hat{p} > \frac{45}{74} \mid p = 0.50) = 0.031$$

So there are really two ways to visualize the P-value: one is by using the test statistic (Z), which centers at 0; and the other is by using the sample proportion ($\hat{p}$) itself, which is centered at the parameter p.



It is the second graph that is used for the interpretation of P-value in this case. What is the connection between the two graphs? It's the test statistic that transforms $\hat{p}$ to the standard normal Z, which tells us the two tail areas are the same:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

- Example 2 (from team homework #8): Suppose we have a sample mean of 8.5, and we are testing whether the mean is significantly different from 10 by using $\sigma = 2$ (the Z statistic = -2.37)

From the wording of "significantly different", we should use $H_0 : \mu = 10; H_a \mu \neq 10$ as our hypotheses.

Being a two-tailed test, the P-value here is P(Z < -2.37 or Z > 2.37) = 2*P(Z<-2.37) = 0.034. Interpretation of P-value is a bit more nuanced than the one-tailed test as well:

"If the population mean is indeed equal to 10, then the probability that the sample mean will be less than 8.5 or greater than 11.5 is 0.034."

What's somewhat unusual here is that the sample mean of 11.5 is actually not seen in the data. However, since the two-tailed alternative hypothesis is supported by both small and large sample means, we will need to "make up" the unseen evidence by using the actual sample mean (8.5) and the claimed population mean (10), as well as the symmetry of the sampling distribution. Since the alternative hypothesis is two-tailed, both very large and very small sample means are considered evidence against H_0, and they are symmetric across the parameter $\mu = 10$ in H_0, according to the Central Limit Theorem (see my notes on this topic). Again, there are really two ways to visualize the P-value: one is by using the test statistic (Z), which centers at 0; and the other is by using the sample mean $\bar{x}$, which is centered at the parameter $\mu = 10$. Here we will only look at the second version, which is what we used to interpret the P-value.

Using the conditional probability to express this sentence, P-value can be stated as:

$$P(\bar{x} < 8.5 \ \ or \ \ \bar{x} > 11.5 \mid \mu = 10) = 0.034$$

This rather long notation explains why P-value is quite cumbersome to state. If it's helpful, you can break it up during the addition rule, since the two tails are mutually exclusive:

$$P(\bar{x} < 8.5 \mid \mu = 10) \ + \ P(\bar{x} > 11.5 \mid \mu = 10) = 0.034$$

- Example 3: On a test of numerical literacy, a random sample of 35 young men had a mean score of 272. By testing the hypothesis that the mean score for all young men is less than 275 (the national average), it was found that the t statistic = -1.45 and P-value = 0.078.

The test statistic has switched from Z to t (by using the sample standard deviation instead of σ), and the test is left-tailed (because of the alternative hypothesis states that $H_a : \mu < 275$). Notice that the sample mean of $\bar{x} = 272$ is also located to the left of the mean according to the null hypothesis. So we modify the generic statement of "*seeing the evidence more extreme than the data"* to reflect the left-tailed test as follows:

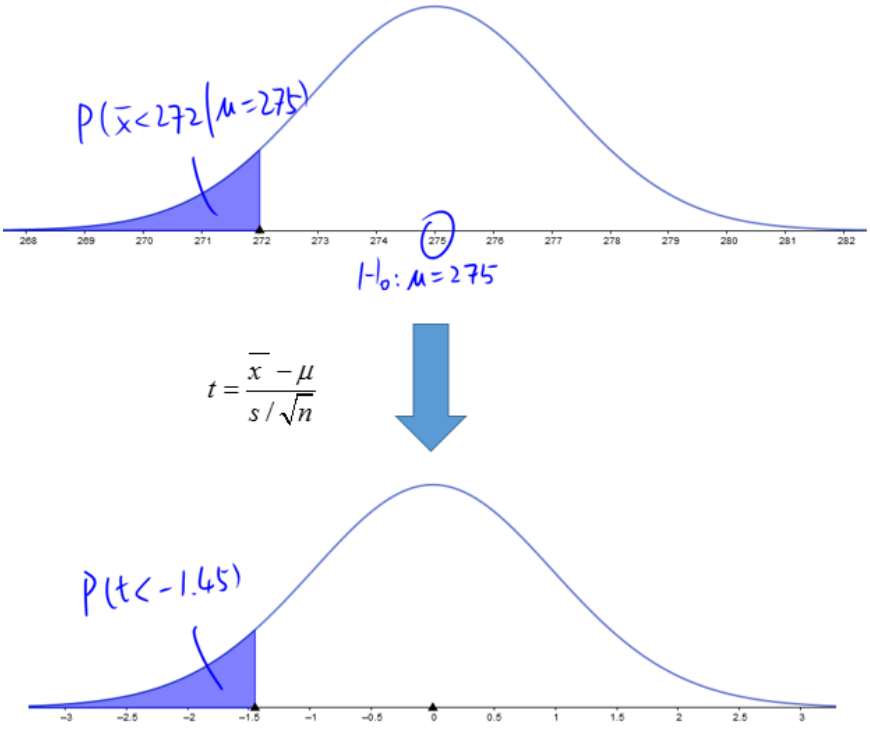
$$P(\bar{x} < 272 \mid \mu = 275) = 0.078$$

Put into words, and using a slightly different sentence structure, we can state the P-value interpretation as:

"If the population of all young men's scores is 275, then there is 7.4% chance that the sample mean of another 35 randomly selected young men will be less than 272."

The following figure illustrates the interpretation of P-value and its relationship with the t statistic. The top graph is helpful for interpreting the P-value, while the bottom graph is how we computes the P-

value based on the Student's t distribution. The two shaded areas are equal, and they are related through the t statistic formula in the middle.



Common Misinterpretations

Because P-value is such a complex idea, and it requires you to write a rather long sentence that you normally wouldn't use unless you are taking a course in philosophy, many people attempt to take shortcuts that turn out to be something else. I want to write about them so that you don't make similar mistakes in your work.

We'll use the Example #3 above to illustrate the various mistakes. Recall that the P-value in Example #3 can be described as a conditional probability:

$$P(\bar{x} < 272 \mid \mu = 275) = 0.078$$

Mistake #1: P-value as the probability of the Null Hypothesis

This is by far the most common mistake. Using the second example above, it will take the following form:

"The probability that $\mu = 275$ *is 0.078."*

People who use it tend to draw the correct conclusions: a small P-value leads to the rejection of the null, they reasoned, because the null is improbable. To see why this is incorrect, you can simply review the chapter on conditional probability. Just as we know that:

$P(A \mid B) \neq P(B)$

For the same reason:

P(evidence | null hypothesis) $\neq$ P(null hypothesis)

Since the null appears as the condition that is given, the P-value never really describes the probability of the null.

Mistake #2: P-value as the probability of the data

This mistake was mostly due to its incompleteness. Using the second example above, some people will say:

"The probability that $\bar{x} < 272$ is 7.8%."

What's missing here is the reference to the assumption that H_0 is assumed to be true. Without this reference, there is no logical reason why H_0 should be rejected under a small P-value.

Mistake #3: Switching the statistic and parameter

Occasionally people get the $\bar{x}$ and μ backwards, confusing what is suggested by your data and the null hypothesis:

"If $\mu = 272$, there is 7.8% chance that $\bar{x} < 275$."

An easy way to detect this error is that the P-value does not make sense given the tail of the test – if the sample mean is 275, and the population mean is 272, then we are expecting the P-value to be close to 1, instead of a small value like 0.078.

Summary

Since P-value is at the heart of statistical inference, you should know how to interpret it correctly for the exams as well as any upper-division courses that use the statistics we learned in this course. To avoid the mistakes with interpretation, I would recommend that you follow these steps:

- Start with the generic definition based on the conditional probability:

  P(evidence more extreme than current data | null hypothesis is true)

- Replace "Given…" with the appropriate H0 in your context
- Replace "evidence…" with the appropriate statistic from your data, together with the appropriate tail.

If your sentence is rather short, then chances that you have omitted one or more of the important pieces.

More on Type I and Type II Errors

After you have been doing hypothesis testing for a while, you may start to get a little wary: after call, this is not some sort of bullet-proof strategy for making decision, and there are two types of errors that could occur. At the back of your mind, you probably have the same question that people often ask:

- Is it better to make a Type II error than Type I error?

The short answer is that Type I Error is not necessarily "better" error than Type II. They are just simply two different animals. Let me use an example to illustrate how they are different. Suppose that you buy a 12-oz cup of coffee from Starbucks daily, but after 10 days, you think on average, you are getting less than what you paid for, since the sample mean is $\bar{x} = 11$ oz. So you would have the hypotheses:

$H_0 : \mu \geq 12$ oz

$H_a : \mu < 12$ oz

Suppose H_0 is true, i.e. you are getting an average of 12 oz or more, but you insist that you are getting less (i.e. rejecting a true H_0), you would have made a Type I Error. You can think of Type I error as the one made by the "overly critical" person, who sees something wrong within the smallest evidence. If it's helpful, you can use the image of a "conspiracy theorist" to remember that.

What about Type II error? (The probability of a Type II error is represented by another greek letter, β). If somehow the baristas at Starbucks were not properly trained, and they got used to putting less than 12 oz in your cup, then H_0 is false, and H_a is true. However, if your hypothesis test shows that the P-value > α, then you will not be able to reject H_0 (i.e. concluding that you are getting less than what you paid for). If someone makes Type II error all the time, then s/he might be a bit too oblivious of things that go wrong. You would probably not want to have this clueless person as your babysitter!

Among the two types of errors, Type I error is easier to control, since we can just choose a smaller α (significance level) so that we become less "critical" in finding something wrong. The smaller the α, the smaller the chance of a Type I error will be. In the coffee example, sufficiently small α values would not allow you to complain to the manager if your sample mean is 11 oz, but would prompt you to complain if your sample mean is 10oz. Using smaller α might make you look less aggressive, since you are giving them the benefit of doubt that it could be occurring by chance.

However, using extremely small α has an unintended consequence. If due to some systematic manipulation, the customers are only getting an average of 11.5 oz (i.e. $\mu = 11.5$, so H_a is true), with a very small α (say $\alpha = 0.0001$), it becomes extremely difficult to reject the false H_0. So the high chance of Type II errors means that coffee shops can get away with serving people less coffee than advertised, since you will have to have a very small sample mean (say $\bar{x} = 6$ oz) in order to reject H_0.

Our example shows that using α alone to control both type of errors is simply impossible, and it's not exactly useful to make the chance of Type I Error as small as possible either. How can we control the Type II Error (β) then? The answer to this, not mentioned in the textbook, is that we will have to increase the sample size (guess what? More data never hurts!), which leads to the increase of "**power**" of the hypothesis test, defined as $1-\beta$. You will study more about the power of the test, as well as how you use power to decide an appropriate sample size, in another advanced statistics course. The web applet we used for last week's discussion can be used to simulate how the sample size affects power of the test, if you are interested.

Here is the summary of the decisions again, together with the concept of power:

		Reality	
		H_0 is true	H_0 is false
Decision	Reject H_0	Prob of Type I: α	Power = $1-\beta$
	Do not reject H_0	Prob of correct null decision: $1-\alpha$	Prob of Type II: β

When Non-significant Results Matter: Two Stories about Cancer Treatments

Null v.s. Significant Results

In the past two weeks, we are gradually becoming familiar with hypothesis testing as a way to make decisions based on data. One of the most puzzling aspect of this process is the null result, i.e. "there is not enough evidence to support the alternative hypothesis / reject the null." As I emphasized in a previous post, the failure to reject the null is NOT the same conclusion as proving the null, as you can recognize from the saying "the absence of evidence is not the same as the evidence of the absence."

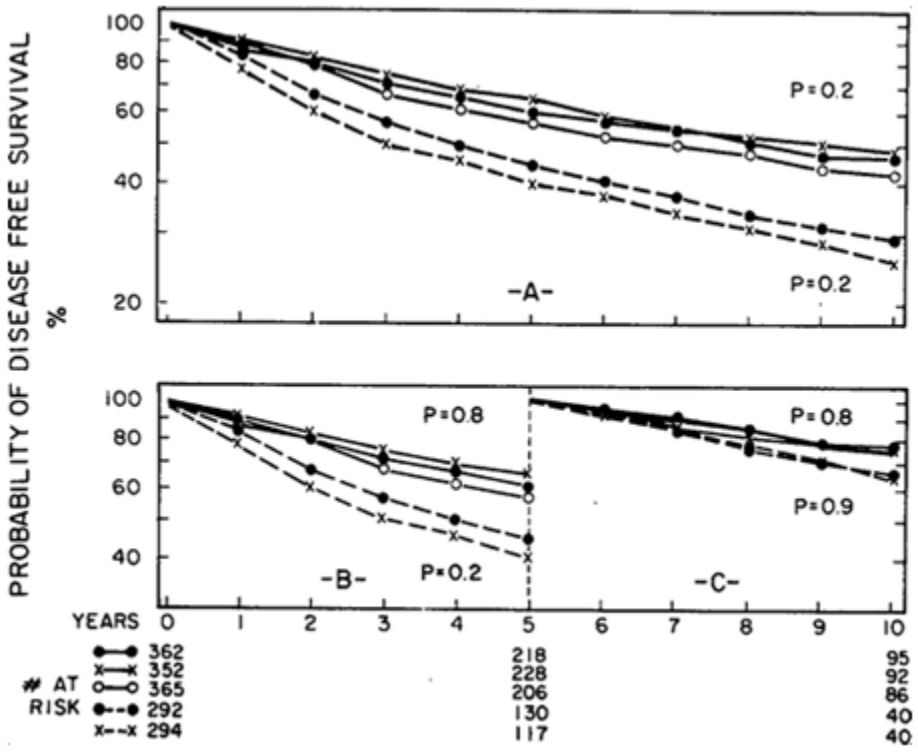
It's useful to recognize that the null result is generally bad news for the researcher: if you are working on your master's thesis in psychology, this could mean an extra semester conducting another experiment to support the main conclusion in your thesis; if you run a pharmaceutical company, a non-significant result in clinical trial could translate to zero return for millions of dollars of investment in Research & Development.

Although in actual research, it's far more likely to encounter null results than significant results, the public awareness is often drawn to the significant results for a simple reason -- it's difficult to publish/publicize results that are not significant. Earlier in the semester, I asked everyone to search for studies related to your favorite health foods, diets, or supplements, and not many people identified null results from this exercise. The lack of publically available null results presented an unfortunate problem: if a significant result is not replicated several times, then how do we know that there was no Type I error (rejecting H0 when it's true)? For every study that shows that green tea significantly lowers the risk of heart attack, how many studies did not find any significant difference?

Case 1: Lumpectomy

There are exceptions to the general rule, however. In rare occasions, null results that have a high impact on public health do attract the publicity it deserves. You may have heard of a new documentary series from PBS titled "Cancer: The Emperor of Maladies". One of the stories mentioned in the documentary is the story of radical mastectomy in the treatment of breast cancer. Radical mastectomy is a brutal surgery that is based on the mistaken notion that cancer spreads in concentric circles centered around the tumor. Based on the belief that "cutting more is better", surgeons often removed large areas of the body, including muscles in the chest and upper arm. When a doctor named Bernard Fisher first proposed a much more conservative surgical approach that only removes the areas immediately near the lump (known as "lumpectomy"), cancer doctors were skeptical, since many suspected that many

more patients would die if they did not receive radical mastectomy. It's not until the clinical trial for lumpectomy went for 10 years, enough data started to appear to convince people that lumpectomy did not reduce the survival rate of the patients. The paper by Fisher (http://www.nejm.org/doi/full/10.1056/NEJM198503143121102) included the following graph:



The three curves in Figure (A) correspond to the survival rates of patients receiving radical mastectomy, versus Fisher's more conservative surgery techniques. Since the significance level used in such trials was 0.05, a P-value of 0.20 (Fisher used a two-tailed test) would indicate a null result, i.e. the patients receiving lumpectomy did NOT have a significantly lower survival rate than the group receiving radical mastectomy. After Fisher's results were published, radical mastectomy was quickly abandoned as an unnecessary surgery that brought little additional benefits to patients but terrible consequences.

Case 2: High-Dose Chemotherapy

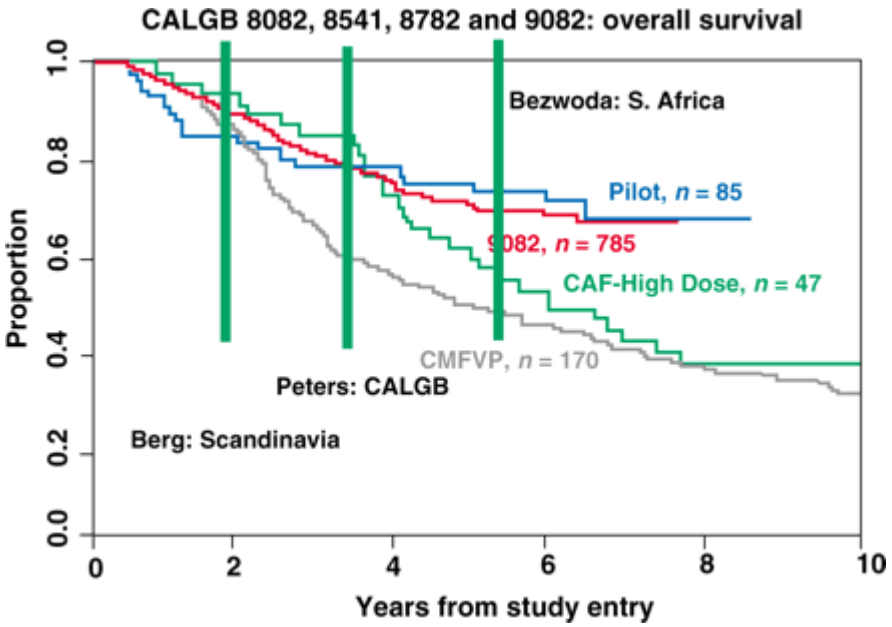
Although the story of lumpectomy represents the triumph of a null result, there were tragedies as well. Consider the case of High Dosage Chemotherapy (also discussed in the documentary), which uses dosages as high as 5 times of the previously accepted safe levels in treating cancer patients. Because the dosage was so high, patients had to have their bone marrows extracted and stored, so that they can receive their own bone marrow after being "knocked out" by the high dosage chemo. Although the initial small-scale trials showed promising results, larger clinical trials that were conducted in a 10-year period failed to produce significant results to show that the High Dosage chemo has extended the

patients' lives. But the news of the null results was too late for the many patients who were attracted to the clinical trial -- many of them died of complications from the dangerous treatment itself.

The findings from the clinical trials of High Dosage Chemotherapy was summarized in this 1999 paper:

http://theoncologist.alphamedpress.org/content/5/1/1.full#sec-1

The survival rate of patients a number of years after they received the therapy was plotted in the following graph (the gray and red curves were the control groups receiving regular dosage; while the green and blue were HDCT groups)



Although the authors attempted to analyze their data in various ways to find support for HDCT, the basic conclusion was fairly stark:

“However, due to limited sample size the *p* value is 0.1, not statistically significant.”

I know that cancer is not on most people's minds until it afflicts someone you know. But since someone we know or we ourselves will eventually die of cancer, I would like to suggest that an essential skill for improving the quality of your life is the understanding of probability and statistics. As an example, a friend of mine has been diagnosed with 3rd stage breast cancer. Among her extended family, there was also a history of various forms of cancer. What that the chance that her cancer was caused by a particular genetic mutation in the family? If she is confirmed to carry such a mutation, what is the chance that she will respond well to a particular "targeted" therapy that attacks the cancerous gene? These are questions that involve conditional probability. As she undergoes the standard chemotherapy after her surgery, her oncologist will present her with some difficult decisions: what are the pros and cons of a particular treatment program? What are the quality-of-life issues that affect the patients after their treatment is over? Decades of research has produced a large body of data, sometimes with

conflicting evidence. Understanding and interpreting these results will require a certain level of statistical literacy.

The progress on difficult diseases such as cancer is painting an increasingly complex picture of the problem, as we as a trove of tools that help us navigate the world of uncertainty. To use these tools effectively, we will need the ability to reason from and make decisions based on data. As a former researcher myself, I believe that this is a skill that most people can learn and use to empower themselves.

In Memory of Mariah Roat, a former Math 15 student at SRJC who lost her battle against cancer. You can read about Mariah's story by going to

http://www.theoakleafnews.com/features/2013/11/21/unbreakable-spirit-mariah-roats-fight/

Team Homework #7: Hypothesis Testing (Part 1)

1. **Identify the random variable in each of the following situations, state the null and alternative hypotheses for each situation below, and indicate whether it's a one or two-tailed test. Be sure to use the correct notations.**

a) The diameter of a spindle in a small motor is supposed to be 5mm. If the spindle is too large or too small, the motor will not work. The manufacturer measures the diameter in a sample of motors to determine whether the mean diameter has moved away from the target.

Quantitative / Categorical?: _____

Hypotheses: _____

Tail of the test: _____

b) In a multiple choice test, each question has 5 choices (ABCDE). A student guessed all of the answers and missed all of them. He claims that he is doing worse than chance.

Quantitative / Categorical?: _____

Hypotheses: _____

Tail of the test: _____

c) The exams in a large business class are scaled after grading so that the mean score is 50. The professor thinks that one teaching assistant is doing a poor job and suspects that his students have a lower mean than the class as a whole. The TA's students this semester can be considered a sample from the population of all students in the course, so the professor compares their mean score with 50.

Quantitative / Categorical?: _____

Hypotheses: _____

Tail of the test: _____

d) The New York Times and CBS News conducted a national poll of 1048 randomly selected 13 to 17-year olds and counted how many owned smart phones. The reporters were trying to decide whether more than half of teenagers own smart phones.

Quantitative / Categorical?: _____

Hypotheses: _____

Tail of the test: _____

2. For scenario (b) and (d) of the previous question, state the Type I and Type II errors in complete sentences.

1(b):
Type I Error:

Type II Error:

1(d):
Type I Error:

Type II Error:

3. Complete the following simulation and answer the questions.

   a. Use a Normal random number generator (GeoGebra or calculator) to generate 10 data values randomly using $\mu = 10, \quad \sigma = 2$. Record their values below:

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

   b. Find the sample mean from these 10 data values:

      $\bar{x}$ = _____inches

   c. At the 0.20 significance level, use your random sample to test the hypothesis that the population mean is not equal to 10. What is the null and alternative hypotheses?

   d. Which hypothesis is true?

e. What is the value of the test statistic (z)? Show your work.

f. What is the P-value of this test? Draw a figure that illustrates the area corresponding to the P-value.

g. State the meaning of P-value in this particular situation.

h. What decision do you arrive at, based on your P-value? State your conclusion in complete sentences.

i. Write your test statistic, P-value and the decision on the whiteboard (if you have a group of 4, please repeat the experiment and show two sets of answers). Record all the decisions that you see on the board and indicate which ones are examples of Type I error.

j. Based on the discussion above, what is the meaning of "the 0.20 significance level"? How is this interpretation related to the concept of sampling distribution?

4. Psychologists wish to investigate the learning ability of schizophrenic people after they have taken a specified dose of a tranquilizer. Thirteen patients were given the drug, and one hour later they were given an exam. Their scores are listed below (data also available in the spreadsheet)

13	20	30	27	24	22	22	17	21	25	23	27	25

   a. Generally, patients score about 20 on the exam without the tranquilizer. Is there statistical evidence that taking a tranquilizer has affected their scores (α=0.05)?

      i. State the null and alternative hypotheses.

      ii. Show the test statistic.

      iii. Find the P-value and interpret it as a probability.

      iv. State the decision of your hypothesis test

      v. State your conclusions in words a non-statistician would understand.

5. (Extra Credit) Using the same procedure as in Problem #3, show how decisions made from certain random samples can result in Type II error.

Team Homework #8: Hypothesis Testing, Part 2

Part I: For each of the following scenarios and the test statistic, find the P-value and state the conclusion given a significance level of 0.05.

a) The diameter of a spindle in a small motor is supposed to be 5mm. The manufacturer measures the diameter in a sample of motors to determine whether the mean diameter has moved away from the target. ($t = 2.14$, $df = 20$)

b) In a multiple choice test, each question has 5 choices (ABCDE). A student guessed all of the answers and missed all of them. He claims that he is doing worse than chance. ($Z = -0.75$)

c) The exams in a large business class are scaled after grading so that the mean score is 50. The professor thinks that one teaching assistant is doing a poor job and suspects that his students have a lower mean than the class as a whole. ($t = -1.04$, $df = 60$)

d) The New York Times and CBS News conducted a national poll of 1048 randomly selected 13 to 17-year olds and counted how many owned smart phones. The reporters were trying to decide whether more than half of teenagers own smart phones. ($Z = 3.17$)

e) Ted's Taxi Company bought 32 cars from the dealership that are advertised to get 28 miles per gallon in fuel efficiency. The company suspects that the average fuel efficiency per car is more than advertised. ($t = 2.01$)

f) Monique wants to test whether at least 70 percent of her employees approve of the paid-time-off policy. ($Z = 3.41$)

Part II: Include all of the following components for each hypothesis test.

a) State the null and alternative hypotheses.
b) Show the test statistic.
c) Show the P-value and interpretation of the P-value as a probability
d) State the decision of your hypothesis test
e) State your conclusions in words a non-statistician would understand.

1. Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 tasters for one new cola recipe (data also available in the spreadsheet):

2.0	0.4	0.7	2.0	-0.4	2.2	-1.3	1.2	1.1	2.3

   Are these data good evidence (at the 10% level) that the cola lost sweetness? (Hint: no loss of sweetness will lead to the value of 0).

2. According to the data released by National Shoe Retailers Association, the average American woman's shoe size is 8.5 (up from 7.5 a decade ago).
   a. Use the class data to test the hypothesis (at the 0.05 level) that the mean shoe size for female students taking Stats class is different from the national average.

b. How is your conclusion useful for shoe stores?

3. A recent report claimed that 20% of all college graduates find a job in their field of study. A survey of a random sample of 500 graduates found that 110 obtained work in their field. At the 5% level, test whether the claim in the report was accurate.

4. In a study of smokers who tried to quit smoking with nicotine patch therapy, 39 were no longer smoking one year after the treatment and 8 were still smoking one year after the treatment.
   a. Use a 0.10 significance level to test the claim that among smokers who try to quit with nicotine patch therapy, less than half are smoking a year after the treatment.

   b. What does your test show with regard to the effectiveness of the therapy?

5. USA Today conducted a nationwide poll of 1156 randomly selected adults. Of these adults, 610 said that they approve of President Obama's performance. Is there good evidence (at 0.01 level) that the majority of American approve of President Obama's performance?

6. In a study of fast-food intake, 3857 customers at three hamburger chains (McDonalds, Burger King, and Wendy's) were approached as they entered the restaurant at lunch and asked to provide their receipt when exiting. The receipts were used to determine the amount of calories consumed by each person. The sample mean was calculated to be 857 calories and the sample standard deviation was 677 calories.
    a. Health experts recommend a target of 750 calories for lunch, based on a daily intake of 2000 calories. Does the study provide convincing evidence that the average carlories consumption by the hamburger chain customers is greater than the recommendation?

    b. If you were a Public Relations spokesperson for the hamburger chains, how would you react to the conclusion of the study?

Confidence Interval for Means

Last time, we discussed the connection between sampling distributions, Central Limit Theorem, and the confidence interval for sample proportions. This week, let's take a look at the problem of estimating the sample means.

Estimating Mean from Quantitative Data

First, we should be clear on how our task has changed when we move from estimating proportions to estimating means. As a parameter, proportion is a value between 0 and 1, while the mean can be any real number. For example, regarding our class data as a sample of the JC student population, you can estimate the proportion of students whose favorite color is blue, but you can also estimate the mean height of students. It would not make much sense for you to estimate the mean from the qualitative data that only has "blue, orange, yellow, …", nor would it make a lot of sense to estimate a proportion from the data that contain the heights of all students (you will have to convert it to qualitative data first, such as "taller than 6 ft (yes/no)", etc.) Whether you are estimating a proportion or a mean depends on you data, and the question you are asking. The difference because proportion v.s. mean goes all the way back to the difference between qualitative and quantitative data in Chapter 1. Quite often, students fail to differentiate which problem they are solving, and end up missing the entire problem in the exams.

Other than heights, other possible population means that can be estimated using our class data include AGE, SHOE, PETS, and CAR. In each case, the data is quantitative, and you can think of the task of estimating the mean as a way to answer the question: "what is the average … of the students who attend SRJC?".

One you have identified the task is to estimate the population mean, you will notice that our text book devotes two separate sections in Chapter 8 to two different scenarios. The primary difference between the two scenarios is whether the population standard deviation is known. To give you some context, let us consider another game that I sometimes play with students in class.

Example: Guessing Age Given Population Standard Deviation

In this game, everyone is asked to guess how old I am. Based on the responses, we can construct a confidence interval to answer the question:

- "How old does Dr. Lin look, on average?"

Since I have been doing this activity for many years in my stats class, I have noticed that although the mean of people's guesses keeps going up year after year (I wonder why this is the case :), the standard

deviation is remarkably stable (approximately 2.6 years). If our class is no exception (unless many of you possess supernatural powers or have seen my passport), I can safely assume that the population standard deviation for my perceived age is known to be 2.6 years. Take, for example, one of my classes last semester:

35 33 32 46 33 32 38 38 35 32 32 35 32 37 35 42 32 33 38 38 33 35
34 39 38 35 36 32 34 34 30

(I do not know who said I looked like I was 46, but I assume this person was not joking, since that will constitute a sampling error)

The sample mean of these 31 guesses was 35.1 years. Recall that in Section 6.5 (Central Limit Theorem), if we know the mean μ and the standard deviation σ of the population, then the sample mean $\overline{x}$ will follow a normal distribution, with mean given by:

$$\mu_{\overline{x}} = \mu$$

(In English: the mean of the sample means is the same as the population mean.)

Note we previously used μ_X instead of μ, but here we try to stick with μ to emphasize it is the parameter we are trying to estimate.

And the standard deviation (also known as the standard error of $\overline{x}$) is given by:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

(In English: the standard deviation of the sample means decreases inversely proportionally with the square root of sample size)

What now follows us parallel to the lesson on estimating proportions: the CLT for sample means leads to the following result that is similar to what we saw with sample proportions: if we convert $\overline{x}$ to a standard normal distribution, we will obtain:

$$Z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

Here the denominator is the standard error. Once we multiply the denominator on both sides, we will obtain the Margin of Error formula for estimating the mean:

$$E = Z \cdot \frac{\sigma}{\sqrt{n}} = \overline{x} - \mu$$

What remains to be decided is how large the critical value $Z_{\alpha/2}$ is acceptable, based on your confidence level. Suppose I've decided that I needed a 95% confidence level (if you forgot how to find $Z_{\alpha/2}=1.96$, just review the previous lesson), then by using $\sigma=2.6$, and $n=31$, I can evaluate the margin of error as:

$$E=1.96\cdot\frac{2.6}{\sqrt{31}}=0.9 \text{ years}$$

So to find my confidence interval, I just need to go above and below this margin from the sample mean (point estimate):

$$34.2=\bar{x}-E<\mu<\bar{x}+E=36.0$$

Because of my confidence level is 95% (which means out of 20 classes, only one will miss my age entirely in their interval), this is a pretty good estimate how old I generally look to people. (If you would like to review the meaning of confidence level, please refer to the web applet mentioned at the end of last lesson.)

As a trivial application of this confidence interval, the picture on my webpage is now at least a few years old. So if I want to save the trouble of taking another one, I could do this experiment by inviting a couple of hundreds of people from Facebook to take a look at my photo, and ask they how long I look from the photo. If the confidence interval turns out to be something like 24 to 26, then I'd better take a new picture!

Finding Sample Size

Similar to the sample size for estimating proportions, the formula is related to the margin of error for estimating the mean. Based on:

$$E=Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Squaring both sides, multiplying by n and dividing by E^2 on both sides you obtain the sample size formula (you can review the notes on estimating proportion for the algebra used):

$$n=\frac{(Z_{\alpha/2})^2\cdot\sigma^2}{E^2}$$

So if you want to do my experiment in a birthday party, and you plan to use a 95% confidence level, and 0.5 years as your margin of error, you will need this many people at the party:

$$n = \frac{1.96^2 \cdot 0.9^2}{0.5^2} = 12.45$$

Since it's not possible to include half a person, we'll round this up to 13 since we are trying to keep the margin of error within 0.5, since as E decreases, the sample size n should increase.

Introducing the Student's t distribution

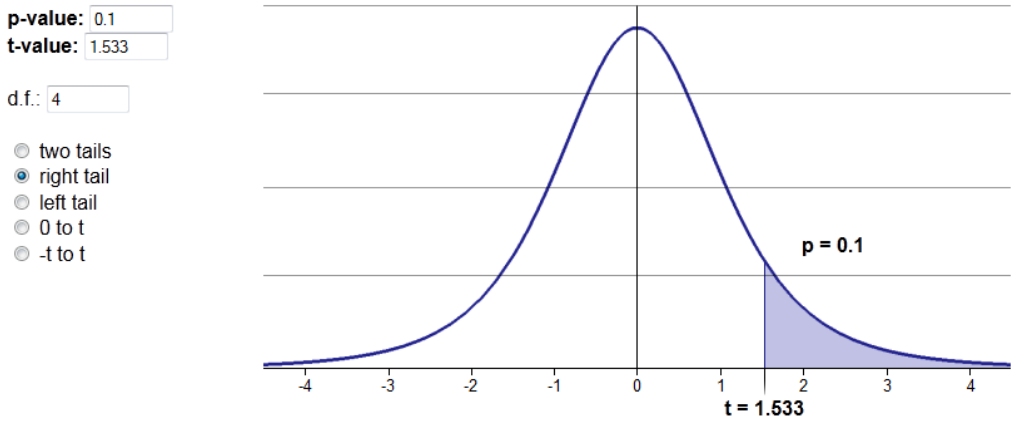
What we did with estimating the mean looks quite reasonable, right? What could go wrong here? Now imagine that I wanted to do this experiment on Halloween, and came to class dressed in a costume. Now this may have some consequences with regard to the standard deviation, since I am not wearing my usual semi-professional outfit anymore, and people's guess can differ wildly. If we look at our original assumption that the population standard deviation is given, this may not be true anymore, and Central Limit Theorem does not apply, since σ is hidden from us.

What can we do instead? It turns out, this was a problem solved by a guy who worked at the Guinness Brewing company that makes the famous beer from barley. What he discovered that was if we replace σ in the formula $Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ by the standard deviation calculated from the sample, the result is a random variable that follows a different distribution, named Student's t-distribution ("student" was the pseudonym used by the author to keep the trade secret for the Guinness Company):

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

(notation review: although both were referred to as "standard deviations", s refers to the statistic, while σ refers to the parameter)

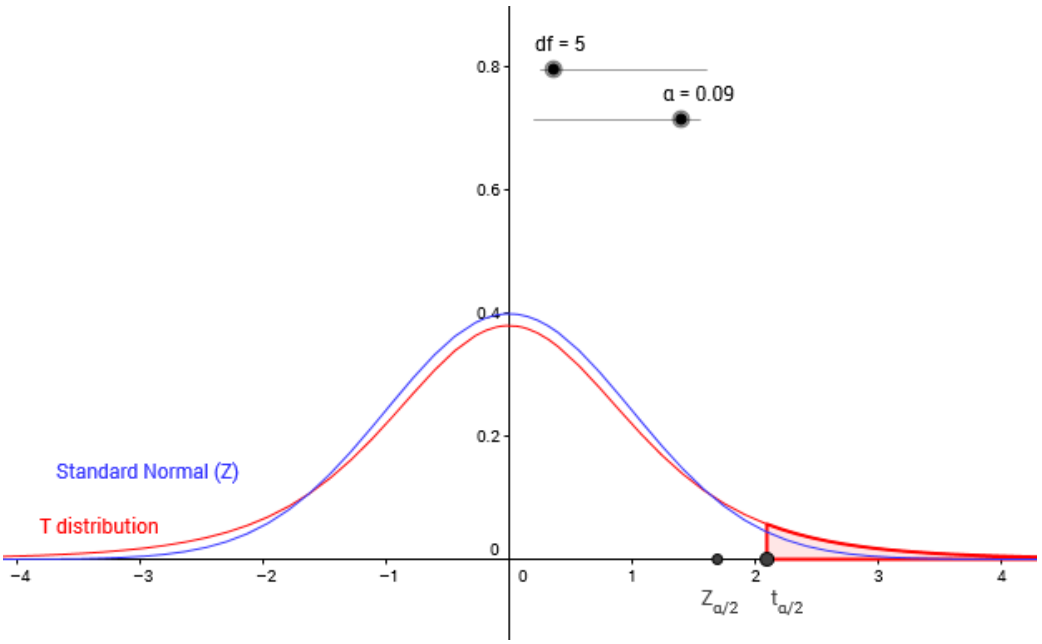
A picture of the t-distribution is shown here, together with the right-tail probability associated with $P(t > 1.533)$



Although the graph of the t-distribution does look like its poor cousin, the standard normal (Z), the main thing that distinguishes the t-distribution from the normal is the so-called "degree of freedom", which is related to the sample size through $df = n - 1$ (remember that the sample standard deviation from Chapter 3: $s = \sqrt{\frac{\sum (x-\mu)^2}{n-1}}$)?

It is useful to compare the t-distribution and the normal distribution side-by-side. From the pictures in the book and in GeoGebra, they appear to be the same. However, there are noticeable differences, especially when the degree of freedom is small. This applet allows you to explore the t-distribution by changing the degree of freedom, and comparing it with the Standard Normal distribution (Z). I have also included a visualization of the critical values under each distribution.

http://www.geogebratube.org/student/m53246



To explore the applet, drag the "df" slider to change the degree of freedom, and "alpha slider to change the value of alpha.

GeoGebra includes a Student's t calculator that we will use for most of the calculations related to the t-distribution. The way you use it to solve for probability/cut-off value is quite similar to the normal calculator. If you can't figure it out by trial and error, this video may help:

http://screencast.com/t/AIacPku2

Example: Guessing Age Without Population Standard Deviation

Going back to our original problem of estimating my average perceived age (still using the 95% confidence level), how would our interval change, based on the t-distribution? Well, it is actually quite straightforward. The main difference in how you find the critical value:

$$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

For the degree of freedom we need ($df = 31 - 1 = 30$), the critical value under the t-distribution is 2.04 (see the following screen shot). If you compare this critical value with its poor cousin $Z_{0.025}$, you can see that $t_{0.025}$ is larger, because of the larger tail under the t-distribution.

These two method return exactly the same critical values for the t distribution, except that if you use the table, your options for α are limited to the common suspects: 0.10, 0.05, 0.01, etc.

Using $t_{0.025} = 2.04$, $s = 3.4$ (you will need to use the original data to calculate the sample standard deviation, and the "Summary Statistics" function in StatCrunch can be handy for that), we have arrived at a new margin of error:

$$E = 2.04 \cdot \frac{3.4}{\sqrt{31}} = 1.2$$

Going above and below this margin from the sample mean (point estimate), we have found the new confidence interval:

$$33.9 = \overline{x} - E < \mu < \overline{x} + E = 36.3$$

If we compare this interval with our earlier result (from 34.2 to 36.0 years), you can see that although both intervals were centered around 35.1, the sample mean / point estimate, the interval from the t-

distribution tends to be more conservative, in that it is based on a higher margin of error (approximately 3 months), thus resulting in a slightly more vague estimate of my perceived age.

You are probably wondering: which is the "right" way of doing it? Well, unlike Algebra, where each equation has a set of solutions no matter how you solve it, Statistics is something a little different. In our case, the two confidence interval could be both "right", depending on what assumptions you are taking. If you want to go for a more accurate estimate (smaller margin of error) by using the standard deviation from the years of data I have accumulated, then the technique in the last section is applicable. However, if you want to be more cautious (as in the Halloween scenario), then play it safe and use the sample standard deviation is the way to go. The key is to check whether your sample indeed comes from the population where σ comes from.

So to sum up, here is how you can show off your newly acquired statistical prowess with some caution: the first time you ask a group of people how old you look, you should use the t-distribution; after a while, when you have got some data to back yourself up, you can impress people even more by using the standard normal Z. Let me know how it has worked for you!

(Photo of William Gosset, the “Student” who invented the t-distribution while working for the Guinness Company)

Summary: Estimating Means

To briefly summarize the process of constructing a confidence interval for a population mean, we do the following:

1. Start from the sample mean $\bar{x}$: this point estimate will be the mid-point of the interval.
2. Choose a sampling distribution: if the population standard deviation σ is given, then we use the standard normal $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$; otherwise, we use the Student's $t = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}$
3. Compute the critical value: either $Z_{\alpha/2}$ or $t_{\alpha/2}$ based on your choice of sampling distribution above. In either case, α = 1 – Confidence Level, and $\alpha/2$ is the area to the right of the critical value.
4. Compute the margin of error using either $E = Z_{\alpha/2} \cdot \dfrac{s}{\sqrt{n}}$ or $E = t_{\alpha/2} \cdot \dfrac{s}{\sqrt{n}}$
5. Add / subtract the margin error from the sample mean to obtain the interval: $(\bar{x} - E, \bar{x} + E)$
6. State the confidence interval in a sentence such as "with (C.L.) confidence, the population mean is between … and …."

Confidence Interval of Proportions

Before you start this new lesson, it's quite helpful to review the previous lesson on Sampling Distribution and Central Limit Theorem, especially the section on sample proportions. As we start introducing the new terminologies of statistics, the image of the hand and pail, used to illustrate the relationship between statistics and probability, will be especially helpful.

Point Estimate for the Population Proportion

when we have a binomial variable, like a 'yes' or 'no,' a 'male' or 'female,' or a 'success' or 'failure', 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 sample of the data set and calculate the **sample proportion** for this category. Let's say we're interested in the fraction of SRJC students who are female. From the Class Data Base, we find that 35 out of 60 are famel. We call 35, the number in our category (female), *x*, using the "p-hat" notation (our book uses p', which is non-standard) introduced in the previous lesson, we have:

$$\hat{p} = \frac{x}{n} = \frac{35}{60} = 0.583$$

Or 58.3%. What should we guess for p, the population proportion of female students? The obvious choice is to just use $\hat{p}$, .since we have some faith in the randomness of our sample. When we guess a parameter (p) by picking a single number like $\hat{p} = 0.583$, we are making what's called a point estimate. If we're not going to use $\hat{p} = 0.583$ as the point estimate, why did we bother to collect the data in the first place?

The Margin of Error and Confidence Level

But what would happen if we chose a different sample of 60 SRJC students? Probably it would have a different number of female students, not exactly 35. We can't say they are wrong and we are right, since our different proportions were just due to the randomness in the samples.

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 level (which we

will go into in great detail) that the population proportion falls between $\hat{p}-E$ and $\hat{p}+E$. We could also use the compound inequality notation of algebra: $\hat{p}-E<p<\hat{p}+E$. In some research papers, people also use the notation $\hat{p}\pm E$. These are all considered **interval estimates** for the parameter, since they specify a range instead of pointing to a single number.

E is called the **margin of error** and is also abbreviated EBP in the book. As you can see, constructing the interval amounts to finding the right E, since we already have the point estimate to start with. In a bit we will look at the 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 level.

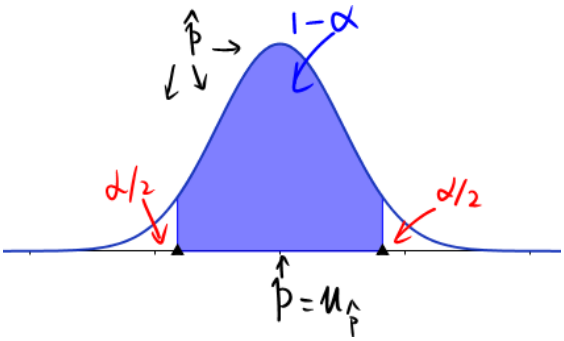
What if I just pick 0.5 as the margin of error for the proportion? In the gender example, that would mean that 0.583-0.5<p<0.583+0.5, or 0.083<p<1.083. How confident am I that this statement is true? One hundred percent confident! Since the proportion is a number between 0 and 1, the statement is of absolutely no use. It gives no information whatsoever about where the true proportion is likely to be.

What about a smaller E, say 0.001? That would mean 0.583-0.001<p<0.583+0.001. This does make it much more precise, but how sure could I be that I really captured p? Not very much. There is clearly a trade-off: the wider the interval, the more confident you are; but the accuracy of the estimate suffers. On the other hand, extremely narrow intervals lead to a low level of confidence and are not exactly useful either.

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%?

The next notation to introduce is α, which is just 1-CL. For example, if the CL is 90%, then $\alpha=1-0.90=0.10$. Since CL represents the probability that you are "right", then α is the probability that you are making a mistake with your estimate.

Think now of the Central Limit Theorem, which says that the sampling distribution of sample proportion $\hat{p}$ will be approximately normal, with the population proportion p as its mean. I want to go left and right from this mean a certain distance along the 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 under the curve on either side of the limits: $\alpha/2$



So the problem of finding E, the margin of error, is really no different from the "inverse probability" problems we saw in Chapter 6 involving finding cut-off values from the area in the middle. Here the area in the middle is CL (hence the area in the tail is $\alpha / 2$), and the cut-off values are the lower and upper limits of the confidence interval.

Let's try to construct a 90% confidence interval from our data that shows 35 out of 60 students are female. Since CL=0.90 (area in the middle), we know that the area of each tail is $\alpha / 2 = 0.05$. So if we have the mean and standard deviation of the sampling distribution above, we are all set.

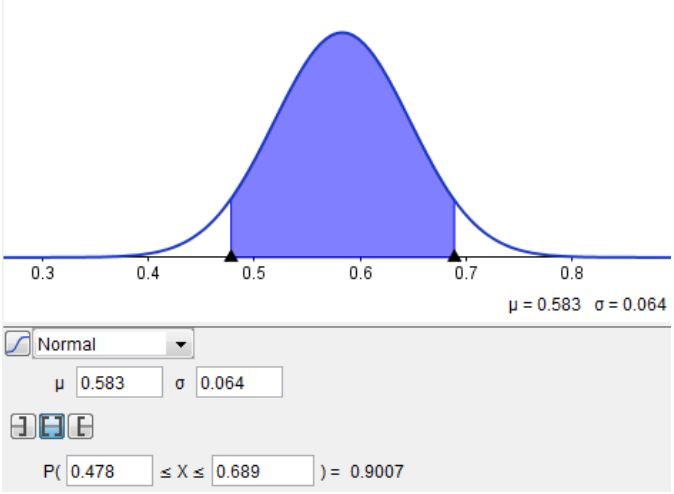
What do we do with the mean? Although p is unknown, we can replace it with the point estimate $\hat{p} = 0.583$. I know it sounds a bit circular, but keep in mind that our $\hat{p}$ is just one of the many such proportions that could emerge from a sample of 60.

What about the standard deviation? (also known as **standard error of sample proportions**, not to be confused with margin of error) The last bit of the Central Limit Theorem for Sample Proportions comes in handy here. It says:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Again, we will have do the same guess for p and q, as we did for the mean: we replace them with the point estimates $\hat{p}$, and $\hat{q} = 1 - \hat{p}$.

So by using the 0.583 as the mean, and $\sigma_{\hat{p}} = \sqrt{\frac{0.583 \cdot (1 - 0.583)}{60}} = 0.064$ as the standard deviation, we can solve the cut-off values that form the 90% confidence interval: $P(0.478 < \hat{p} < 0.689) = 0.90$



What happened to the margin of error E? It seems that we have found the confidence interval without using it at all. But remember the good-old days when everything has to be done through the standard normal Z? (you may want to review these notes) One of the first thing to recognize is that the margin of error is the difference between point estimate $\hat{p}$, and the true parameter p:

$$E = \hat{p} - p$$

On the other hand, since $\hat{p} \sim N(p, \sqrt{\frac{pq}{n}})$, we can also convert it to the standard normal Z:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Multiply the denominator on both sides, we have:

$$Z \cdot \sqrt{\frac{pq}{n}} = \hat{p} - p$$

Comparing this with $E = \hat{p} - p$, we can see that $E = Z \cdot \sqrt{\frac{pq}{n}}$. Notice this equation says that the margin of error is actually a random variable related to the standard normal Z by a constant factor $\sqrt{\frac{pq}{n}}$. But if we want to choose a cut-off value for E, we will need to do two things:

- Guess p and q using their point estimates $\hat{p}$ and $\hat{q} = 1 - \hat{p}$
- Choose a cut-off value for Z that is appropriate for the desired confidence level

Hence the formula you see in the textbook:

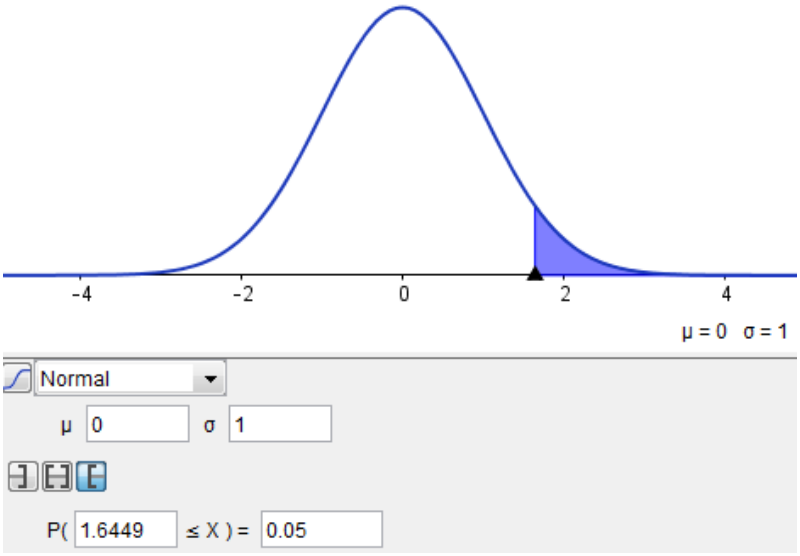
$$E = Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Here $Z_{\alpha/2}$, known as the **critical value**, is the cut-off value that marks the area to the right tail of $\alpha/2$. Notice the use of subscript here is a bit peculiar, but it's kind of stuck as part of the standard notation.

Example: Estimating the Proportion from the Class Data

Let's make things a bit more concrete using our example of class data: from x = 35, n = 60, we already knew that $\hat{p} = 0.583$. We shall see that the process of finding the margin of error E leads to exactly the same interval as the one we solved directly by using Central Limit Theorem.

First, from CL=0.90, we know that $\alpha / 2 = 0.05$. So the critical value $Z_{\alpha/2} = 1.64$, found through GeoGebra or your calculator:



Now using the margin of error formula, we look for E:

$$E = 1.645 \cdot \sqrt{\frac{0.583(1-0.583)}{60}} = 0.105$$

If you are using your calculator to evaluate this but not getting the same answer, I recommend that you check your parentheses to see whether the square root encloses the entire fraction, something like the following will probably work on your calculator: 1.645*sqrt(0.583*(1-0.583)/60)

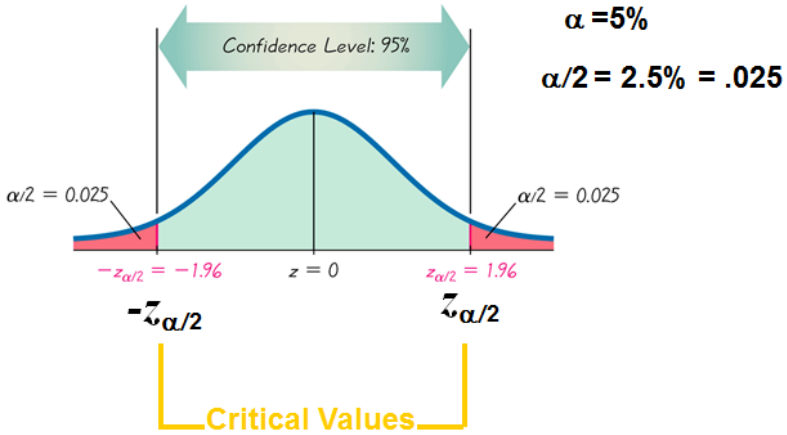
At last, we go above and below the point estimate with this error, arriving at the interval: 0.583-0.105<p<0.583+0.105, or $0.478 < p < 0.689$. There are two other ways to state this interval:

- Using the interval notation in algebra: $p \in (0.478, 0.689)$
- Using the point estimate plus/minus the error: 0.583 ± 0.105

It's useful to state the confidence interval in words that tell people what you just did, and I expect you to do so in the exams. Here is a statement that you can use as a template:

- With 90% confidence, the proportion of female students in SRJC is between 47.8% and 68.9%.

If, for some reason, you would like to use another confidence level, say 95%, then you will just have to re-compute $Z_{\alpha/2}$ as follows: $\alpha = 1 - CL = 0.05$, $\alpha / 2 = 0.025$, and $Z_{\alpha/2} = 1.96$, as illustrated below:



Then your new margin of error will be:

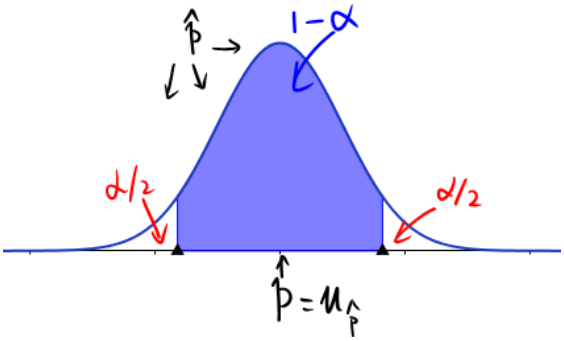
$$E = 1.96 \cdot \sqrt{\frac{0.583(1-0.583)}{60}} = 0.125$$

Which is larger than the 90% confidence interval since you've chosen cut-off values further away from the mean so that the confidence level goes up to 95%.

Simulating Confidence Intervals from a Sampling Distribution

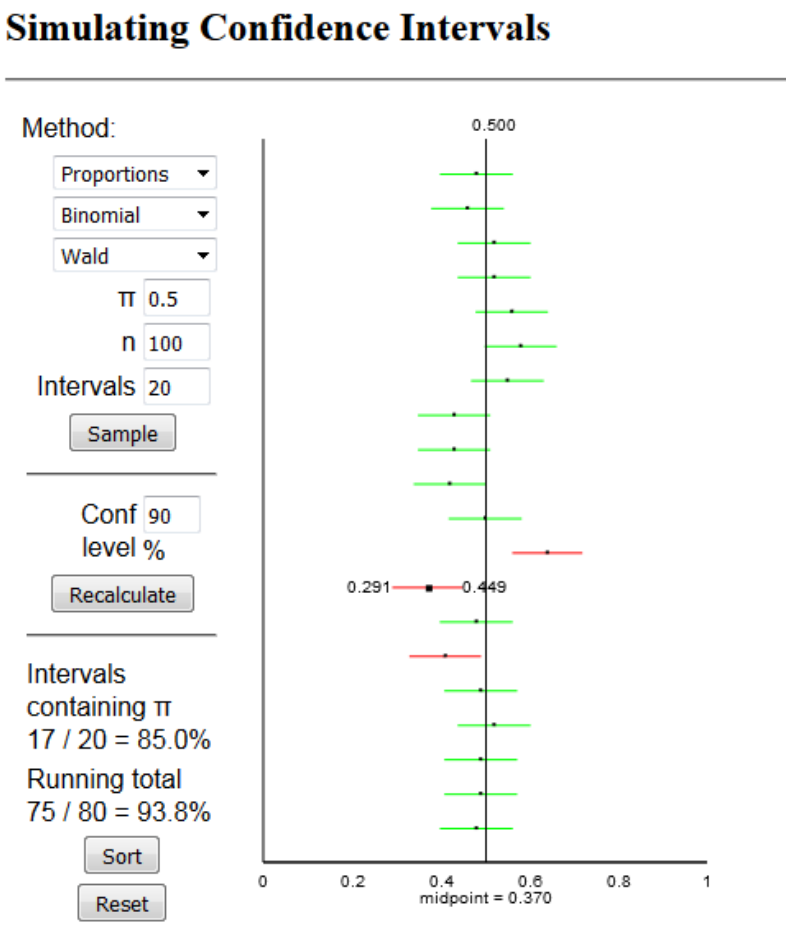
If this feels very complicated to you, don't feel discouraged! It also took me a while to see the connection between sampling distribution and statistical inference when I first studied this subject. But once you see that the idea of sampling distribution and Central Limit Theorem dictates everything we do, the rest of statistics will be much, much easier to understand.

Fortunately, there are now tools that can help you grasp the concepts that were not available when I was a student. For example, what does "95% confident" really mean? Obviously, it has nothing to do with your self-image and self-esteem. We referred to this picture earlier as the correct interpretation of Confidence Level:



It's useful to see if from another point of view, nicely demonstrated by this applet:

http://www.rossmanchance.com/applets/ConfSim.html



Here the bars indicate 20 different confidence intervals constructed from 20 different random samples, each containing 100 values that are binomial. The green bars indicate the intervals that contain the true parameter p=0.50 that is used to generate the random samples; the red bars are the ones that missed the parameter (I've highlighted one of them, 0.291<p<0.449, so that you can see why it has missed). Out of 20 intervals, 3 of them missed the target, which is what you expect with a 90% confidence level.

- *Interpretation of 90% confidence level: if many samples with the same size (60) are drawn from the same population to create confidence intervals, then approximately 90% of these intervals will contain the parameter.*

A useful image is the game of the ring toss game you may have seen in carnivals, except that you are playing it in the dark. The useful analogy here is that the parameter is the pole, an unknown but fixed value. The confidence intervals you create are the rings that you are using to toss at / estimate the pole. Although you don't know where the pole is, the confidence level is the percentage of the rings that will hit the pole.

Finding Sample Size

Our goal here is to produce a formula for the sample size, given the desired margin of error. This problem can emerge as the part of the experimental design: how many people do you have to call/survey in order to accomplish the desired margin of error? Because of the relationship between E and n, all we need to do was rearranging the formula for *E* in estimating means that produced the formula for *n*

$$E = Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Squaring both sides, this leads to:

$$E^2 = (Z_{\alpha/2})^2 \cdot \frac{\hat{p}\hat{q}}{n}$$

Multiplying both sides by n,

$$n \cdot E^2 = [Z_{\alpha/2}]^2 \frac{\hat{p}\hat{q}}{n} \cdot n$$

$$n \cdot E^2 = [Z_{\alpha/2}]^2 \hat{p}\hat{q}$$

Dividing by E^2 on both sides, you obtain the sample size formula:

$$n = \frac{(Z_{\alpha/2})^2 \cdot \hat{p}\hat{q}}{E^2}$$

This derivation only uses skills you learned in Elementary Algebra, and I hope you can try to derive it yourself to help understand it a little further.

There is one final twist: if we have no idea what sample proportion might be like, i.e. $\hat{p}$ is unknown, it's useful to remember from Intermediate algebra that, since $\hat{q} = 1 - \hat{p}$, then the largest value for $\hat{p}(1-\hat{p})$ is 0.25, since the function

$$f(x) = x(1-x)$$

reaches its maximum when $x = 0.50$. If we replace $\hat{p}\hat{q}$ with 0.5*0.5=0.25, the formula for n becomes:

$$n = \frac{(Z_{\alpha/2})^2 \cdot 0.25}{E^2}$$

From another perspective, this shows that when all else is equal (i.e. the same margin of error and confidence level, you need the largest sample when the proportion is around 50%.

As an example, imagine we need to survey the students to see how many percent of them are passionate about climate change, within a margin of error of 3% (the news outlets also use $\pm 3%$ in their report). Suppose we have never done this in the past, we should probably use the most conservative (in the sense of the largest) estimate for sample size. If we use the common confidence level of 90%, which corresponds to $ Z_{\alpha/2}=1.64$, this will give us (be careful to use 0.03, not 3 as E):

$$n = \frac{1.64^2 \cdot 0.25}{0.03^2} = 747.1$$

Obviously we can't survey the 0.1 person, so we should round the sample size to 748 to keep the margin of error **within** 3% (to keep the inequality in the right direction).

You can also use the formula in other ways. For example, why do political opinion surveys always include about 1500 people and a 3% of margin of error? You may have heard about this in the Political Science class, but probably not an explanation. Let's see what the confidence level they are using:

$$1500 = \frac{(Z_{\alpha/2})^2 \cdot 0.25}{0.03^2}$$

Solving for $Z_{\alpha/2}$ gives us

$$Z_{\alpha/2} = \sqrt{\frac{1500 \cdot 0.03^2}{0.25}} = 2.32$$

Using the GeoGebra normal calculator, we see that $\alpha / 2 = P(Z > 2.32) = 0.01$. Hence $\alpha = 0.02$. This tells us the Confidence Level is at 98%. I guess this looks better than what most people use, but it's really a convention that various polls try to adhere to.

Summary: Estimating Proportions

To briefly summarize the process of constructing a confidence interval for a population proportion, we have presented the following recipe:

1. Start from the sample proportion $\hat{p} = \frac{x}{n}$: this point estimate will be the mid-point of the interval.
2. Compute the critical value: find $Z_{\alpha/2}$, the critical value from the standard normal distribution that corresponds to $\alpha/2$ in the right tail. Here α = 1 – Confidence Level, and $\alpha/2$ is the area to the right of the critical value.
3. Compute the margin of error using either $E = Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$
4. Add / subtract the margin error from the sample mean to obtain the interval: $(\hat{p} - E, \hat{p} + E)$
5. State the confidence interval in a sentence such as "with (C.L.) confidence, (the population proportion) is between … and …."

Comparison between Confidence Interval and Hypothesis Testing

Now that we have seen two different ways of doing statistical inference ("what can I say about the parameter based on a sample/statistic?"): Confidence Interval (CI) and Hypothesis Testing (HT), it might be worth taking a moment reflecting on what they have in common, and how they differ.

Key Formulas in One-Sample Statistical Inference

The following chart summarizes the formulas that have appeared in the two chapters on Confidence Interval and Hypothesis Testing (also included in the lecture slides)

	Proportion p	Mean μ	
		σ is given	σ is unknown
Confidence Interval	$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$	$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$E = t_{\alpha/2}\frac{s}{\sqrt{n}}$
Hypothesis Testing	$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

In addition, the formulas for estimating the sample size are based on the formulas for finding margin of error, except we are solving for n instead of E.

Review of Central Limit Theorem

As you can see, the topics we studied went in parallel, and we essentially looked at each scenario twice, once from the perspective of Hypothesis Testing, and the second time from the perspective of Confidence Interval. For example, when we are doing inference on the population mean when σ is given, we used the same assumptions to carry out our inference:

- Central Limit Theorem: if we know the mean μ and the standard deviation σ of the population, then the sample mean $\bar{x}$ will follow a normal distribution, with mean given by: $\mu_{\bar{x}} = \mu$ And the standard deviation (also known as the standard error of $\bar{x}$) is given by: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

What this says is that if we know what is in the "pail" (parameter), then we can predict how our "hand" (statistic) will vary according to the normal distribution. Very powerful result indeed. Because without this, any inference we make is really no different from just another blind guess.

From CLT to Two-Tailed Hypothesis Test

Let's look at how this leads to the Hypothesis Test about μ: say we have a sample of 106 body temperatures with a sample mean of 98.20°F. Assume that the sample is a simple random sample and that the population standard deviation σ is known to be 0.62°F. Does this contradict the the common belief that the mean body temperature of healthy adults is equal to 98.6°F? Apparently, our sample mean is low compared to the popular belief. But is it low "enough" to be significant?

By now you probably have recognized that this is supposed to be a two-tailed test, with $H_0 : \mu = 98.6, H_a : \mu \neq 98.6$. Why do we need H_0? Because without some "ground truth"/assumption about what μ could be, how could we otherwise apply the Central Limit Theorem? The CLT allows us to condense the data AND H_0 into a single number, the test statistic, according to:

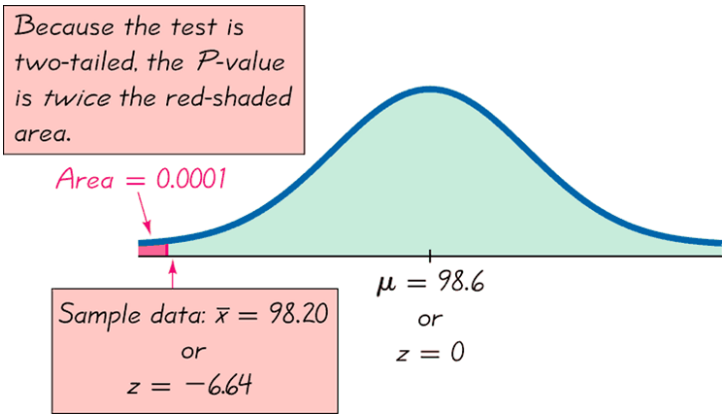
$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{98.2 - 98.6}{0.62 / \sqrt{106}} = -6.64$$

We can illustrate the logic of Hypothesis Testing as follows:

- Given $\bar{x}$ and μ (according to H_0) $\Rightarrow$ do the data contradict H_0? (in other words, support H_a)

How far Z is away from zero thus completely characterizes whether our data is likely to occur by chance, according to H_0. This greatly simplifies how decisions should be made in the face of data.

So when all is said and done, we have the following picture that tells us about the test statistic and P-value:



Notice that there are two distributions described by this picture: one is the normal distribution of $\bar{x}$, the other the distribution of the test statistic Z. Technically they have different means and scales, but the P-value identified by the $\bar{x}$ and Z was exactly the same. The distribution of Z allows us to make a decision relatively quickly, since we just need to compare P-value with α (or comparing the test statistic with the critical value); while the distribution of $\bar{x}$ gives us an intuitive interpretation of what P-value means: when μ is indeed equal to 98.6F, the chance of getting sample mean less than 98.2F or greater than 99.0F (remember the test is two-tailed in this case) is 0.0002, a very unlikely event. This result tell us there is something wrong with our initial assumption of $\mu = 98.6$, otherwise we wouldn't end up with a contradiction. So we reject H_0, and brought strong evidence to debunk the myth that people have average body temperature equal to 98.6F.

From CLT to Confidence Interval

Historically, Hypothesis Testing was the first tool of statistical inference that was created to answer questions like "does this growing method improve the yield of cotton?", or "are babies less likely to have a particular disease if we know the health of their parents?" Confidence Interval, on the other hand, was invented later, and it was used in situations where people don't necessary want a "yes" or "no" answer.

For example, suppose you are not on a mission to bust the myth about body temperature, but just curious about what might be the range of average body temperature of people. How would you use the Central Limit Theory and $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$?

Obviously, this time around, you don't have μ at all, instead, what you are looking for is quite different:

- Given $\bar{x}$ $\Rightarrow$ can I construct an interval to estimate μ ?

Here it's helpful to apply a little Algebra, because we can think of μ as the unknown variable to be solved, and the margin of error (E) can be defined as the maximum difference between the point estimate $\bar{x}$ and the actual parameter μ :

$$E = Z \cdot \frac{\sigma}{\sqrt{n}} = \bar{x} - \mu$$

If we want to keep the error within E, in algebra we can express this as an absolute value inequality:

$$|\bar{x} - \mu| < E$$

Solving this inequality gives us:

$$-E < \bar{x} - \mu < E$$

Subtracting $\bar{x}$, and multiplying (-1) gives us:

$$-\bar{x} - E < -\mu < -\bar{x} + E$$

$$\bar{x} - E < \mu < \bar{x} + E$$

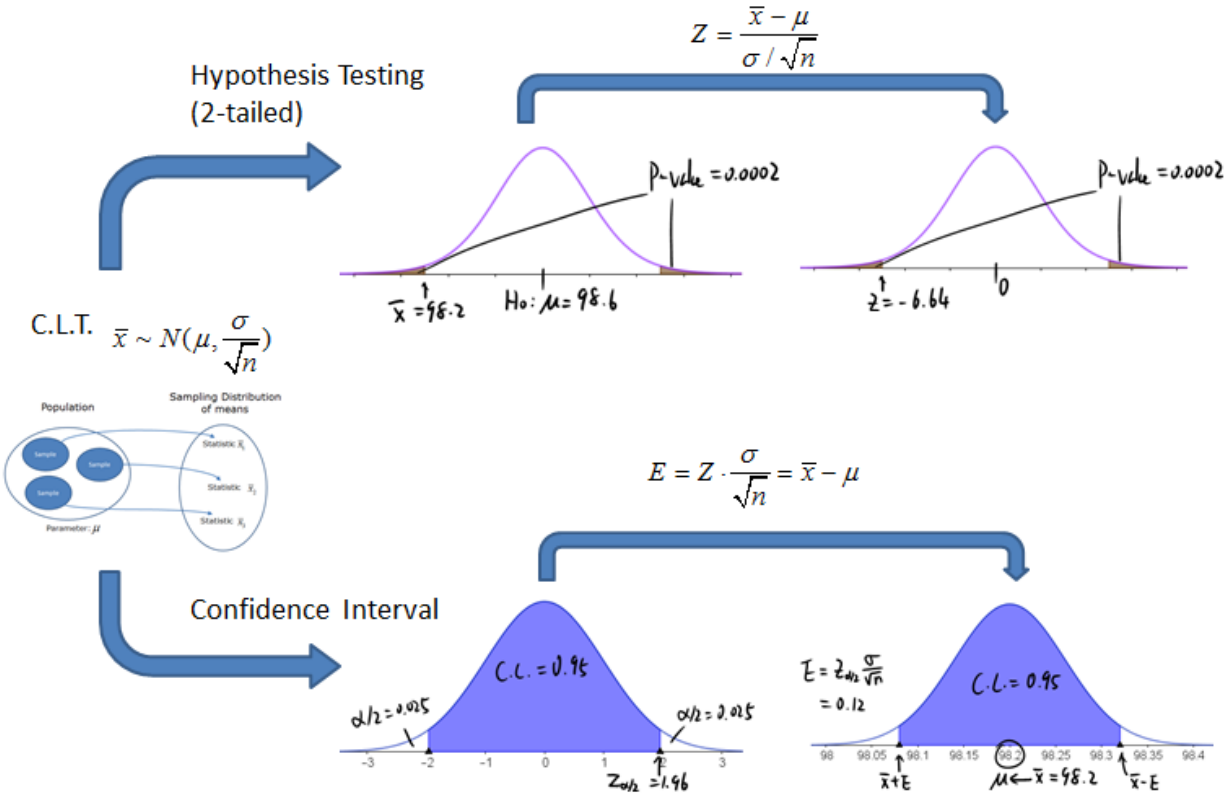
Where $E = Z \cdot \frac{\sigma}{\sqrt{n}}$

Here $\bar{x}$ and σ are given, but Z can vary. To produce an interval, we will just have to choose a confidence level and decide what ranges of Z are acceptable as "not too extreme" (see the section on confidence interval). Not too surprisingly, the critical value $Z_{\alpha/2} = 1.96$ is exactly the same as our two-tailed test, which leads to:

$$E = 1.96 \cdot \frac{0.62}{\sqrt{106}} = 0.12$$

Applying this in $\bar{x} - E$ and $\bar{x} + E$, we found our interval to be $98.08 < \mu < 98.32$.

Let's summarize the discussion above in the following graph, which provides an unified perspective on the two types of statistical inference that we just learned:



So here you have it: confidence interval and hypothesis testing are based on the same set of assumptions about how statistical inference is done: if we know how to predict the hand (statistic) from the pail (parameter), then we will use this knowledge to help us go backwards to predict the pail from the hand. The main difference between the two is that while HT was designed to be conclusive for the person who asks the question, CI leaves the interpretation to the person who receives the interval. As an interesting historical footnote, while HT was invented by British statisticians (Karl Pearson and Ronald Fisher, see my previous notes), CI was invented by the polish-born statistician Jerzy Neyman, who moved to the United States to work at UC Berkeley, which was a relatively unknown university in 1937. Neyman also helped found the Department of Statistics at Berkeley.

Team Homework #9: Confidence Intervals

Instructions: please show your work on separate sheets of paper and attach the cover sheet with your submission. The written homework must contain writings from all team members. Label your problem properly.

1. Complete the following simulation and answer the questions.
    a. From the class data, load the AGE data in GeoGebra. Regard this data set as the population and record its parameters below:

    μ = _____years, σ = _____ years

    b. Use a random number generator to select 10 data values randomly. Record their values below:

_____ _____ _____ _____ _____

_____ _____ _____ _____ _____

    c. Find the sample mean from these 10 data values:

    $\bar{x}$ = _____years

    d. Construct an 80% confidence interval and show all steps. Write your interval on the whiteboard when you are done.

    e. Compare all the confidence intervals on the board with the parameter. What is the meaning of "80% confident"?

2. A manufacturer of pharmaceutical products analyzes a specimen from each batch of a product to verify the concentration of the active ingredient. The chemical analysis is not perfectly precise. Repeated measurements on the **same** specimen give slightly different results. The results of repeated measurements are known to follow a normal distribution quite closely. The standard deviation of this distribution is known to be $\sigma = 0.0068$ grams per liter. Management asks the laboratory to produce results accurate to within ± 0.005 with 95% confidence. How many measurements must be averaged to comply with this request?

3. Answer the following questions based on the t-distribution:
    a. Compare the tail probability under the standard normal distribution and the t-distribution by using the following table.

	df = 3	df = 10	df = 100
$P(t > 2)$			
$P(Z > 2)$			

    b. Compare the critical values $t_{\alpha/2}$ (use df = 3) and $z_{\alpha/2}$ by using the following table (C.L. stands for confidence level)

	90% C.L.	95% C.L.	99% C.L.
$t_{\alpha/2}$			
$z_{\alpha/2}$			

4. Assuming that our class represent a simple random sample of all students at SRJC, use the class data to construct a 95% confidence interval for the average shoe size of female students at SRJC.

5. The level of various substances in the blood of kidney dialysis patients is of concern because kidney failure and dialysis can lead to nutritional problems. A researcher performed blood tests on several dialysis patients on six consecutive clinic visits. One variable measured was the level of phosphate in the blood. Phosphate levels for an individual tend to vary normally over time. The data on one patient, in milligrams of phosphate per deciliter (mg/dl) of blood are given below.

5.1	4.6	4.8	5.7	6.4	5.6

Give a 90% confidence interval for this patient's mean phosphate level.

6. How common is behavior that puts people at risk for AIDS? The National AIDS Behavioral Surveys interviewed a random sample of 2673 adult heterosexuals. Of these, 170 had more than one sexual partner in the past year. That's 6.36% of the sample. Based on these data, what can we say about the proportion of all adult heterosexuals who have multiple partners?
    a. Using the 99% confidence level, calculate the margin of error. Show your work.

    b. Give a 99% confidence interval for the true proportion p

7. Based on the class data, find a (i) 90% confidence interval, and (ii) 99% confidence interval for the true proportion of SRJC students whose favorite color is blue. Show your detailed work for the 90% confidence interval. What is the effect of increasing confidence level?

8. Gloria C. and Ronald F. are candidates for political office in a large city. You are planning a sample survey to determine what percent of the voters plan to vote for Gloria C. You want to estimate p with 95% confidence and a margin of error no greater than 3%, or 0.03.
    a) How large a sample do you need?

    b) How large a sample would you need for a 2.5% margin of error? For 2%?

9. (Bonus) Gallup tracks daily the percentage of Americans who approve or disapprove of the job Barack Obama is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults; Margin of error is ±3 percentage points. What is the confidence level used in Gallup's polls?

Math 15 Test #2 Study Guide (Chapter 7 – Chapter 9)

The test will be made of 4 sections: multiple-choice, brief-response, short calculations, and full-response problems. You are allowed to use a one-sided 8.5''x11'' sheet for notes.

The following team assignments should be reviewed before the test: Team Homework #6, #7, #8, #9, and Chapter Quizzes #7, #8, and #9.

Concepts and vocabulary:

- Given a parent distribution, how do you create a sampling distribution of means?
- When and how do you apply the Central Limit Theorem?
- Which sampling distributions have been applied in one-sample statistical inference?
- What is the meaning of confidence level (e.g. 95% confident), and how it is related to the sampling distribution?
- How does confidence level affect the interval estimate?
- How does significance level affect the outcome of hypothesis test?
- What is the relationship between sample size, confidence level and margin of error (or the accuracy of the interval estimate)?
- How are the interval estimate and point estimate related?
- Given a research question, what are the appropriate null and alternative hypotheses (specify the parameter and tail of the test)?
- What decisions can be made in a hypothesis test?
- What are the Type I and Type II errors in a particular hypothesis test?
- How do you find the P-values in one-tailed and two-tailed tests?
- How do you make decisions in a hypothesis test, based on the p-value method and traditional method?
- What are the implications of rejecting null / failure to reject the null hypothesis for the person conducting the hypothesis test?
- What is the interpretation of P-value in a particular hypothesis test?
- Why does a small P-value lead to the rejection of the null hypothesis?
- What is the difference between the t-distribution and standard normal distribution (Z)? How would this difference impact the critical value and the P-value?
- Which test statistics are appropriate for doing inference with which parameters?
- How is the concept of sampling distribution related to both Confidence Interval and Hypothesis Testing?

Practice problems from Team Homework:

(You may refer to individual team assignments for answers to these problems.)

1. Based on a population of SRJC students, describe how you would construct a sampling distribution of means for n = 20.

2. Using the Reese's pieces as your population, describe how you would construct a sampling distribution of the proportion of orange pieces, for n = 100.

3. A bottling company uses a filling machine to fill plastic bottles with cola. The bottles are supposed to contain 300 milliliters (ml). In fact, the contents vary according to a normal distribution with mean 298 ml and standard deviation 3 ml.
    a. What is the probability that an individual bottle contains less than 295 ml?
    b. What is the probability that the mean contents of the bottles in a six-pack is less than 295 ml?

4. Assume the passengers who use a particular elevator have normally distributed weights with a mean of 150 lb and a standard deviation of 35 lb. The elevator has a weight limit of 1600 lb and maximum carrying capacity of 10 passengers. What is the probability that the elevator exceeds its weight limit when it is filled to the capacity?

5. The wait time at the popular Russian River Brewing company follows a uniform distribution between 0 and 45 minutes. If 30 customers are randomly surveyed, what is the probability their average wait time is less than half an hour?

6. The level of various substances in the blood of kidney dialysis patients is of concern because kidney failure and dialysis can lead to nutritional problems. A researcher performed blood tests on several dialysis patients on six consecutive clinic visits. One variable measured was the level of phosphate in the blood. Phosphate levels for an individual tend to vary normally over time. The data on one patient, in milligrams of phosphate per deciliter (mg/dl) of blood are given below.

    5.1 4.6 4.8 5.7 6.4 5.6

    Give a 90% confidence interval for this patient's mean phosphate level.

7. Census Bureau data show that the 1250 out of 3540 families make less than $50,000 per year. A market research firm questions shoppers at the mall. The researchers suspect more than 30% of the families makes less than $50,000 per year. Test the claim using α=0.05.

8. How common is behavior that puts people at risk for AIDS? The National AIDS Behavioral Surveys interviewed a random sample of 2673 adult heterosexuals. Of these, 170 had more than one sexual partner in the past year. That's 6.36% of the sample. Based on these data, what can we say about the proportion of all adult heterosexuals who have multiple partners? Give a 99% confidence interval for the true proportion *p*

9. For each of the following scenarios, compute the P-value from the test statistic, and interpret P-value as a probability.
    a. Ted's Taxi Company bought 32 cars from the dealership that are advertised to get 28 miles per gallon in fuel efficiency. The company suspects that the average fuel efficiency per car is better than advertised. ($\bar{x} = 31.5mpg$, t = 2.01)
    b. In a multiple choice test, each question has 5 choices (ABCDE). A student guessed all 30 of the answers and missed all of them. He claims that he is doing worse than chance. (Z = -0.75)

10. The New York Times and CBS News conducted a nationwide poll of 1048 randomly selected 13- to 17-year olds. Of these teenagers, 692 had a TV in their room and 189 named FOX as their favorite network. Is there good evidence (at 0.01 level) that
    a. more than half of all teenagers have a TV in their room?
    b. Less than 20% of all teenagers named FOX as their favorite network?

11. On a NAEP assessment instrument, a score of 275 points or higher means a person can balance a checkbook. A random sample of 840 young men had a mean score of 272, and a standard deviation of 60 points. The hypothesis is that the mean for all young men is less than 275.
    a. What is the value of the test statistic?
    b. What is the P-value of this test?
    c. Is this sample good evidence, say at significance level 0.05 that the mean for all young men is less than 275?
    d. State the meaning of the P-value in terms of conditional probability, and explain why you reached the conclusion in (c) from the perspective of P-value.

12. Based on the class data (available in Canvas), find a (i) 90% confidence interval, and (ii) 99% confidence interval for the true proportion of SRJC students whose favorite color is blue.

13. According to the data released by National Shoe Retailers Association, the average American woman's shoe size is 8.5 (up from 7.5 a decade ago). Use the class data to test the hypothesis (at the 0.05 level) that the mean shoe size for female students taking Stats class is different from the national average.

14. Assuming that our class represent a simple random sample of all students at SRJC, use the class data to construct the following:
    a. A 90% confidence interval for the average height of male students at SRJC.
    b. A 95% confidence interval for the average shoe size of female students at SRJC.

15. Gloria C. and Ronald F. are candidates for political office in a large city. You are planning a sample survey to determine what percent of the voters plan to vote for Gloria C. This is a population proportion p. You will contact an SRS of registered voters in the city. You want to estimate p with 95% confidence and a margin of error no greater than 3%, or 0.03.
    a) How large a sample do you need?
    b) How large a sample would you need for a 2.5% margin of error? For 2%?

16. In a study of smokers who tried to quit smoking with nicotine patch therapy, 39 were no longer smoking one year after the treatment and 8 were still smoking one year after the treatment. Use a 0.10 significance level to test the claim that among smokers who try to quit with nicotine patch therapy, less than half are smoking a year after the treatment. Do these results suggest that the nicotine patch therapy is effective?

17. Bottles of cola are supposed to contain 300 ml of cola. There is some variation due to imprecision in filling machinery. An inspector suspects that the bottles are being overfilled and takes a sample of six bottles:

299.4, 297.9, 301.2, 301.8, 300.2, 302.1

Is this convincing evidence that the mean is deviating from 300 ml (assume α=0.10)?

18. In order to have enough staff on duty in an emergency room, a hospital emergency room wants to estimate the mean number of patients that are treated each day. In a random sample of 48 days, it is found that an average of 38 patients were seen each day in the emergency room with a standard deviation of 4 patients.
    a) Find a 99% confidence interval for the true mean number of patients that were seen in the emergency room per day. Write out the conclusion in complete sentences.
    b) What is meant when we say that we are 99% confident?
    c) Find the sample size that will give a margin of error no larger than 1 for the 99% confidence interval.

19. A manufacturer of pharmaceutical products analyzes a specimen from each batch of a product to verify the concentration of the active ingredient. The chemical analysis is not perfectly precise. Repeated measurements on the **same** specimen give slightly different results. The results of repeated measurements are known to follow a normal distribution quite closely. The laboratory analyzes each specimen three times and reports the mean result. Three analyses of one specimen give concentrations:
    0.8403, 0.8363, 0.8447.

    a. Give a 99% confidence interval for the true mean of the concentration, μ
    b. Now suppose the manufacturer is content with a 90% confidence interval. Will this increase or decrease the margin of error?
    c. Management asks the laboratory to produce results accurate to within $\pm$0.005 with 95% confidence. How many measurements must be averaged to comply with this request? Write down the formula, and then calculate.

20. The health of the bear population in Yellowstone National Park is monitored by periodic measurements taken from anesthetized bears. A sample of 54 bears has a mean weight of 182.9 lb, and a standard deviation of 121.8 lb, find a 99% confidence interval estimate of the mean of the population of all such bear weights.

21. A study was conducted to estimate hospital costs for accident victims who wore seat belts. Twenty randomly selected cases have a distribution that appears to be bell-shaped with a mean of $9004 and a standard deviation of $5629. Construct the 99% confidence interval for the mean of all such costs.

22. Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 tasters for one new cola recipe.

    2.0 0.4 0.7 2.0 -0.4

    2.2 -1.3 1.2 1.1 2.3

    Are these data good evidence (at the 10% level) that the cola lost sweetness?

23. In a paranormal experiment done in the 1970's, a subject is asked to guess which one of the four letters (ABCD) was written in the other side of the room. The subject made 20 correct guesses out of 50 trials. Researchers claimed that this subject has a success rate significantly different from chance. (Test the claim using α=0.05.)

24. A recent report claimed that 20% of all college graduates find a job in their field of study. A survey of a random sample of 500 graduates found that 110 obtained work in their field. At the significance level of 5%, answer the question: "Is there statistical evidence to refute the claim?"

25. A laboratory is asked to evaluate the claim that the concentration of active ingredient in a specimen is over 0.86 mg. The lab makes 3 repeated analyses of the specimen with mean result $\bar{x}$ = 0.87, and s = 0.007. Is there enough evidence (at the 1% level) that supports the claim?

26. The average power consumption of 25 randomly selected families in a community for a given period is 125.6 kilowatt-hours and the standard deviation is 20.3 kilowatt-hours. If we assume that kilowatt usage is normally distributed, is there evidence that the mean usage for the whole community less than 130 kilowatt-hours? Test your hypothesis at the significance level of 0.10.

27. Psychologists wish to investigate the learning ability of schizophrenic people after they have taken a specified dose of a tranquilizer. Thirteen patients were given the drug, and one hour later they were given an exam. Their scores are listed below:

    13 20 30 27 24 22 22 17 21 25 23 27 25

    a. Generally, patients score about 20 on the exam without the tranquilizer. Is there statistical evidence that taking a tranquilizer has affected their scores (α=0.05)?

    b. Construct a 95% confidence interval based on the same data. How does your conclusion compare with part (a)?