

Introduction to Statistical Inference

*Dr. Tom Pierce
Department of Psychology
Radford University*

What do you do when there's no way of knowing for sure what the right thing to do is? That's basically the problem that researchers are up against. I mean, think about it. Let's say you want to know whether older people are more introverted, on average, than younger people. To really answer the question you'd have to compare all younger people to all older people on a valid measure of introversion/extroversion – which is impossible! Nobody's got the time, the money, or the patience to test 30 million younger people and 30 million older people. So what do you do? Obviously, you do the best you can with what you've got. And what researchers can reasonably get their hands on are samples of people. In my research I might compare the data from 24 older people to the data from 24 younger people. And the cold hard truth is that when I try to say that what I've learned about those 24 older people also applies to all older adults in the population, I might be wrong. As we said in the chapter on descriptive statistics, samples don't have to give you perfect information about populations. If, on the basis of my data, I say there's no effect of age on introversion/extroversion, I could be wrong. If I conclude that older people are different from younger people on introversion/extroversion, I could still be wrong! Looking at it this way, it's hard to see how anybody learns anything about people.

The answer is that behavioral scientists have learned to live with the fact that they can't "prove" anything or get at "the truth" about anything. You can never be sure whether you're wrong or not. But, there is something you *can* know for sure. Statisticians can tell us exactly what the **odds** are of being wrong when we draw a particular conclusion on the basis of our data. The means that you might never know for sure that older people are more introverted than younger people, but your data might tell you that you can be **very confident** of being right if you draw this conclusion. For example, if you know that the odds are like one in a thousand of making a mistake if you say there's an age difference in introversion/extroversion, you probably wouldn't lose too much sleep over drawing this conclusion.

This is basically the way data analysis works. There's never a way of knowing for sure that you made the right decision, but you can know exactly what the odds are of being wrong. We can then use these odds to guide our decision making. For example, I can say that I'm just not going to believe something if there's more than a 5% chance that I'm going to be wrong. The odds give me something concrete to go on in deciding how confident I can be that the data support a particular conclusion. When a person uses the odds of being right or wrong to guide their decision making they're using statistical inference.

Statistical inference is one of the most powerful tools in science. Practically every conclusion that behavioral scientists draw is based on the application of a few pretty

simple ideas. Once you get used to them – and they do take some getting used to – you’ll see that these ideas can be applied to practically any situation where researchers want to predict and explain the behavior of the people they’re interested in. All of the tests we’ll talk about – t-tests, Analysis of Variance, the significance of correlation coefficients, etc – are based on a common strategy for deciding whether the results came out the way they did by chance or not. Understanding statistical inference is just a process of recognizing this common strategy and learning to apply it to different situations.

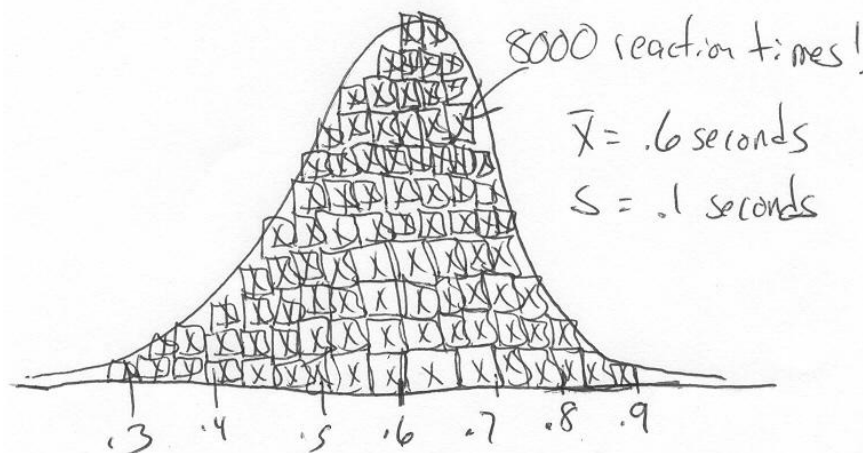
Fortunately, it’s a lot easier to give you an example of statistical inference that it is to define it. The example deals with a decision made that a researcher might make about a bunch of raw scores – which you’re already familiar with. Spend some time thinking your way through this next section. If you’re like most people, it takes hearing it a couple of times before it makes perfect sense. Then you’ll look back and wonder what the fuss was all about. Basically, if you’re okay with the way statistical inference works in this chapter, you’ll understand how statistical inference works in every chapter to follow.

An example of a statistical inference using raw scores

The first thing I’d like to do is to give you an example of a decision that one might make using statistical inference. I like this example because it gives us the flavor of what making a statistical decision is like without having to deal with any real math at all.

One variable that I use in a lot of my studies is reaction time. We might typically have 20 younger adults that do a reaction time task and 20 younger adults that do the same task. Let’s say the task is a choice reaction time task where the participants are instructed to press one button if a stimulus on a computer screen is a digit and another button if the stimulus is a letter. This task might have 400 reaction time trials. From my set of older adults I’m going to have 400 trials from each of 20 participants. That’s 8000 reaction times from this group of people. Now, let’s say for the sake of argument that this collection of 8000 reaction times is normally distributed. The mean reaction time in the set is .6 seconds and the standard deviation of reaction times is .1 seconds. A graph of this hypothetical distribution is presented in Figure 3.1.

Figure 3.1



One problem that I run into is that the reaction times for three or four trials out of the 8000 trials are up around 1.6 seconds. ***The question I need to answer is whether to leave these reaction times in the data set or to throw them out.*** They are obviously **outliers** in that these are scores that are clearly different from almost all the other scores, so maybe I'm justified in throwing them out.. However, data is data. Maybe this is just the best that these subjects could do on these particular trials; so, to be fair, maybe I should leave them in.

One thing to remember is that the instructions I gave people were to press the button on each trial as fast as they could while making as few errors as they could. This means that when I get the data, I only want to include the reaction times for trials when this is what was happening – when people were doing the best they could – when nothing went wrong that might have gotten in the way of their doing their best. So now, I've got a reaction time out there at 1.6 seconds and I have to decide between two options, which are:

1. The reaction time of 1.6 seconds belongs in the data set because this is a trial where nothing went wrong. It's a reaction time where the person was doing the task the way I assumed they were. Option 1 is to keep the RT of 1.6 seconds in the data set. What we're really saying is that ***the reaction time in question is really a member of the collection of 8000 other reaction times that makes up the normal curve.***

Alternatively...

2. The reaction time does not belong in the data set because this was a trial where the subject wasn't doing the task the way I assumed that they were. Option 2 is to throw it out. What we're saying here is that ***the RT of 1.6 seconds does NOT belong with the other RTs in the set This means that the RT of 1.6 seconds must belong to some other set of RTs – a set of RTs where the mean of that set is quite a bit higher than .6 seconds.***

In statistical jargon, Option 1 is called the **null hypothesis**. The null hypothesis says that ***our one event only differs from the mean of all the other events by chance.*** If the null hypothesis is really true, this says there was no reason or cause for the reaction time on this trial to be this slow. It just happened by accident. The symbol HO is often used to represent the null hypothesis.

In statistical jargon, the name for Option 2 is called the **alternative hypothesis**. The alternative hypothesis says that ***our event didn't just differ from the mean of the other events by chance or by accident – it happened for a reason.*** Something caused that reaction time to be a lot slower than most of the other ones. We may not know exactly what that reason is, but we can be pretty confident that SOMETHING happened to give us a really slow reaction time on that trial -- the event didn't just happen by accident. The alternative hypothesis is often symbolized as H1.

Now, of course, there is no way for both the null hypothesis and the alternative hypothesis to both be true at the same time. We have to pick one or the other. But there's no information available to help us to know for sure which option is correct. This is something that we've just got to learn to live with. Psychological research is never able to prove anything, or figure out whether an idea is true or not. We never get to know for sure whether the null hypothesis is true or not. There is nothing in the data that can tell us for sure whether that RT of 1.6 seconds really belongs in our data set or not. It is certainly possible that someone could have a reaction time of 1.6 seconds just by accident. There's no way of telling for sure what the right answer is. So we're just going to have to do the best we can with what we've got. We've got to accept the fact that whichever option we pick, we could be wrong.

The choice between Options 1 and 2 basically comes down to whether we're willing to believe that we could have gotten a reaction time of 1.6 seconds just by chance. If the RT was obtained just by chance, then it belongs with the rest of the RTs in the distribution, and we should decide to keep it. If there's any reason other than chance for how we could have ended up with a reaction time that slow -- if there was something going on besides the conditions that I had in mind for my experiment, then the RT wasn't obtained under the same conditions as the other RTs -- and I should decide to throw it out.

So what do we have to go on in deciding between the two options? Well, it turns out that the scores in the data set are normally distributed. And we know something about the normal curve. We can use the normal curve to tell us exactly what the odds are of getting a reaction time this much slower than the mean reaction time of .6 seconds.

For starters, if you convert the RT of 1.6 seconds to a standard score, what do you get? Obviously, if we convert the original raw score (a value of X) to a standard score (a value of Z), we get...

$$Z_x = \frac{X - \bar{X}}{S} = \frac{1.6 - .6}{.1} = \frac{1.0}{.1} = 10.0$$

...a value of 10.0. The reaction time we're making our decision about is 10.0 standard deviations above the mean. That seems like a lot! The symbol Z_x translates to "the standard score for a raw score for variable X".

So what does this tell us about the odds of getting a reaction that far away from the mean just by chance alone? Well, you know that roughly 95% of all the reaction times in the set will fall between the standard scores of -2 and $+2$. Ninety-nine percent will fall between -3 and $+3$. So automatically, we know that the odds of getting a reaction time with a standard score of $+3$ or higher must be less than 1%. And our reaction time is ten standard deviations above the mean. If the normal curve table went out far enough it would show us that the odds of getting a reaction time with a standard score of 10.0 is something like one in a million!

Our knowledge of the normal curve combined with our knowledge of where our raw score falls on the normal curve gives us something solid to go on when making our decision. We know that the odds are something like one in a million that our reaction time belongs in the data set.

What would the odds have to be to make you believe that the score doesn't belong in the data set? An **alpha level** is a set of odds that the investigator decides to use when deciding whether one event belongs with a set of other events. For example, an investigator might decide that they're just not willing to believe that a reaction time really belongs in the set (i.e., it's different from the mean RT in the set just by chance) if the odds of this happening are less than 5%. If the investigator can show that the odds of getting a certain reaction time are less than 5% then it's different enough from the mean for them to bet that they didn't get that reaction time just by chance. It's different enough for them to bet that the reaction time must have been obtained when the null hypothesis was false.

So how far away from the center of the normal curve does a score have to be before it's in the 5% of the curve where a score is least likely to be? In other words, how far above or below the mean does a score have to be before it fits the rule the investigator has set up for knowing when to reject the null hypothesis?

So far, our decision rule for knowing when to reject the null hypothesis is: Reject the null hypothesis when the odds that it's true are less than 5%. Our knowledge of the normal curve gives us a way translating a decision rule stated in terms of odds into a decision rule that's expressed in terms of the scores that we're dealing with. What we'd like to do is to throw reaction times out if they look a lot different from the rest of them. One thing that our knowledge of the normal curve allows us to do is to express our decision rule in standard score units. For example, if the decision rule for rejecting the null hypothesis is that we should "reject the null hypothesis when the odds of its being true are 5% or less", this means that we should reject the null hypothesis whenever a score falls in the outer 5% of the normal curve. In other words, we need to identify the 5% of scores that one is least likely to get when the null hypothesis is true. ***How many standard deviations away from the center of the curve to we have to go until we get to the start of this outer 5%?***

In standard score units, you have to go 1.96 standard deviations above the mean and 1.96 standard deviations below the mean to get to the start of the extreme 5% of values that make up the normal curve. So, ***if the standard score for a reaction time is above a positive 1.96 or is below a negative 1.96, the reaction time falls in the 5% of the curve where you're least likely to get reaction times just by chance.***

The decision rule for this situation becomes: Reject H_0 if $Z_x \geq +1.96$ or if $Z_x \leq -1.96$. The reaction time in question is 10.0 so the decision would be to reject the null hypothesis. The conclusion is that the reaction time does not belong with the other reaction times in the data set and should be thrown out.

The important thing about this example is that it boils down to a situation where one event (a raw score in this case) is being compared to a bunch of other events that occurred when the null hypothesis was true. If it looks like the event in question could belong with this set, we can't say we have enough evidence to reject the null hypothesis. If it looks like the event doesn't belong with a set of other events collected when the null hypothesis was true, that means we're willing to bet that it must have been collected when the null hypothesis is false. You could think of it this way: ***every reaction time deserves a good home. Does our reaction time belong with this family of other reaction times or not?***

The Z-Test

The example in the last section was one where we were comparing one raw score to a bunch of other raw scores. Now let's try something a little different.

Let's say you've been trained in graduate school to administer the I.Q. test. You get hired by a school system to do the testing for that school district. On your first day at work the principal calls you into their office and tells you that they'd like you to administer the I.Q. test to all 25 seventh graders in a classroom. The principal then says that all you have to do is to answer one simple question. ***Are the students in that classroom typical/average seventh graders or not?***

Now, before you start. What would you expect the I.Q. scores in this set to look like? The I.Q. test is set up so that the mean I.Q. for all of the scores from the population is 100 and the standard deviation of all the I.Q. scores for the population is 15. So, if you were testing a sample of seventh graders from the general population, you'd expect the mean to be 100.

Now, let's say that you test all 25 students. You get their I.Q. scores and you find that the mean for this group of 25 seventh graders is 135. 135! Do you think that these were typical/average seventh graders or not. Given what you know about I.Q. scores, you probably don't. But why not. What if the mean had turned out to be 106? Are these typical/average seventh graders. Probably. How about if the mean were 112? Or 118? Or 124? At what point do you change your mind from "yes, they were typical/average seventh graders" to "no, they're not". What do you have to go on in deciding where this cutoff point ought to be? At this point in our discussion your decision is being made at the level of intuition. But this intuition is informed by something very important. It's informed by your sense of ***the odds of getting the results that you did***. Is it believable that you could have gotten a mean of 135 when the mean of the population is 100? It seems like the odds are pretty low that this could have happened.

Whether you've realized it or not, your decisions in situations like these are based on odds of what really happened. In an informal way, you were making a decision using statistical inference. Tools like t-tests work in exactly the same way. The only thing that makes them different is the degree of precision involved in knowing the relevant odds. Instead of knowing that it was ***pretty unlikely*** that you'd tested a group of typical average

seventh graders, a tool like a t-test can tell you *exactly how unlikely it is* that you tested a group of typical/average seventh graders.

Just like in the example with the reaction time presented above, the first step in the decision process is in defining the two choices that you have to pick between: the null and alternative hypotheses. In general, the null hypothesis is that the things being compared are just different from each other by accident or that the difference is just due to chance. There was no reason for the difference, really. It just happened by accident. In this case the null hypothesis would be that the mean of the sample of 25 seventh graders and the population mean of 100 are just different from each other by accident.

The alternative hypothesis is the logical opposite of this. The alternative hypothesis is that there is something going on other than chance that's making the two means different from each other. It's not an accident. The means are different from each other for a reason.

So how do you pick between the null and the alternative hypotheses? Just like in the example with the reaction times, it turns out that the only thing we can know for sure are the odds of the null hypothesis being true. We have to decide just how unlikely the null hypothesis would have to be before we're just not willing to believe that's it's true anymore. Let's say you decide that if we can show that the odds are less than 5% that the null hypothesis is true, you'll decide that you just can't believe it anymore. When you decide to use these odds of 5%, this means that you've decided to use an alpha level of .05.

So how do you figure out whether or not the odds are less than 5% that the null hypothesis is true? The place to start is by remembering where the data came from. They came from a **sample** of 25 students. There's an important distinction in data analysis between a sample and a population. ***A population is every member of the set of people, animals, or things. etc., that you want to draw a conclusion about.*** If you only wanted to draw a conclusion about the students that attend a certain school, the students that go to that school make up the population of people we're interested in. If you're interested in drawing a conclusion about all older adults in the United States then you might define this population as every person in the U.S. at or above the age of 65. ***A sample is a representative subset of the population you're interested in.*** A sample often consists of just a tiny portion of the whole population. The assumption is that the people that make up the sample contain roughly the same characteristics as the characteristics seen in the whole population. If the population is 65% female and 35% male, then the sample should be 65% female and 35% male. The sample should look like the population. If it doesn't, the sample may not be a representative subset of the population. The whole point of using a sample is that you can use the information you get from a sample to tell you about what's going on in the whole population. Investigators do whatever they can to make sure that their samples are unbiased – that is, the samples don't give a distorted picture of what the people that make up the population look like.

Samples are supposed to tell you about populations. This means that the numbers you get that describe the sample are intended to describe the population. ***The descriptive statistics you get from samples are assumed to be unbiased estimates of the numbers you'd get if you tested everyone in the population.*** Let's look at that phrase – unbiased estimate. The estimate part comes from the fact that every time you calculate a descriptive statistic from a sample, it's supposed to give you an estimate of the number for everyone in the whole population. By unbiased, we mean that the descriptive statistic you get from a sample has an equal chance of being too high or too low. No one assumes that estimates have to be perfect. But they're not supposed to be systematically too high or too low.

So a sample mean has only one job – to give you an unbiased estimate of the mean of a population. In the situation presented above we've got the mean of a sample (135) and we're using it to decide whether a group of 25 seventh graders are members of a population of typical, average seventh graders. The mean of the population of all typical average seventh graders is 100. So the problem basically comes down to a yes or no decision. ***Is it reasonable to think that we could have a sample mean of 135 when the mean of the population is 100? Just how likely is it that we were sampling from the population of typical, average seventh graders and ended up with a sample mean of 135, JUST BY CHANCE?*** If the odds aren't very good of this happening then you might reasonable decide that your sample mean *wasn't* an estimate of this population mean. Which says that you're willing to bet that the children in that sample aren't members of that particular population.

This is exactly the kind of question that we were dealing with in the reaction time example. We started out with one reaction time (a raw score) and we decided that if we could determine that the odds were less than five percent that this one raw score belonged with a collection of other reaction times (a group of other raw scores), we'd bet that it wasn't a member of this set. The basic strategy in the reaction time example was to compare one event (a raw score) to a bunch of other events (a bunch of other raw scores) to see if it belonged with that collection or not. If the odds were less than five percent that it belonged in this collection our decision would be to reject the null hypothesis (i.e., the event belongs in the set) and accept the alternative hypothesis (i.e., the event does not belong in the set).

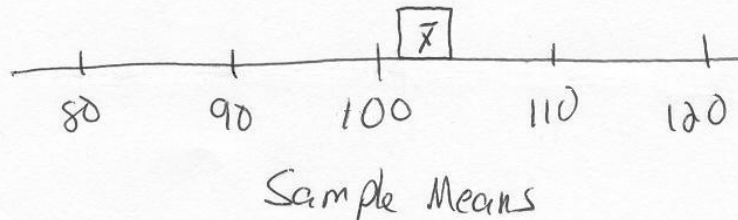
So how do we extend that strategy to this new situation? How do we know whether the odds are less than 5% that the null hypothesis is true? The place to start is to recognize that the event we're making our decision about here is a sample mean, not a raw score. Before, we compared one raw score to our collection of other raw scores. If we're going to use the same strategy, ***it seems like we would have to compare one sample mean to a bunch of other sample means***, and that's exactly how it works. BUT WAIT A MINUTE, I've only got just the one sample mean. Where do I get the other ones? That's a good question, but it turns out that we don't really need to collect a whole bunch of other sample means – sample means that are all estimates of that population mean of 100. What the statisticians tell us is that because we know the mean and standard deviation of the raw scores in the population ***we can imagine what the sample means would look like***

if we kept drawing one sample after another one and every sample had 25 students in it.

For example, let's say that you could know for sure that the null hypothesis was true – that the 25 students in a particular class were drawn from the population of typical, average students. What would you expect the mean I.Q. score for this sample to be. Well, if they're typical average students, and the mean of the population of typical, average seventh graders is 100, then you'd **expect** that the mean of the sample will be 100. And, in fact, that's the single most likely thing that would happen. But does that sample mean have to be 100? If the null hypothesis is true and you've got a sample of 25 typical, average seventh graders, does the sample mean have to come out to 100? Well, that sample mean is just an estimate. It's an unbiased estimate so it's equally likely to be greater than or less than the number it's trying to estimate, but it's still just an estimate. And estimates don't have to be perfect. So the answer is no. The mean of this sample doesn't necessarily have to be equal to 100 when the null hypothesis is true. The sample mean could be (and probably will be) at least a little bit different from 100 just by accident – just by chance.

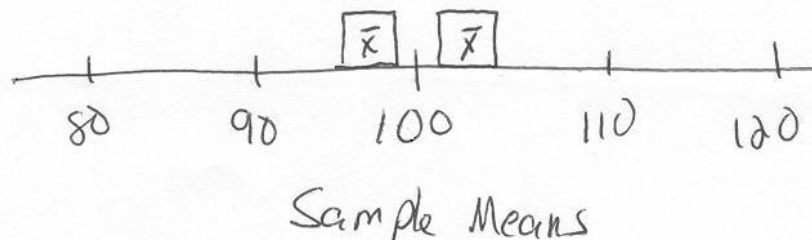
So let's say that, hypothetically, you can know for sure that the null hypothesis is true and you go into a classroom of 25 typical average seventh graders and obtain the mean I.Q. score for that sample. Let's say that it's 104. We can put this sample mean where it belongs on a scale of possible sample means. See Figure 3.2.

Figure 3.2



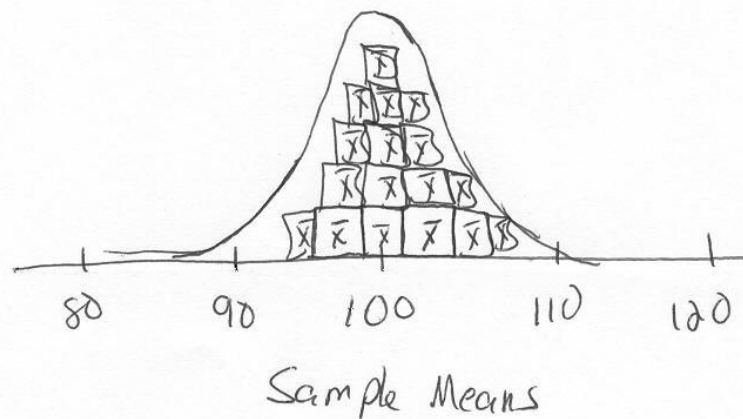
Now, hypothetically, let's say that you go into a second classroom of typical, average seventh graders. The null hypothesis is true -- the students were drawn from a population that has a mean I.Q. score of 100 – but the mean for this second sample of students is 97. Now we've got two estimates of the same population mean. One was four points too high. The other was three points too low. The locations of these two sample means are displayed in Figure 3.3.

Figure 3.3



Now, let's say that you go into fifteen more classrooms. Each classroom is made up of 25 typical, average seventh graders. From each of these samples of people you collect one number – the mean of that sample. Now we can see where these sample means fall on the scale of possible sample means. In Figure 2.4 each sample mean is represented by a box. When there's another sample mean that shows up at the same point on the scale (i.e., the same sample mean) we just stack that box up on top of the one that we got before. The stack of boxes presented in Figure 3.4 represents the frequency distribution of the 17 sample means that we've currently got. The shape of this frequency distribution looks more or less like the normal curve.

Figure 3.4



Now, let's say that, hypothetically, you went into classroom after classroom after classroom. Every classroom has 25 typical, average seventh graders and from every classroom you obtain the mean I.Q. score for this sample of students. If you were to get means from hundreds of these classrooms – thousands of these classrooms – and then put each of these sample means where they belong on the scale, the shape of this distribution of numbers would look exactly like the normal curve. The center of this distribution is 100. The average of all of the sample means that make up this collection is 100. This sort of makes sense because every one of those sample means was an estimate of that one population mean of 100. Those sample means might not all have been perfect estimates, but they were all unbiased estimates. Half of those estimates were too high, half of them were too low, but the average of all of these sample means is exactly equal to 100. ***All of these numbers – these sample means – were obtained under a single set of conditions: when the null hypothesis is true.***

Now we've got a collection of other sample means to compare our one sample mean to. The sample means in this collection show us how far our estimates of the population mean of 100 can be off by just by chance. This collection of sample means is referred to as **the sampling distribution of the mean**. It's the distribution of a bunch of sample means collected when the null hypothesis is true – when all of the sample means that make up that set are estimates of the one population mean we already know about – in this case, 100.

So how does the sampling distribution of the mean help us to make our decision? Well, the fact that the shape of this distribution is normal makes the situation exactly like the one we had before, when we were trying to decide whether or not to toss a raw score out of a data set of other raw scores.

If you remember...

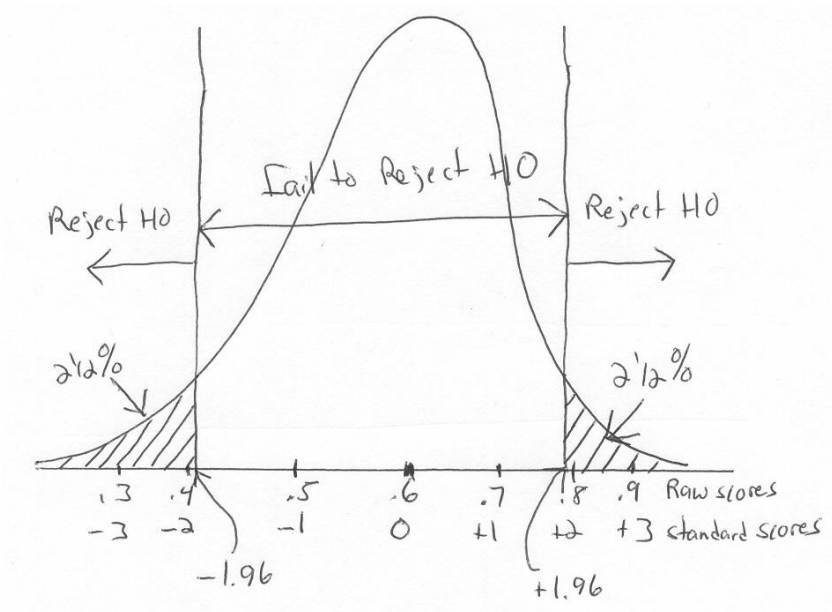
- We made a decision about a single number – in this case a single raw score. The decision was about whether that raw score was collected when the null hypothesis was true or whether the null hypothesis was false.
- The only thing we had to go on were the odds that the raw score was obtained when the null hypothesis was true. We had these odds to work with for two reasons. First, we had a collection of other raw scores to compare it to – raw scores that were all obtained under one set of conditions – when the null hypothesis is true. Second, the shape of the frequency distribution for these raw scores was normal.
- We said that unless we had reason to think otherwise, we'd just have to go with the null hypothesis that the raw score really did belong in the collection. We would only decide to reject the idea that the reaction time belonged in the set when we could show that the odds were less than 5% that it belonged in that set.
- We said that if we converted all of the raw scores in the collection to standard scores we could use the normal curve to determine how far above or below a standard score of zero you'd have to go until you hit the start of the extreme 5% of this distribution. In other words, ***how far above or below zero do you have to get to until you hit the start of the least likely 5% of reaction times that really do belong in that collection.***
- The decision rule for knowing when to reject the null hypothesis became: Reject the null hypothesis if Z_X is greater than or equal to +1.96 or if Z_X is less than or equal to -1.96.
- The only thing left was to take that raw score and convert it to a standard score.

Now we've got exactly the same kind of situation...

- We're making a decision about a single number. The only difference is that now this number is the mean of a sample of raw scores, rather than a single raw score. The decision is whether that number was collected when a null hypothesis was true or when it was false.
- There's no way of knowing for sure whether this sample mean was collected when the null hypothesis was true or not. The only thing that we can know for sure are the odds that the sample mean was collected when the null hypothesis was true. We can know these odds because we have a collection of other sample means to compare our one sample mean to. These sample means were all collected under the same circumstances – when the null hypothesis was true.
- Unless we have reason to think otherwise we'll have to assume that our sample mean was collected when the null hypothesis was true – when the mean of the sample really was an estimate of that population mean of 100.

- Specifically, we can decide to reject the null hypothesis only if we can show that the odds are less than 5% that it's true. We can decide to only reject the idea that the sample mean belongs in the set when we can show that the odds are less than 5% that it belongs in our hypothetical collection of other sample means.
- If we imagine that we're able to convert all of the sample means that make up our normal curve to standard scores, we can use our knowledge of the normal curve to determine how far above or below a standard score of zero you'd have to go until you hit the start of the most extreme 5% of this distribution. In other words, how far above or below zero do you have to go until you hit the start of the least likely 5% of sample means that you're likely to get when the null hypothesis is true.
- If we take our one sample mean and convert it to a standard score our knowledge of the normal now tells us that ***if this standard score is greater than or equal to +1.96 or if this standard score is less than or equal to -1.96 we'll know that our sample mean falls among the 5% of sample means that you're least likely to get when the null hypothesis is true.*** We would know that the odds are less than 5% that our 25 seventh graders (remember them) were members of a population of typical, average seventh graders. See Figure 3.5.
- The only thing left at this point is to convert our sample mean to a standard score.

Figure 3.5



To convert a number to a standard score you take that number, subtract the mean of all the other numbers in the set, and then divide this deviation score by the standard deviation of all the scores in the set. A standard score is the deviation of one number from the mean divided by the average amount that numbers deviate from their mean. The equation to convert a raw score to a standard score was...

$$Z_X = \frac{X - \bar{X}}{\sigma}$$

It's the same thing to take a sample mean and convert it to a standard score. We need to take the number that we're converting to a standard score (our sample mean of 135), divide it by the average of all the sample means in the set, and then divide this number by the standard deviation of all the sample mean in the set. The equation for the standard score we need here becomes...

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$Z_{\bar{X}}$ represents the standard score for a particular sample mean. M is the sample mean that is being converted to a standard score. μ represents the mean of the population (or the average of all the sample means). $\sigma_{\bar{X}}$ represents the standard deviation of all the sample means.

So what's the average of all the sample means that you could collect when the null hypothesis is true? Well, you know that every one of those sample means was an estimate of the same population mean. They were all estimates of the population mean of 100. If we assume that these sample means are unbiased estimates then ***half of those estimates end up being less than 100 and half of those estimates end up being greater than 100, so the mean of all those estimates – all those sample means – is 100!*** The average of all of the values that make up the sampling distribution of the mean is also the mean of the population. Okay, that was easy.

Now, how about the standard deviation of all of the sample means. Well, we've got a collection of numbers and we know the mean of all of those numbers. So we should be able to calculate the average amount that those numbers deviate from their mean. Unfortunately, the sampling distribution of the mean contains the means of, hypothetically, every possible sample that has a particular number of people in it. So doing that is kind of out. But this is where those spunky little statisticians come in handy. Some smart person – probably on a Friday night when everyone else was out having fun – nailed down the idea that the standard deviation of a bunch of sample means is influenced by two things: (1) the standard deviation of the raw scores for the population mean and (2) the number of people that make up each individual sample. ***The Central Limit Theorem tells us that to calculate the standard deviation of the sample means you take the standard deviation of the raw scores in the population (sigma) and then divide it by the square root of the sample size.*** The equation for calculating the standard deviation of the sample means ($\sigma_{\bar{X}}$) becomes...

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

The symbol $\sigma_{\bar{x}}$ simply reflects the fact that we need *the standard deviation of a bunch of sample means*. The term for this particular standard deviation is the **Standard Error of the Mean**. One way of thinking about it is to say that it's the average or "standard" amount of sampling error you get when you're using the scores from a sample to give an estimate of what's going on with everyone in the population.

- From this equation it's easy to see that if the standard deviation of all the raw scores in the population is larger (if the raw scores are more spread out around their mean) the more spread out the sample means get.
- Also, the more people you have in each sample, the less spread out the sample means will be around their average. This makes sense because the sample means are just estimates. The more scores that contribute to each estimate, the more accurate they ought to be – the closer they ought to be on average to the mean of the population.

So, in our example, the standard error of the mean becomes...

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3.0$$

With a standard deviation of the raw scores of 15 and sample sizes of 25, the average amount that sample means are spread out around the number they're trying to estimate is 3.0.

So now we've got everything we need to convert our sample mean to a standard score.

$$Z_{\bar{x}} = \frac{135 - 100}{3} = \frac{35}{3} = 11.67$$

Our sample mean is 11.67 standard deviations above the population mean of 100. Our decision rule already told us that we'd be willing to reject the null hypothesis if the sample mean has a standard score that is greater than or equal to +1.96. So our decision is to reject the null hypothesis. This means that we're willing to accept the alternative hypothesis, so our conclusion for this decision is that "The seventh graders in this classroom are not typical/average seventh graders". That's a lot to think about to get to say one sentence!

Directional (one-tailed) versus non-directional (two-tailed) tests

Now let's say that we change the question a little bit. Instead of asking whether the 25 kids in the class are typical/average seventh graders, let's say that researcher wants to know whether this is a class of gifted and talented seventh graders. In other words, instead of asking whether the mean I.Q. of the sample is significantly *different from* the population mean of 100, we're now asking whether the mean of the sample is significantly *greater than* the population mean of 100.

How does this change the problem? It doesn't change anything about the number crunching. The standard error of the mean is still 3.0. The sample mean of 135 is still 11.667 standard deviations above the population mean of 100. However, the conclusion the researcher can draw on the basis of this standard score is going to change because the research question has changed. One way to think about this is to consider that the only reason for getting our value for Z is help us to decide between two statements: the null hypothesis and the alternative hypothesis. You're doing the statistical test to see which of the two statement you're going to be willing to believe. The alternative hypothesis is the prediction made by the researcher. The null hypothesis is the opposite is this prediction. So, obviously, if you change the prediction you change both the null and alternative hypotheses.

In the example in the previous section, the prediction was that the mean of the sample would be significantly different from the mean of the population, so that was the alternative hypothesis. The null hypothesis was the logical opposite of the alternative hypothesis: the mean of the sample of 25 seventh graders is not significantly different from the population mean of 100. This is said to be a **non-directional prediction** because the statement could be true no matter whether the sample mean was a lot larger or a lot smaller than the population mean of 100.

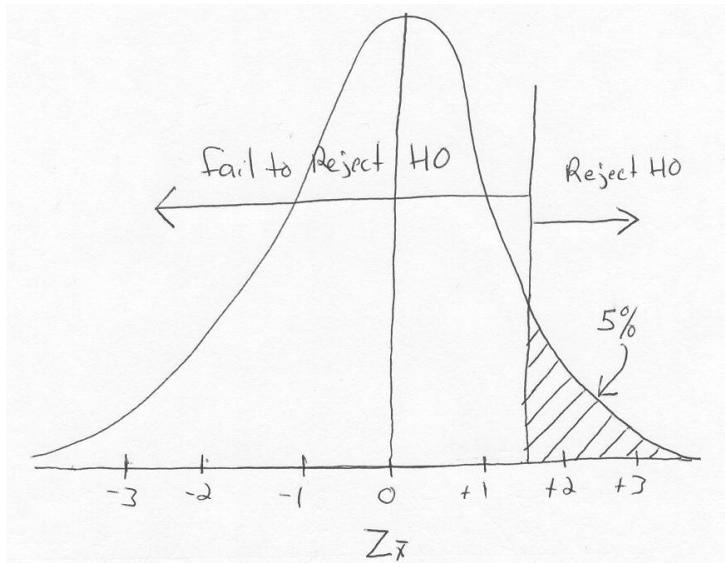
In the current example, the prediction is that the mean of the sample is significantly greater than the mean of the population. From this it follows that the alternative hypothesis is that the mean of the sample of 25 seventh graders is significantly greater than the population mean of 100. The null hypothesis is the logical opposite of this statement: the mean of the sample of 25 seventh graders is not significantly greater than the population mean of 100. The researcher is said to have made a **directional prediction** because they're being specific about whether they think the sample mean will be above or below the mean of the population. The null and alternative hypotheses for the directional version of the test are stated below:

H₀: The mean of the sample of 25 seventh graders is not significantly greater than the population mean of 100.

H₁: The mean of the sample of 25 seventh graders is significantly greater than the population mean of 100.

Okay, so if the null and alternative hypotheses change, how does this change the decision rule that tells when we're in a position to reject the null hypothesis? Do we have to change the alpha level? No. We can still use an alpha level of .05. We can still say that we're not going to be willing to reject the null hypothesis unless we can show that there's less than a 5% chance that it's true. Do we have to change the critical values? Yes, you do. And here's why. Think of it this way. With the non-directional version of the test we were placing a bet – “don't reject the null hypothesis unless there's less than a 5% chance that it's true – and we split the 5% we had to bet with across both sides of the normal curve. We put half (2½%) on the right side and half (2½%) on the left side to cover both ways in which a sample mean could be different from the population mean (a lot larger or a lot smaller than the population mean). In the directional version of the test, a sample mean below 100 isn't consistent with the prediction of the experimenter. It just doesn't make sense to think that a classroom of gifted and talented seventh graders has a mean I.Q. below 100. The question is whether our sample mean is far enough above 100 to get us to believe that these are gifted and talented seventh graders. So if we still want to use an alpha level of .05, we don't need to put half on one side of the curve and half on the other side. We can put all 5% of our bet on the right side of the curve. See Figure 3.6.

Figure 3.6



If we put all 5% on the right side of the normal curve, how many critical values are we going to have to deal with? Just one. The decision rule changes so that it tells us how far the standard score for our sample mean has to be above a standard score of zero before we can reject the null hypothesis.

If $Z_{\bar{x}} \geq (\text{some number})$ reject H_0

The only thing left is to figure out what the critical value ought to be. How about +1.96? Think about it. Where did that number come from? That was how many standard deviations about zero you had to go to hit the start of the upper 2½% of the values that

make up the normal curve. But that's not what we need here. We need to know how many standard deviations you have to go above zero before you hit the start of the upper 5% of the values that make up the normal curve. How do you find that? Use the normal curve table! If 45% of all the values that make up the curve are between a standard score of zero and the standard score that we're interested in, that means that the standard score for our critical value is 1.645! So the decision rule for our directional test becomes...

If $Z_{\bar{x}} \geq +1.645$ reject H_0

Obviously, the standard score for our sample mean is greater than the critical value of +1.645 so our decision is to reject the null hypothesis. This means that we're willing to accept the alternative hypothesis. Our conclusion, therefore, is that ***the mean of the 25 seventh graders in the class is significantly greater than the population mean for typical/average seventh graders.***

Advantages and disadvantages of directional and non-directional tests. The decision of whether to conduct a directional or non-directional test is up to the investigator. The primary advantage of conducting the directional test is that (as long as you've got the direction of the prediction right) the critical value to reject the null hypothesis will be a lower number (e.g., 1.64) than the critical value you'd have to use with the non-directional version of the test (e.g., 1.96). This makes it more likely that you're going to be able to reject the null hypothesis. So why not always do a directional test? Because if your prediction about the direction of the result is wrong there is no way of rejecting the null hypothesis. In other words, if you predict ahead of time that a bunch of seventh graders are going to have an average I.Q. score that is significantly greater than the population mean of 100 and then you find that their mean I.Q. is 7.0 standard deviations below 100 can you reject the null hypothesis? No! It doesn't matter if the standard score for that sample mean is 50 standard deviations below 100. In this case, no standard score below zero is consistent with the prediction that the students in that class have an average I.Q. that is greater than 100. Basically, if you perform a directional test and guess the direction wrong you lose the bet. You're stuck with having to say that the sample mean is not significantly greater than the mean of the population. What the researcher absolutely should not do is change their bet after they have a chance to look at the data. The decision about predicting a result in a particular direction is made ***before*** the data is collected. After you place your bet, you just have to live with the consequences,

So if a theory predicts that the result should be in a particular direction, use a directional test. If previous research gives you a reason to be confident of the direction of the result, use a directional test. Otherwise, the safe thing to do is to go with a non-directional test.

There are some authors who feel that there is something wrong with directional tests. Apparently, their reasoning is that directional tests aren't conservative enough. It is certainly true that directional tests can be misused, especially by researchers that really had no idea what the direction of their result would be, but went ahead and essentially cheated by using the lower critical value from a directional test. However, the logic of a directional test is perfectly sound. An alpha level of 5% is an alpha level of 5%, no

matter whether the investigator has used that alpha level in the context of a directional or a non-directional test. If you've got 5% worth of bet to place, it ought to be up to the researcher to distribute it in the way they want – as long as they're honest enough to live with the consequences. I personally think it's rather self-defeating to test a directional question, but use a critical value based on having a rejection region of only 2½% on that side of the curve. The reality of doing that is that the researcher has done their directional test using an alpha level of .025, which puts the researcher at an increased risk of missing the effect their trying to find (a concept we'll discuss in the next section).

Errors in decision making

When you make a decision, like the one the made above, what do you know for sure? Do you know that the null hypothesis is true? Or whether the alternative hypothesis is true? No. We don't get to know the reality of the situation. But we do get to know what our decision is. You know whether you picked the null hypothesis or the alternative hypothesis. So in terms of the outcome of your decision there are four ways that it could turn out.

		<u>Reality</u>	
		HO False	Ho True
<u>Your Decision</u>	Reject HO		
	Fail to Reject HO		

There are two ways that you could be right and there are two ways that you could be wrong. If you decide to reject the null hypothesis and, in reality, the null hypothesis is false you made the right choice – you made a correct decision. There was something there to find and you found it. Some people would refer to this outcome as a “Hit”

		<u>Reality</u>	
		HO False	Ho True
<u>Your Decision</u>	Reject HO	Correct Decision “Hit”	
	Fail to Reject HO		

If you decide not to reject the null hypothesis and, in reality, the null hypothesis is true then again you made the right choice: you made a correct decision. In this case, there was nothing there to find and you said just that.

		<u>Reality</u>	
		HO False	Ho True
<u>Your Decision</u>	Reject HO	Correct Decision "Hit"	
	Fail to Reject HO		Correct Decision

Now, let's say that you decide to reject the null hypothesis, but the reality of the situation is that the null hypothesis is true. In this case, you made a mistake. Statisticians refer to this type of mistake as a **Type I error**. Basically, a Type I error is saying that there was something there when in fact there wasn't. Some people refer to this type of error as a "False Alarm". So what are the odds of making a Type I error? This is easy. The investigator decides how much risk of making a Type I error they're willing to run before they even go out and collect their data. The alpha level specifies just how unlikely the null hypothesis would have to be before we're not willing to believe it anymore. An alpha level of .05 means that we're willing to reject the null hypothesis when there is still a 5% chance that it's true. This means that even when you get to reject the null hypothesis, you're still taking on a 5% risk of making a mistake – of committing a Type I error.

		<u>Reality</u>	
		HO False	Ho True
<u>Your Decision</u>	Reject HO	Correct Decision "Hit"	Type I Error "False Alarm"
	Fail to Reject HO		Correct Decision

Finally, let's say that you decide that you can't reject the null hypothesis, but the reality of the situation is that the null hypothesis is false. In this case, there was something there to find, but you missed it! The name for this type of mistake is a Type II error. Some people refer to this type of mistake as a "Miss".

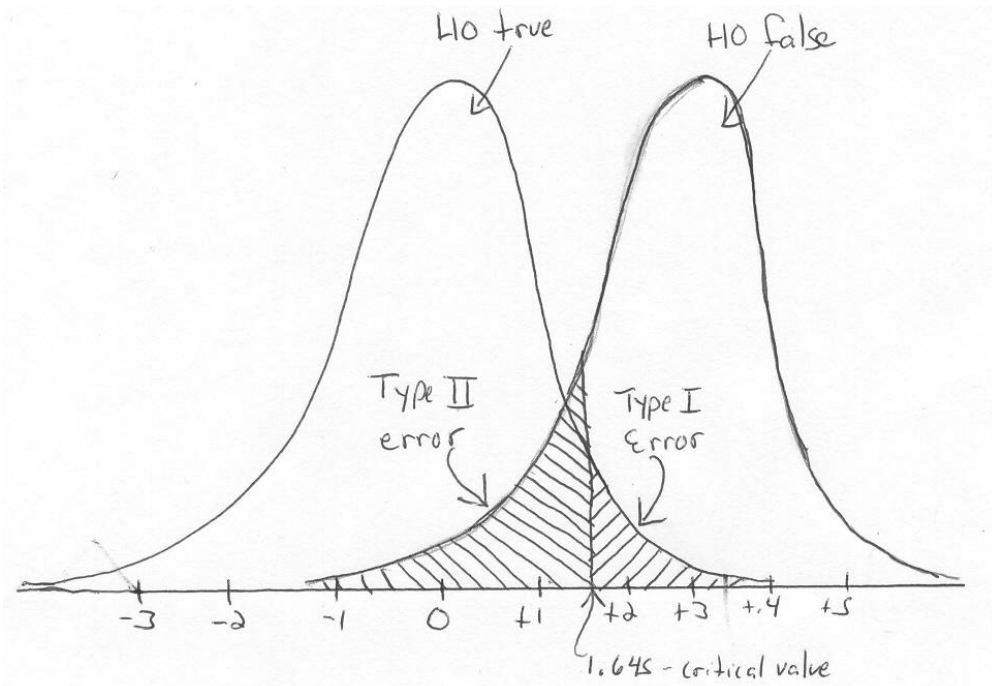
		<u>Reality</u>	
		HO False	Ho True
<u>Your Decision</u>	Reject HO	Correct Decision "Hit"	Type I Error "False Alarm"
	Fail to Reject HO	<i>Type II Error</i> <i>"Miss"</i>	Correct Decision

So what are the odds of committing a Type II error. This one's not as easy. The one thing that it's not is 95%! Just because the risk of making a Type I error is 5%, that doesn't mean that we've got a 95% chance of making a Type II error. But one thing that we do know about the risk of a Type II error is that it is inversely related to the risk of a Type I error. In other words, when an investigator changes their alpha level they're not only changing the risk of a Type I error, they're changing the risk of a Type II error at the same time.

- If an investigator changes their alpha level from .05 to .01 the risk of making a Type I error goes from 5% to 1%. They're changing the test so that it's more difficult to reject the null hypothesis. If you make it more difficult to reject the null hypothesis, you're making it more likely that there might really be something there, but you miss it. ***If you lower the alpha level to reduce the risk of making a Type I error, you'll automatically increase the risk of making a Type II error.***
- If an investigator changes their alpha level from .05 to .10 the risk of making a Type I error will go from 5% to 10%. They're changing the test so that it's easier to reject the null hypothesis. If you make it easier to reject the null hypothesis, you're making it less likely that there could really be something out there to detect, but you miss it. ***If you raise the alpha level and increase the risk of making a Type I error, you'll automatically lower the risk of making a Type II error.***

The graph below shows where the risks of both a Type I error and Type II error come from. The example displayed in the graph is for a directional test.

Figure 3.7



The curve on the left is the sampling distribution of the mean we talked about before. Remember, this curve is made up of sample means that were all collected when the null hypothesis is true. The critical value is where it is because this point is how far above the mean of the population you have to go to hit the start of the 5% of sample means that you're least likely to get when the null hypothesis is true. Notice that 5% of the area under the curve on the left is in the shaded region. The percentage of the curve on the left that is in the shaded region represents in risk of committing a Type I error.

Now for the risk of committing a Type II error. Take a look at the curve on the right that is labeled "HO False". This curve represent the distribution of a bunch of sample means that would be collected if the null hypothesis is false. That's why it's labeled "HO False". Let's say that the reality of the situation is that the null hypothesis is false. In other words, when we collected our sample mean it really belonged in this collection of other sample means. Now, let's say that the standard score for our sample mean turned out to be +1.45. What would your decision be (reject HO or don't reject HO)? Of course, the observed value is less than the critical value so you're decision is going to be that you fail to reject the null hypothesis. Are you right or wrong? We just said that the null hypothesis is false so you'd be wrong. Any time the null hypothesis really is false, but any time you get a standard score for your sample mean that is less than +1.645 you're going to be wrong. Look at the shaded region under the "HO False" curve. The percentage of area that falls in the shaded region under this curve represents the odds of committing a Type II error. All of the sample means that fall in this shaded region

produce situations where the research will decide to keep the null hypothesis when they should reject it.

Now let's say that the researcher had decided to go with an alpha level of .025. The researcher has done something to make it more difficult to reject the null hypothesis. How does that change the risks for both the Type I and a Type II error?

Well, if we perform a directional test using alpha level of .025 what will the critical value be? +1.96, of course. On the graph, the critical value will move to the right. What percentage of area under the curve labeled "HO True" now falls to the right of the critical value? 2½% . The risk of committing a Type I error has gone down. And if the critical value moves to the right, what happens to the risk of committing a Type II error? Well, now the percentage of area under the curve on the right – the one labeled "HO False" – has gone way up. This is why, when a researcher uses a lower alpha level, the risk of making a Type II error goes up. When you move the alpha – when you move the critical value – you change the risk for both the Type I and the Type II error.

The choice of alpha level. Okay. So why is 90-something percent of research in the behavioral sciences conducted using an alpha level of .05? That alpha level of .05 means that the researcher is willing to live with a five percent chance that they could be wrong when they reject the null hypothesis. Why should the researcher have to accept a 5% chance of being wrong? Why not change the alpha level to .01 so that now the chance of making a Type I error is only 1%? For that matter, why not change the alpha level to .001, giving them a one in one-thousand chance of making a Type I error? Or .000001? The answer is that if the risk of a Type I error is the only thing they're worried about, that's exactly what they should do. But of course, we just spent some time saying that the choice of an alpha level determines the levels of risk for making both a Type I or a Type II error. Obviously, if one uses a very conservative alpha level like .001, the odds of committing a Type I error will only be one in one-thousand. However, the investigator has decided to use an alpha level that makes it so hard to reject the null hypothesis that they're practically guaranteeing that if an effect really is there, they won't be able to say so – the risk of committing a Type II error will go through the roof.

It turns out that in most cases an alpha level of .05 gives the risk a happy medium in terms of balancing the risks for both types of errors. The test will be conservative, but not so conservative that it'll be impossible to detect an effect if it's really there.

In general, researchers in the social and behavioral sciences tend to be a little more concerned about making a Type I error than a Type II error. Remember, the Type I error is saying that there's something there when there really isn't. The Type II error is saying there's nothing there when there really is. In the behavioral sciences there are often a number of researchers who are investigating more or less the same questions. Let's say that 20 different labs all do pretty much the same experiment. And let's say that in this case the null hypothesis is true – there's nothing there to find. But if all 20 labs conduct their test using an alpha level of .05, what's likely to happen? The alpha level of .05 means that the test is going to be significant one out of every 20 times just by chance. So if 20 labs do the experiment, one lab will find the effect by accident and publish the

result. The other 19 labs will, correctly, not get a significant effect and they won't get their results published (because null results don't tend to get published). The one false positive gets in the literature and takes the rest of the field off on a wild goose chase. To guard against this scenario researchers tend to use a rather conservative alpha level, like .05, in their tests. The relatively large risk of making a Type II error in a particular test is offset by the fact that if one lab misses an effect that's really there, one of the other labs will find it. Chance determines which labs win and which labs lose, but the field as a whole will still learn about the effect.

Now, just to be argumentative for a moment. Can you think of a situation where the appropriate alpha level ought to be something like .50? .40?! That means the researcher is willing to accept a 40% risk of saying that there's something there when there really isn't – a 40% chance of a false alarm. The place to start in thinking about this is to consider the costs associated with both types of errors. What's the cost associated with missing something when it's really there (the Type I error)? What's the cost associated with saying that there's nothing there where there really is (the Type II error)?

Consider this scenario. You're a pharmacologist searching for a chemical compound to use as a drug to cure the AIDS. You have a potential drug on your lab bench and you test it to see if it works. The null hypothesis is that people with AIDS who take the drug don't get any better. The alternative hypothesis is that people with AIDS who take the drug do get better. Let's say that people who take this drug get very sick to their stomach for several days. What's the cost of committing a Type I error in this case? If the researcher says the drug work when it really doesn't at least two negative things will happen: (a) a lot of people with AIDS are going to get their hopes up for nothing and (b) people who take the drug are going to get sick to their stomachs with getting any benefit from the drug. There are certainly significant costs associated with a Type I error. But how about the costs associated with making a Type II error? A Type II error in this case would mean that the researcher had a cure for AIDS on their lab bench – they had it in their hands – they tested it, and they decided it didn't work. Maybe this is the ONLY drug that would ever be effective against AIDS. What are the costs of saying the drug doesn't work when it really does? The costs are devastating. Without that drug millions of people are going to die. So which type of mistake should the researcher try not to make? The Type II error, of course. And how can you minimize the risk of a Type II error? By setting the alpha level so high, that the risk of a Type II error drops down to practically nothing.

It seems like the researcher's choice of an alpha level ought to be based on an assessment of the costs associated with each type of mistake. If the Type I error is more costly, use a low alpha level. If the Type II error is more costly, use a higher alpha level. Using an alpha level of .05 out of habit, without thinking about it, strikes me as an oversight on the part of investigators in many fields where costly Type II errors could easily have been avoided through a considered use of alpha levels of .10 or higher. The "one size fits all" approach to selecting an alpha level is particularly unfortunate considering the ease with which software packages for data analyses allow researchers to adopt whatever alpha level they wish. When you read the results section of a paper, ask yourself what the costs

of Type I and Type errors are in that particular situation and ask yourself whether you think the author's choice of an alpha level is justified.

Statistical inference is gambling. A researcher places her or his bet and then has to be willing to live with the consequences. It may not be perfect, but it's the best we can do!