

Independent samples t-test

*Dr. Tom Pierce
Radford University*

- The logic behind drawing causal conclusions from experiments
- The sampling distribution of the difference between means
- The standard error of the difference between means
- The equation for t as a standard score

Research in psychology is based on the collection of data. Usually the researcher has some idea of what they expect the data to look like. For example, a theory may predict that one group will have higher scores on average than another group. In a situation like that the researcher is asked to decide whether the results they actually get fit the pattern of results predicted by the theory.

This decision could be made by looking at the means for the two groups and thinking, in a subjective or intuitive way, about whether the mean for group A is far enough above the mean of group B for it to be believable that the prediction made by the theory has been supported by the evidence. But psychologists, as scientists, would rather have a way of making these decisions that is more precise than this. Certainly, one person's subjective decision might be different from another person's subjective decision. Tools of statistical inference are useful because they are based on very clearly defined rules for deciding whether or not the predictions of an experimenter are supported by the data. In the sections to follow, we will see how psychologists (a) define the decision that has to be made (b) state a very clear rule for making the decision, and (c) compute the necessary information to use this clear and unambiguous decision rule. First, let's see how researchers define their decisions as a choice between two options: a null hypothesis and an alternative hypothesis.

Null and alternative hypotheses

Statistical inference begins by recognizing that research questions can be stated in terms of a choice between two very clear and mutually exclusive options. One option holds that the predicted difference between comparison groups does not exist out in the real world (the population). The only reason that the two group means are different from each other is *chance*. This first option is known as the **null hypothesis**. The other option is that the predicted difference does exist out in the real world. There *is* a reason (other than chance) why the two group means are different from each other. This second option is known as the **alternative hypothesis**.

The alternative hypothesis is almost always the outcome that the researcher predicts will happen. For example, if the researcher predicts that ***younger adults will make more errors than a group of older adults*** the alternative hypothesis for the test of this idea is that "*the mean number of errors for the younger group will be significantly higher than the mean number of errors for the older group*". The null hypothesis is the *opposite* of

the alternative hypothesis. The null hypothesis in this situation is that "*the mean number of errors for younger adults is not significantly greater than the mean number of errors for the older group*".

In this case the researcher has phrased the alternative hypothesis as a *directional* question. It is a directional question or hypothesis because there is a specific direction to the predicted difference between the two groups. If the researcher had phrased their prediction (the alternative hypothesis) as "the mean number of errors by younger adults will be significantly *different* from the mean number of errors made by older adults", we would say that this alternative hypothesis is in the form of a *non-directional* question or hypothesis. The null hypothesis in this case would, again, be the opposite of the alternative hypothesis: "The mean number of errors made by the younger group will *not be significantly different* from the mean number of errors made by the older group".

The choice of whether the alternative hypothesis is stated in the form of a directional or a non-directional hypothesis is something that's left up to the experimenter. Certainly, the directional hypothesis is more precise, but as we'll see later, it may not be advisable to make a directional hypothesis when the researcher isn't really sure that the direction is going to turn out like they think.

In any case, *you* will know whether another researcher has made a directional or a non-directional prediction on the basis of how their predictions are phrased. If the researcher uses the word "different" in the prediction (as in "one group mean is different from another group mean") the prediction is non-directional. If either the phrase "less than" or "greater than" is used, the prediction is directional. You will need to know how to tell the difference for your exam.

In sum, the two-sample (independent samples) t-test is a choice between two possibilities; a null hypothesis and an alternative hypothesis. The null hypothesis is that the researcher's prediction is not true. The alternative hypothesis is that the researcher's predicted difference is true.

Statistical inference: a decision based on the odds

So, the two sample t-test gives us a way to decide between a null hypothesis and an alternative hypothesis. But, unfortunately, this doesn't mean that the test will tell us *the truth* about which choice is correct. It turns out that in none of the tests used by psychologists do we get to know *anything* for certain. We never get to know the "truth" about our research questions. But we do get to know something that is not that far from the truth. We do get to know *something* about how good our predictions are. We get to know the odds of making a mistake if we decide to choose the alternative hypothesis. We'll talk about the reason for this in a bit. But just for the moment, take my word for it that we can have *direct knowledge* about the *odds of the null hypothesis being true*. This doesn't mean you're necessarily going to be correct if you decide the null hypothesis isn't true, but it does mean that you would at least know the odds of your being wrong if you

decide the null hypothesis isn't true. You **would** know how much risk you're running if you reject the possibility that the null hypothesis is true.

The two-sample t-test (and all other tests of statistical inference) is a technique for making a decision based on the odds. Specifically, the two sample t-test is a technique based on the odds of the *null hypothesis being true*. The researcher can decide ahead of time that if the odds of the null hypothesis being true are less than a certain amount (say, 5%) then they will decide that *the null hypothesis is not true*.

This is the logic of statistical inference. Because we can only know the odds of the null hypothesis being true, we can make a rule that says that "if there is less than a 5% chance that the null hypothesis is true, reject the idea that the null hypothesis is true, and accept the alternative hypothesis". The alternative hypothesis, the statement the experimenter would **like** to be able to make, can only be accepted after it's opposite, the null hypothesis, has been rejected. This certainly seems counter-intuitive. You can only get to say what you want by rejecting the possibility of its opposite being true. For example, if I think that younger adults make more errors than older people, I can only claim that my data give me evidence for this by *rejecting* the null hypothesis that younger adults do not make more errors than older adults.

It's a strange and different way of thinking. But it has to be this way because the *only thing we can know for sure are the odds of the null hypothesis being true*.

In sum, the two-sample test is a decision based on the odds of the null hypothesis being true. Before I collect any data, I can write my rule for knowing when to reject the null hypothesis. For example, my decision rule might be that I will "reject the null hypothesis when the odds of its being true are less than five percent".

The next thing we have to deal is to make this more concrete. How do I know if the odds are less than five percent that the null hypothesis is true?

Testing a difference between two treatment means

Let's say that I plan an experiment in which I intend to compare the mean of one group or sample of people against the mean of another group or sample of people. Specifically, I believe that the mean of five younger adults on the number of errors they make on a RT test will be significantly *different* than the mean number of errors made by five older subjects. This is a non-directional test because I am not making a specific prediction about which group will have higher or lower scores. Let's say that I'm giving a new kind of measure and I really don't have any reason to predict a specific direction; I just think that the two groups are going to be different on this dependent variable. I decide that my decision rule, in general, will be that I will reject the null hypothesis if I find that the odds are less than five percent that it is true.

The first point to make in describing how the test works is that every sample mean is an estimate of a population mean. Remember, there's a difference between a population and

a sample. A population is made up of *every* member of the set of people you want to know something about. A sample is a *representative subset* of the set of people you want to study. People study samples only when they can't get scores from everyone in the population. For example, I would much rather know the mean number of errors made by a population of older adults than the mean number of errors made by a sample of 5 older adults. This is because I want to be able to draw conclusions about all older adults, not just five older adults. But psychologists just don't have the time or the money to study all older adults. So the best we can do is to learn what we can from samples. The mean of my sample of older adults is as close as I can get to the mean of what I really want, the mean of the population. The mean of the sample is an estimate of that population mean.

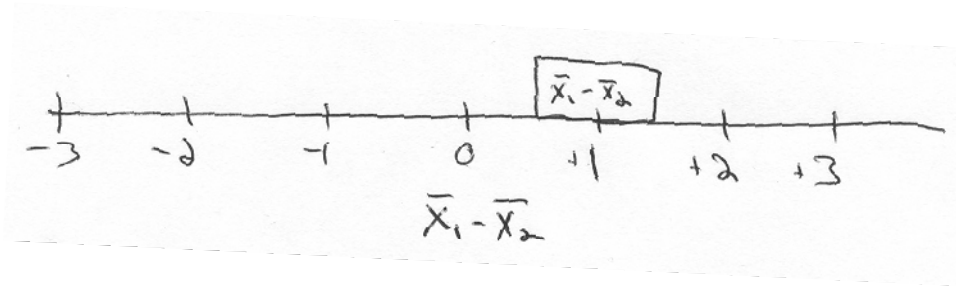
In our experiment, we have two sample means; the mean of five older adults and the mean of five younger adults. Either one of these two sample means *might not be a perfect estimate* of the population means they are trying to estimate. A sample mean might be higher than it should be *by chance*. A sample mean might be lower than it should be *by chance*.

So, what if the null hypothesis for this experiment really were true? Younger adults do not differ from older adults in the number of errors they make on a task. What should the relationship between the two sample means be? It seems logical to think that if the null hypothesis is true and there is no reason for there to be a difference between the two sample means, then the *difference between the two sample means should be zero*. But, does it have to be zero? Remember, we're dealing with two sample means. Sample means are just estimates. They can be too high or too low by chance. So, when the null hypothesis is true, do the two sample means have to be the same, giving us a difference between sample means of zero? No. We could have gotten a difference between the two sample means that's different from zero *just by chance*.

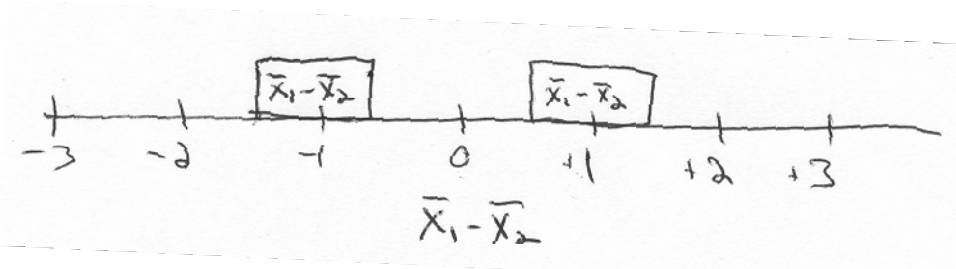
So, the question then becomes "how different from zero would the difference between the two sample means have to be before it's just not believable that you could get a difference that large by chance alone?" That's the questions that the t-test answers. If it's not believable you could get a difference that large by chance, you would be justified in deciding that it wasn't that large by chance alone. It was that large due to a reason in addition to chance. There must have been something about your independent variable that was forcing the sample means away from each other. And thus you could conclude that the independent variable had an effect on the dependent variable.

There could be a difference between the two sample means even when the null hypothesis is true. We talked about that above. Now its time to talk about how we can know the odds of the null hypothesis being true. Think of it this way. Imagine that you could know for sure that the null hypothesis is true. If the null hypothesis is true and there's no reason for there to be a difference between the two sample means, you would expect the difference between the two sample means ($\bar{X}_1 - \bar{X}_2$) to be equal to zero. But we already discussed the fact that this doesn't have to turn out this way when you're comparing sample means.

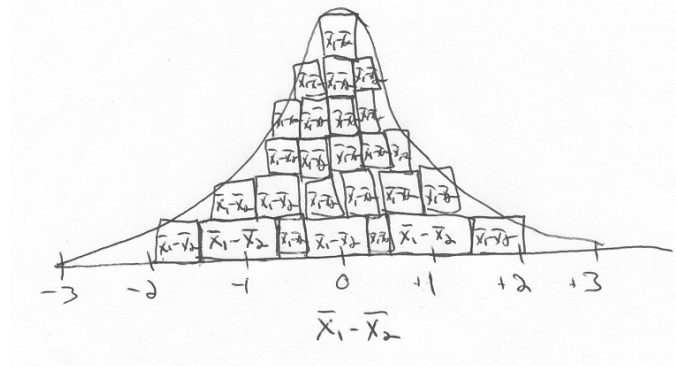
Let's say you could know for a fact that the null hypothesis is true (which you could never know in reality) and you do the experiment a first time. You get the mean of five older adults and you get the mean of five younger adults. You subtract the mean of the older adults from the mean of the younger adults. You now have **one number**, one ***difference between sample means***. Let's say that this difference turns out to be +1.0, and you draw a box representing that difference where it belongs on a scale of possible differences. See below.



Now, you do the same experiment again. Again, you know somehow that the null hypothesis is true. This time the mean for group A, the younger adults is a little less than the mean from Group 2, the older adults, by chance. Let's say that this gives us a difference in this experiment of -1.0. And you plot this second difference between sample means where it belongs on the scale. See below.



Let's say that you do the same experiment over and over and over. Every time you do the experiment the null hypothesis is true. Every time you do the experiment, the only reason for getting a difference other than zero is *chance*. If you could collect enough of these differences (differences obtained when the null hypothesis is true) the shape of this distribution will come to closely resemble the normal curve. It won't look exactly like the normal curve, but close enough for us to think of it as the normal curve for right now.



Now you have a frequency distribution of a very large number of differences between sample means when the null hypothesis is true. The name for this set of things is the **sampling distribution of the difference between means**. The sampling distribution of the difference between means is at the very heart of how the t-test works. If you don't understand what this sampling distribution is and what it's used for, you don't understand the t-test.

In a two-sample t-test, you compute one difference between sample means from the data you collected. You don't know whether this difference was obtained when the null hypothesis is true or not (that's why you're doing the test). The sampling distribution of the difference between means gives you a set of differences collected *when the null hypothesis is true*. This sampling distribution gives you a set of values to compare the differences from your experiment against.

This frequency distribution of differences between means can be used to give you the odds that your one sample mean belongs in that set. After all, we know from the normal curve that if you go one standard deviation from the center of the curve (out to standard scores of -1 and +1) you will include approximately 68% of all the values in that set. If you go two standard deviations from the center of the curve you will include approximately 95% of all that values that make up that curve. So, we should be able to find exactly how many standard deviations away from the center of the curve our one difference would have to be in order for the odds to be less than five percent that it belongs in that set of differences, *the set of differences where the null hypothesis is true*. If we convert our difference to a standard score, how many standard scores above or below zero would our difference need to be before we reject the null hypothesis?

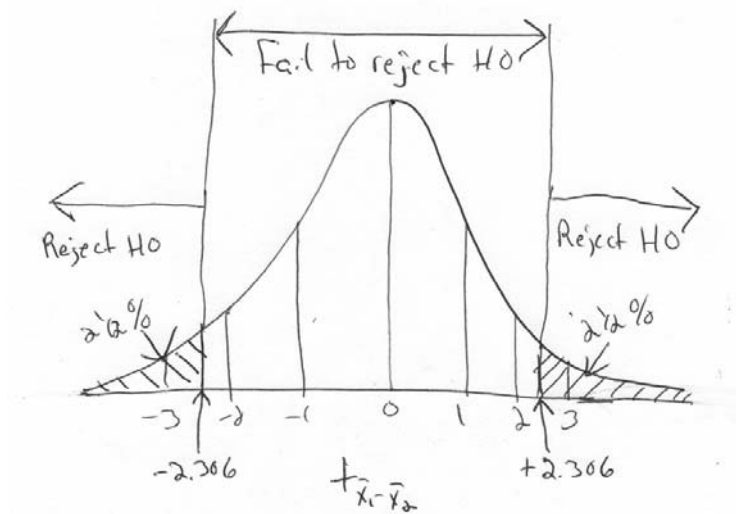
We can be a little more concrete in how our decision rule reads at this point: "If the standard score for our difference between sample means is either less than a particular negative standard score or greater than a particular positive standard score, reject the null hypothesis". The only two things we need are the two comparison standard scores that correspond to the two points on the curve (one on the negative side of the curve and one on the positive side of the curve) and the standard score that we get for the difference between sample means in our data.

Now we just have to find these standard scores and our decision rule will be useable. The comparison standard scores come from Table C in the back of your stats book. The values in the table are referred to as values of t rather than values of Z (the usual symbol for a standard score) because we can only really get *estimates of Z* in this situation. So we're going to compare one value for t that we compute from our data against comparison values for t that we're going to get from table C. That makes this a *t-test*.

There are three things you need to know in order to find the standard scores you need. First you need to know whether you're doing a directional (one-tailed) or a non-directional (two-tailed) test. Next you need to know the odds you're using to reject the null hypothesis. We're saying we'll reject the null hypothesis if the odds are less than five percent of it's being true. In this table these odds are expressed as a proportion (a number between zero and one). The odds we're using correspond to a proportion of .05. So the standard score we're looking up is in the .05 column for a non-directional test. This is the column starting at the top with a 12.706.

Now we need to know the row to look in. This is based on the number of *degrees of freedom* for the experiment. This is equal to the number of people in the first group (5) plus the number of people in the second group (5) minus 2. In symbols the number of degrees of freedom is equal to $N_1 + N_2 - 2$, which in this experiment is equal to 8.0. The comparison standard score from the body of the table for a non-directional test at the .05 level and eight degrees of freedom is 2.306.

The curve is symmetrical so that the comparison value is equally as far above the center of the curve (+2.306) as it is below to center of the curve (-2.306). The location of the comparison values (or critical values as they are often called) are displayed below. The decision rule should now read "If the standard score for the difference between sample means is greater than +2.306 or less than -2.306, reject the null hypothesis".



The critical values for t don't come out of nowhere. They get something very specific done. They tell us how far away from a standard score of zero (the value you should get

when the null hypothesis is true) that the standard score for a difference between means has to be before the odds are *less than five percent* that this difference belongs in the set of differences one would obtain if the null hypothesis were true. In other words, you have to go out that far from the center of the curve, in standard score units, before you get to the point where only five percent of values belonging in that curve fall outside of these values by chance. If you get a value that falls outside of these critical values, then the odds are less than five percent that you got a value that far from zero by chance. If this is true then you're on pretty firm ground in deciding that you didn't get that result by chance. You got the result because *the null hypothesis isn't true and your difference between sample means doesn't belong in the set of differences between sample means that you get when the null hypothesis is true*. This is a very important concept and, again, if you don't understand this concept, you don't understand how the two-sample t-test works. Don't feel strange about spending a significant amount of time struggling with these concepts. It's not easy stuff, but it does make sense after you've thought your way through it a few times.

Now that we know where the critical values in the decision rule come from, the only thing left is to find the standard score that corresponds to our difference between sample means. This standard score will tell us how many standard deviations our difference falls from the mean of this set of differences. Remember, in general, a standard score tells us how many standard deviations one value falls from the mean of a set of values. In general, the equation for a standard score is one number minus the mean of all the numbers in the set divided by the standard deviation of all the numbers in the set. Here, instead of comparing one score to the mean of a set of scores, we're comparing one difference between sample means to the average of a very large set of other differences between sample means.

Let's say that the mean number of errors for the younger group is 10.0 and the standard deviation of scores in the younger group is 2.0. The mean number of errors for the older group is 6.0 and the standard deviation of scores in the older group is 3.0. Our difference between the two sample means is 10-6, or 4.0. We need to take our one difference between the means and then subtract the average of all the differences between means that you'd get if the null hypothesis were true. So what's the mean that we need here? When the null hypothesis is true you're supposed to get a difference between the means of zero. When you do the same experiment over and over and over when the null hypothesis is true, half the time the differences between the means are greater than zero and half the time the differences between the means are less than zero. ***But the average of all of these differences between means is exactly equal to zero!*** So the equation for t , up to this point, is equal to one difference between sample means (4.0) minus the average of all the differences between sample means that you get if the null hypothesis were true (zero). Once we know how far our difference between means is above zero all we have to do is to ***divide that number by the average amount that differences between means deviate from zero***. We need to divide by the standard deviation of all the differences between means that made up our curve. This standard deviation is referred to as the **standard error of the difference between means**.

The equation for t in this case becomes, essentially, one difference between means minus the average of a bunch of differences between means divided by the standard deviation of that same bunch of differences between means, or ...

$$t_{\bar{X}_1 - \bar{X}_2} = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_{\bar{X}_1 - \bar{X}_2}}$$

The equation for the standard error of the difference between means is...

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2} \left(\frac{1}{N_1} + \frac{1}{N_2} \right)}$$

Plugging these values into the equations gives us the value for t observed in our experiment.

Younger	Older
-----	-----
$N_1 = 5$	$N_2 = 5$
$\bar{X}_1 = 10$	$\bar{X}_2 = 6$
$S_1 = 2.0$	$S_2 = 3.0$

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(5 - 1)(4) + (5 - 1)(9)}{5 + 5 - 2} \left(\frac{1}{5} + \frac{1}{5} \right)} = 1.61$$

$$t_{\bar{X}_1 - \bar{X}_2} = \frac{(10 - 6) - 0}{1.61} = \frac{4}{1.61} = 2.48$$

When we take this observed value for t and compare it to the critical values for t specified in the decision rule we find that the observed value of +2.48 is greater than +2.306, therefore the researcher should decide to reject the null hypothesis. Their conclusion

would be that “Older adults make significantly more errors than younger adults, $t(8) = 2.48, p < .05$ ”.

Directional tests

The alternative hypothesis for the experiment above was that the mean number of errors for the younger group would be different from the mean number of errors made by the older group. This is a non-directional prediction. What if the researcher had a reason to make an educated guess about which group would do more poorly? What if the researcher predicted that subjects in the younger group would have more errors than subjects in the older group? This is an example of a *directional hypothesis*. And the t-test for a directional t-test is slightly different than that for a non-directional t-test.

The calculation of the observed value for t doesn't change a bit. The equations are exactly the same. The only thing that changes is the decision rule.

In this directional test the researcher is *predicting* that the mean of the younger group (group 1) will be higher than the mean of the older group (group 2). This means that the researcher is predicting that they'll get a *positive* value for t. Therefore, there's no need to distribute the 5% worth of risk to both sides of the curve. We can put it all on the positive side of the curve. This will give us a different critical value on the positive side of the curve and *no* critical value on the negative side of the curve. In other words, with this directional hypothesis here's no possible way of rejecting the null hypothesis if the investigator gets a negative value for t.

The researcher would look up the critical value by finding the column for a *one-tailed* test at the .05 level and eight degrees of freedom. The number from table C is 1.860. So, the decision rule for this directional test is "if the observed value for t is greater than 1.860, reject the null hypothesis".

Note that as long as the investigator guesses the direction correctly, it is easier to reject the null hypothesis. The observed value for t doesn't have to be as far above zero as in the non-directional test (1.860 compared to 2.306). This is the primary advantage to performing a directional test rather than a non-directional test. However, if the researcher guesses the direction wrong, then there is *no way* of rejecting the null hypothesis, no matter how different the two means are from each other.