

**Statistical Hypothesis Testing:  
how far from the tree can an acorn fall before we say it came from a different tree?**

**Authored by:**

**Paul F. Doruska (CF #1213)  
College of Natural Resources  
University of Wisconsin-Stevens Point**

**Upon Request of the  
Society of American Foresters Forest Science and Technology Board**

---

---

“An acorn never falls far from the tree” is a phrase most of us have heard or used, but probably not in the context of statistical hypothesis testing. Actually, we can gain insights into the process of statistical hypothesis testing by considering that simple phrase. If you continue to read what follows, please be aware that I am not a silviculturalist, an expert in stand dynamics, or an ecologist, so please don’t try to read too much into the example that follows, if we keep it simple, hopefully it will all make sense...

There is this mast producing oak tree in the woods, several actually, all of the same species, but we’ll start out just focusing on one of these trees. There are also some oak seedlings growing on the ground in this stand, again of that same species. An inquisitive person might begin to wonder from which tree did the acorn that produced a particular seedling originate (let’s leave those squirrels out of the picture at this point...).

I think many of us would agree that the seedlings closest to that chosen oak tree probably originated from acorns from that tree. We can’t be 100% sure, but that’s most likely the case. Now let’s move a little farther from that oak tree. From which tree did those seedlings originate? The chosen oak tree, or perhaps another, it gets a little tougher to decide (shy of genetic testing of course...) How far from the chosen tree are you? How close to the next oak tree are you? Those kinds of questions probably start arising.

We now move even farther from that originally chosen tree and consider the seedlings there. Then we move even farther from the original tree and consider seedlings there. Sooner or later, we will get to the point where we will eventually decide that the seedling we are looking at did not originate from an acorn from that originally selected tree, it most likely came from another oak tree that we are probably closer too at this point. We basically draw a line in the ground and say “that’s it, seedlings any farther than this probably did not originate from that tree, rather, they probably originated from a tree closer to where we are.”

Is there a chance a seedling this far away from the initially selected tree originated from an acorn from that tree (hello Mr. Squirrel!) – you bet there is, but the chances are pretty low. Similarly, is there a

chance a seedling extremely close to the initially selected tree originated from an acorn from a different tree (hello Ms. Squirrel!) –you bet there is, but the chances are pretty low. We will use our line on the ground to lead us to our conclusion.

With that simplistic example in place, we can gain insights into statistical hypothesis testing – as most of the process is contained in that example. Let's see how.

The statement being examined when we reached each seedling was: “this seedling originated from an acorn from the chosen oak tree.” That is an important statement as it relates to hypothesis testing as we are basically looking for evidence that makes that statement false, if we cannot find such evidence, we'll have to assume that seedling did indeed originate from an acorn from that tree. How does that statement relate to hypothesis testing? We've just identified the null hypothesis ( $H_0$  as it is often appears when written out). A key to the hypothesis testing process is the realization that the process of hypothesis testing assumes the null hypothesis is correct, and one looks for evidence (via your measurements or experiment) that the statement is false. All of this is done via numerical probabilities based on statistical distributions (terms known as the alpha-level [ $\alpha$ ] and a p-value), but even those are somewhat relatable to the acorn example. Let's see how.

In the example we used a line on the ground to make our call as to whether a given seedling originated from an acorn from the chosen tree. There was nothing set in stone about where that line occurred, it is something we selected based on the original tree and a distance from it. Recall that once we crossed that line, we felt that any seedling we came across probably did not come from that tree. Note that this is something we could have chosen without even looking at any seedlings, the location of that line is simply something that we chose. What's the link to hypothesis testing then? Well, we have just gotten a sense of  $\alpha$  as  $\alpha$  and the null hypothesis determine where that line falls for any hypothesis test – it has NOTHING to do with the data you collect or observe.

The  $\alpha$ -level is a numerical probability (often set to 0.05) that is used to determine how far away from the null hypothesis your observed data can be before you say “that's it, the null hypothesis is probably false”. Please note that the  $\alpha$ -level is not the actual distance away; rather it is used to SET the distance away. Said another way, the probability of being farther away from that chosen oak tree and still have a seedling originate from an acorn from that tree is less than 0.05 (or the chosen  $\alpha$ -level). That would make that pretty darn rare, so it probably is not the case (the null hypothesis is probably false, or reject the null).

If you ever have worked with or come across what is called a “critical value” for a hypothesis test, guess what, that line on the ground (distance away from the tree) in the example represents the “critical value”. If you are farther away from the null hypothesis than that (actually farther away from 0 on a number line than that once all the statistical math is performed) you are ready to say “that's it, the null hypothesis is probably false”. If you are closer, you don't have enough evidence to conclude the null hypothesis is false (fail to reject the null). Note that you can be close to that distance away, often close enough to want to say the null hypothesis is false, but statistically speaking, until you actually cross that line, you cannot conclude the null is false.

Hopefully you are still with me, and hopefully all of the above has not been so bad. So what is that darn p-value concept? We just talked through an example and hopefully got a handle on hypothesis testing and the notion of a p-value never came up – what is the big deal about the p-value? Basically, it's just another way to go about the hypothesis testing process, but the one most often used as most computer packages provide the p-values for hypothesis tests and have the practitioners draw their conclusions from them. Let's go back to our example and find out what that p-value concept is.

As you will recall, we had an oak tree, a line on the ground (critical value) that was a set distance from the tree essentially set by our chosen  $\alpha$ -level (rareness threshold from a probability perspective), and made our conclusion based on where the seedling fell (analogous to an observed or calculated test statistic) in relation to that line. But what if that line was not there? Could we still come to a conclusion? We absolutely can as long as we remember an important concept we've highlighted before: hypothesis tests are performed (numerical probabilities found) assuming the null hypothesis is correct.

In the material above, we chose the  $\alpha$ -level to determine that distance on the ground, and compared where our seedlings (our data) fell to that distance to draw our conclusion. Rather than doing that, what if we could find the numerical probability (on a 0-1 scale, close to 0 being rare, close to 1 being very probable) of those seedlings (or observing our actual data) originating from acorns from that tree assuming that the null hypothesis was true (essentially assuming the seedlings came from that chosen tree, but realizing the farther away from the tree the seedlings occurred the less likely, or the lower the probability, that was). If we could find that numeric probability and it was relatively large, we would not have enough evidence to conclude the null was probably false (we would be unable to say the seedlings probably did not originate from acorns from that tree). But if we could find that numeric probability and it was pretty small, I think you would agree that we do have evidence to conclude the null was probably false (we could say the seedlings probably came from a different tree). That numeric probability we just talked about in the previous two sentences IS the p-value.

The p-value is the probability of observing our actual data or data more extreme (farther away from the tree in this case) again assuming the null hypothesis is correct (essentially assuming the seedlings came from that chosen tree, but realizing the farther away from the tree the seedlings occurred the less likely, or the lower the probability, that was). Go back and read that last sentence again – it is a very important one. Finding the p-value involves our observed data and the null hypothesis, it has NOTHING to do with the chosen  $\alpha$ -level. However, in order to draw a statistical conclusion, we need both the p-value and the  $\alpha$ -level. Here is why.

We read earlier that if the seedlings being examined (our data) crossed that line on the ground we would conclude the null hypothesis was probably false. The distance away from the tree to that line on the ground was determined by the  $\alpha$ -level (again that number is typically 0.05). Here is another way to interpret that line. If the seedlings we examined (our data) fell exactly on that line, the probability of observing that data or data more extreme than that (farther away from the tree than that) AND have those seedlings originate from acorns from the selected tree (the null hypothesis is assumed to be correct!) would equal 0.05 (or the chosen  $\alpha$ -level). If you are still with me, then we should be able to answer the following two questions.

If the seedlings we examined (our data) were closer to the chosen tree than the distance to that line, by now we know we would not be able to say the null hypothesis was false, BUT what would be the numeric probability of having those seedlings be from that tree (again assuming the seedlings came from that chosen tree, but realizing the farther away from the tree the seedlings occurred the less likely, or the lower the probability, that was). Hopefully you have determined that numeric probability would be something LARGER than the chosen  $\alpha$ -level: larger than 0.05 in this case.

If the seedlings we examined (our data) were farther away from the chosen tree than the distance to that line, by now we know we would say the null hypothesis was probably false, BUT what would be the numeric probability of having those seedlings be from that tree (again assuming the seedlings came from that chosen tree, but realizing the farther away from the tree the seedlings occurred the less likely, or the lower the probability, that was). Hopefully you have determined that numeric probability would be something SMALLER than the chosen  $\alpha$ -level: smaller than 0.05 in this case.

Well then, we just drew statistical conclusions using the p-value. Anytime the p-value from a hypothesis test is less than the chosen  $\alpha$ -level, one can say the null is probably false – the only way for that to happen in the example is for the observed seedlings to be farther away from the chosen tree than that line on the ground, the distance to which was set by  $\alpha$ .

The above represents the basics of hypothesis testing. Hypothesis tests are performed by finding the probability of obtaining what we observed through our experiment or data collection process assuming what is stated in the null hypothesis is true. If what we observed (our data) is “not so rare” (the seedling did not cross that line) given what is stated in the null hypothesis we do not have enough evidence to state the null hypothesis is probably false (we fail to reject the null). However, if what we observed is “rare” (the seedlings did cross that line) given what is stated in the null hypothesis we do have enough evidence to state that the null hypothesis is probably false (we reject the null). The chosen  $\alpha$ -level sets that “rareness” threshold.

To complete the picture, there are a few more concepts we need to discuss. First, for every null hypothesis we need an alternative hypothesis (often denoted by  $H_1$  or  $H_A$ ). If our conclusion is to reject the null, we can state our data support the alternative hypothesis.

It is here where the acorn analogy tends to fall apart unless we limit the direction the acorns can fall to two directions – we need to envision a straight line to keep the discussion going. Basically picture the chosen oak tree on a number line, the seedlings occurring on the number line in either direction from the tree, and all other mast-producing oak trees of that species in that stand lining up on that number line. In reality the other trees and seedlings can obviously be in any direction from the selected oak tree, but the concepts herein are best explained when we limit ourselves to a number line.

Recall our null hypothesis states that the seedlings originated from acorns from the chosen oak tree (mathematically think of this as an equal sign, so equal signs go into the null hypothesis). We can actually have three different alternative hypothesis paired with this null, and when performing a hypothesis test, only one can be chosen and used. One alternative hypothesis is that the seedlings originated from an oak tree to the left of the chosen tree (mathematically think of this as a less than

sign, this is known as a one-sided lower tail test). Another alternative hypothesis is that the seedlings originated from an oak tree to the right of the chosen tree (mathematically think of this as a greater than sign, this is known as a one-sided upper tail test). The final alternative hypothesis is that the seedlings originated from an oak tree on either side of the chosen tree (mathematically think of this as a not equal to sign, this is known as a two-sided test).

One role of the alternative hypothesis is to identify the DIRECTION AWAY from the chosen tree to draw our line (the distance of which is set by  $\alpha$ ) needed to make our comparisons. For a one-sided lower tail test, we would draw the line to the left of the chosen tree and consider seedlings (our data) to the left of that line as “rare” and seedlings (data) to the right of that line as “not so rare”. For a one-sided upper tail test, we would draw the line to the right of the chosen tree and consider seedlings (our data) to the right of that line as “rare” and seedlings (data) to the left of that line as “not so rare”. For a two-sided test, lines are drawn in both directions (left and right) and comparisons are made in both directions. Just be aware that for two-sided tests, the lines drawn in either direction actually have to be farther away from the chosen tree when compared to the line drawn if only a one-sided test in a given direction was used (the reason for that is tied up in what happens with the chosen  $\alpha$ -level in two-sided tests when translated into numerical probabilities).

In general, we put what we are trying to show in the alternative hypothesis. In the case of the example, we would be trying to show that the seedling (our data) most likely came from a tree to the left of the chosen tree (one-sided lower) a tree to the right of the chosen tree (one-sided upper) or from a tree on either the left or the right of the chosen tree (two-sided). Again, if what we observed is rare (low numerical probability) when we assume the seedlings (our data) came from the chosen tree (the null hypothesis), we conclude the null hypothesis is probably false and conclude the seedlings came from another tree (one to left, to right, or either the left or the right for a one-sided lower, one-sided upper, or two-sided test, respectively).

Here is a brief quiz to test your knowledge of the above concepts, this time using a more realistic hypothesis test. We want to see if fertilizing stands increase tree diameter growth over the current average of 0.4 inches per year. We setup and performed an experiment to collect data to evaluate diameter growth in relation to the 0.4 inches per year number. We want to perform this test at  $\alpha=0.05$ . With that  $\alpha$ -level and the condition we will correctly put in the null hypothesis, we find that the critical value is 1.645. First, try to write out the null and alternative hypothesis in words. Then, what would our conclusion be if our data led us to a calculated test statistic of 1.555? Here’s one more to try: using those same hypotheses, what would our conclusion be if instead of a calculated test statistic of 1.555, we found a p-value equal to 0.03? (Note there is no link between these two outcomes – they are completely independent questions.) The hypotheses and correct conclusion for both outcomes are presented at the very end of this article.

Now that we have a handle on null and alternative hypotheses as well as the hypothesis testing process (particularly how we assume the null hypothesis is true, base probabilities off of that, and conclude the null hypothesis is probably false if the chances of what we saw with our data are low if the null was

indeed correct), let's focus on why I used kept using the word "probably" when mentioning the null hypothesis was "probably false" (hello Mr. and Ms. Squirrel).

In the case where the seedling (our data) might have crossed the line on the ground, leading us to reject the null and state the condition in the alternative, we have to realize that there is always the chance that no matter how far away from the original tree our observed seedlings occurred, they still might have originated from that chosen tree (the squirrel factor if we will), we will just never know if this actually happened. Through our chosen  $\alpha$ -level, we are diligent to keep that probability low (that's why those  $\alpha$ -levels are typically 0.10 or less, with 0.05 being most common), but the chance always exist.

Likewise in the case where the seedling (our data) did not cross that line, leading us to fail to reject the null and state that we cannot say the null is false, there is always the chance that they originated from acorns from a different tree (the darn squirrel factor again).

In both cases, the "squirrel factor" might have led us to incorrect conclusions. Drawing an incorrect conclusion (compared to the unknown reality, often called "true state of nature" in statistical terminology) is known as making an error, and the two types described above have specific names. In the case where we reject the null hypothesis when we should not have, that error is called a Type I Error in statistical terminology, and for any given hypothesis test, the probability of making a Type I Error is  $\alpha$ . We should be able to infer that relationship from the acorn example.

In the case where we fail to reject the null when we should have (our seedlings [data] did not cross the line even though they came from a different tree), that error is called a Type II Error in statistical terminology, and for any given hypothesis test, that probability can also be calculated and is represented by beta ( $\beta$ ). Please know that even though  $\beta$  can be calculated, it is unfortunately not easily explained, so I will not attempt to do so here. Ideally one would also like this to be small, but the important concept here is that  $\alpha$  and  $\beta$  are mathematically linked and are in essence on either end of a tug of war rope. Pull on one end to make one of those smaller (reduce the probability of making that type of error) and the other one automatically becomes numerically larger (increasing the probability of making that type of error). That relationship is an unfortunate reality in hypothesis testing.

My students often have a hard time grasping the concept Type I and Type II Errors, but the following analogy works pretty well for them, so I will try it here too. We are trying to determine if our pants are on fire. Since in hypothesis testing you tend to put what you are trying to show in the alternative hypothesis, our alternative hypothesis is "our pants are on fire" and the null hypothesis then becomes "our pants are not on fire". We will conclude "our pants are on fire" if what we observe (our data) suggest the null hypothesis is probably false (lots of smoke versus little to no smoke if we will!). What the heck does this have to do with Type I and Type II Errors? It's simple really and here's the link. In this scenario, a Type I Error is saying "our pants are on fire" when in reality they are not, and a Type II Error would be saying "our pants are not on fire" when they actually are.

I have never had a student miss that concept after they have adopted that analogy, so hopefully it will always work for the reader as well. That example also introduces the notion that sometimes the error to

worry more about is the Type II Error (the probability of which is  $\beta$ ) as opposed to the Type I Error (the probability of which is  $\alpha$ ) that often draws the most attention.

I will now very cautiously raise and discuss one more concept that arises in hypothesis testing, but if anyone wishes to stop reading right here feel free to absolutely do so, it will get a little deep here compared to what appears above. That one final concept is the called the power of the test. The power of a test is the “ability or probability a test can lead you to reject the null and conclude the alternative hypothesis if the null hypothesis truly is false or incorrect”. Kindly go back and read that last sentence again – I mentioned it would get a bit deep at this point. Since probabilities are always numbers between 0 and 1, and since the probability of something and its opposite by definition have to sum to 1, we can actually come up with a numerical expression for the power of a test. Can you figure it out?

Let’s go back to the type I and type II error concepts and specifically re-evaluate the type II error concept (the probability of which is represented by  $\beta$ ). Recall that the type II error was saying “our pants are on fire” when they actually are not (concluding the alternative when we should not have). The opposite of that would be saying “our pants are on fire” when they actually are. The reader should kindly note that would be the correct decision in that case (saying the null hypothesis is probably false when in reality it is false). How can we represent that (correctly saying the null hypothesis is probably false given reality) numerically? Using the tidbits mentioned previously, hopefully we got it and see that correctly saying the null hypothesis is probably false (rejecting the null) when in reality it is false is numerically represented by  $1 - \beta$ . With that step, then, we have identified the formula for finding the power of a test as  $1 - \beta$ . (Do recall that calculating  $\beta$  can be difficult and not addressed herein, but the good news is that once  $\beta$  is determined, calculating the power of a test is pretty simple from there). In general, we want the power of a test to be relatively large (closer to 1 than to 0). We want to employ tests that have a relatively large probability of saying the null is probably false when in reality it is false!

Calculating power is one thing but the more important concept, at least in my opinion, is understanding the impact of low power – particularly when we fail to reject the null (we are unable to conclude the null hypothesis is probably false). If we fail to reject the null and the test had high power, I am generally quite comfortable with the results of the test. Under such a scenario, we were unable to say the null hypothesis was probably false even though the test has a relatively large probability of telling us it was false if that truly was the case. In other words, our data basically supported the null hypothesis and that is just how things probably are.

The more frightening scenario is when we fail to reject the null hypothesis of a test and the test had very low power. Under such a scenario we were unable to say the null hypothesis was probably false using our data, but this time we do not know if that is because that was truly the case OR because our test had a low probability of saying the null hypothesis was false EVEN IF IT TRULY WAS FALSE. I mentioned it would get pretty deep here and it surely has here, but hopefully the reader now understands why a test with low power and which leads us to conclude “fail to reject the null” does us no good whatsoever, again, it had very little chance of telling us the null was probably false even if it was false! (As an aside, I worry much less about power when I reject the null). The moral to this story is that the power of a test can be calculated, we should consider the power of a test during its planning stages and we should

never spend the time and money to perform a test that has little to no chance of ever telling us the null is probably false even if it is truly false.

So there we go. We have hit upon many of the key concepts of hypothesis testing basically by considering how far an acorn might fall from the tree, and then expanded the concepts a bit from there. That was not so bad was it?

(The previously promised answers appear on the next page)

Quiz answers:

Hypotheses:

$H_0$ : Average diameter growth of fertilized trees equals 0.4 inches per year (statistically written as  $\mu=0.04$  inches per year)

$H_1$ : Average diameter growth of fertilized trees is greater than 0.4 inches per year (statistically written as  $\mu>0.04$  inches per year)

$\alpha=0.05$ .

For this hypothesis test, we found a critical value of 1.645 (based on the statistical distribution applicable to the null hypothesis and our chosen  $\alpha$ ).

+++++

For the scenario where wound up with a calculated test statistic of 1.555, our correct conclusion would be:

*Fail to reject the null, not enough evidence to show the average diameter growth of fertilized exceeds 0.4 inches per year.*

Did you get that one correct??? Note that the average diameter growth in our data did in fact exceed 0.4 inches per year (inferred because the calculated test statistic is greater than 0 in this case, even though the actual diameter growth was not provided in the example) it was just not far enough away from the 0.4 number to say it was statistically larger – the acorn did not fall far enough away from the tree!

+++++

For the second scenario where wound up with a p-value of 0.03, our correct conclusion would be:

*Reject the null, data indicate the average diameter growth of fertilized trees exceeds 0.4 inches per year.*

Did you get that one correct??? Note that the average diameter growth in our data did in fact exceed 0.4 inches per year (admittedly harder to infer from the p-value approach alone) but in this case it was far enough away from the 0.4 number to statistically say it was larger – the acorn fell far enough away from the tree!

+++++