[MUSIC PLAYING]

JOHN WILLOUGHBY: Hello and welcome. This is John Willoughby, and this is a course on the Basics of Statistical Analysis for Use in Natural Resource Management. This is module four, Inferential Statistics Comparing Two or More Population Parameters. This is the last of four modules in this course. Hopefully you have viewed the first three. As I mentioned, this one will talk about the use of samples to compare two or more population parameters to see if the populations are the same.

And I'd like to thank my colleagues Dan Salzer and Marie Denn for providing some of the slides that I use in this presentation. So we'd like to compare two population parameters. This is a common situation in monitoring where we estimate a population parameter, for example, a mean or a total or a proportion at a particular site and in each of two years. And the two years don't necessarily have to be consecutive. They could be, for example, five years apart.

But we measure this parameter in two years, and now we want to compare those estimates to determine if the population has changed. And we'll discuss methods of determining this. And these methods, though, can also be used to determine if two populations that are spatially separate are different in their population parameters.

I'm going to emphasize the use of methods for detecting change in the slides that follow, but just remember, these methods can be used to detect any difference between populations, not just those caused by change over time. So here, we have an area. This is a plant example. So say we have a population of a rare plant species in a large area. So here's a large macroplot containing that plant species. We've delineated this plot around the plant population for ease of sampling. And we've decided that we're going to sample this macroplot using rectangular quadrats.

So here's the area overlaid by a population of these quadrats. Remember, that in a statistical parlance, a statistical population is a collection of all the possible sampling units that could fit in the population without overlap. And so here are sampling units or quadrats in our statistical population is that set of quadrats that we could put in that population. But, we're using these quadrats to make inferences about the biological population, the statistical population is the set of quadrats.

So here, at two different time periods, we take random samples of 25 quadrats. Now note that these are temporary quadrats, so we've taken a random sample at time one, but we've taken a new random sample at time two. And so the samples are independent. Granted, we're sampling the same area because that's the population for which we want to make inferences, but because we've taken a different random sample at each time, the samples are independent.

What we want to know is if the population has changed from time one to time two. So let's say that our sample mean at time one is 17.44 plants for quadrat. And then at time two, our sample mean is 20.72 plants per quadrat. So based on these samples, it appears that the mean has increased from time one to time two. But remember. there's inevitable sampling error associated with any sample, and so it could be that the population hasn't changed at all and simply by chance we obtain two samples with different means. So that's sampling error.

And so when we were dealing with a single population parameter, we quantify this inevitable sampling error with a confidence interval. To compare two population parameters on the other hand, we're going to use statistical tests. It turns out you can also use confidence intervals for this purpose, but we're going to emphasize statistical tests in this module. So when comparing population parameters, there are two types of possible sampling errors. A false-change error and a missed-change error. And false-change errors are also called Type I errors and missed-change errors are also called Type Ii errors.

So let's look at these in a little more detail. So here are the possible errors when you're monitoring for change. And again, these errors also apply if we're comparing two populations that are spatially separate from one another. But again, we're going to emphasize monitoring in this module. So we'll talk about the possible errors when monitoring to detect a change. So here's the true situation. No real change is taking place or there has been a real change.

Now, we're using the samples to estimate this, so we won't know whether a real change has actually taken place or not. We're going to use the sample to tell us whether there's been a change. And so if the monitoring system detects a change and, in fact, there's been a real change, then no error has been committed. On the other hand, if monitoring detects a change and no real change has actually taken place, then that's a false-change error. That's one of the two types of errors that can take place.

And this is also called a Type I error, as I mentioned, and it's also uses the symbol alpha when we set a value for the false-change error rate that we're willing to live with, which we'll talk about. So if the monitoring system detects no change, and in fact no change has taken place, then again no error has been committed. On the other hand, if the monitoring system detects no change but there has really been a real change, we just missed it, then that's a missed-change error. That's the second kind of error that we can make.

So you can control the magnitude of the false change error that you're willing to accept by specifying that level prior to sampling. So that threshold, then, is used in a significance test to determine whether a change has occurred. So we're going to focus on the false-change area for the next several slides because that's the error that's used in statistical tests. So the false-change error is called either alpha or the Type I error when you look at the literature. But in a monitoring context, we prefer to call it false-change error because it just seems clearer than just saying Type I error what we're talking about.

And it's the probability of falsely concluding the true population parameter has changed when, in fact, it hasn't. It was just sampling error. We just happened to take two samples that were extreme enough from the same population of the population that hadn't changed. And they were extreme enough that the tests showed us that there had been a change when, in fact, there wasn't. And so we're going to focus, as I said, on this type of sampling error for the next several slides because it's the only type of error used when conducting significance tests.

Later, we'll see that we also need to worry about the missed-change error, which is the probability of missing a true change should it occur. So we need to determine what false-change error rate we're willing to live with before we do the sampling. And the most common rate you'll see in the scientific literature is 0.05. But this is often too low for natural resource monitoring programs. As we'll see later, the lower the false-change error rate, the higher the missed-change error rate. And missing a true change is often more important than falsely concluding a change has taken place when it hasn't.

So missed-change error rates are important, too, and later we'll see how the false-change error rate and the missed-change error rates are related. But the bottom line is that if making the false-change error rate smaller, makes it more likely that you'll make a missed-change error.

So let's explore how the false change here is used in practice. So I take a sample of 25 quadrats from a population in year one. The sample mean of the first year is 20 plants per quadrat, and the standard deviation is 10 plants per quadrat. In year two, I take another sample of 25 quadrats, again, using different random coordinates, so these two samples are independent. And so this is from the same population. And so the second year's mean is 25 plants per quadrat. And the second year's standard deviation is 10 plants per quadrat.

What I want you to determine is whether the fact that the sample mean has increased from 20 to 25 indicates that the population has changed from year one to year two. Remember, it could be that these samples came from the same population. It's just that by chance I got 20 the first time I sampled and 25 the second time. So we do that by estimating the sampling distribution. So we call it a sampling distribution of means can be estimated based on a single sample.

The distribution is centered on the sample mean and the standard error is calculated by dividing the sample standard deviation by the square root of the sample size. So remember this formula that we introduced in module three. The standard deviation divided by the square root of the sample size. So in our example, the year one sample mean was 20 and the standard deviation from year one was 10 and the sample size was 25, so here's our calculation of the standard error for year one. And it's two. So the standard error is two plants per quadrat.

So using the mean and standard error, we can now estimate the sampling distribution of sampling means for year one. And here is the normal curve. Remember, the normal distribution is described by two parameters, the mean, which is 20, and the standard deviation of the sampling distribution, which is the standard error of two. What's the probability of getting a sample mean as large or larger than this one? So 20. Well, since the distribution of the sampling distribution of means is centered on a mean of 20, it obviously wouldn't be a surprise to get a sample mean close to 20.

It turns out that the probability of getting a mean of 20 or larger is 50% or 0.5. And recall that the standard deviation of a sampling distribution is the standard error. So that we take the standard deviation of our one sample, and then we calculate the standard error from that standard deviation. And then that approximates the standard deviation of the sampling distribution, which is shown here.

So in this case, the standard deviation of the sampling distribution is two, two plants per quadrat and the standard error is the same thing, so two plants per quadrat. So what's the probability of getting a sample mean that's larger than this one? This is 22. Well, it's not going to be as likely as getting one near to the center, but it's still rather likely. The probability is about 16% of getting a value as large or larger than this because you can see there's still quite a bit of area in this tail from this point on.

And remember, we talked about that most of the data, or 68% of the data, is contained between plus or minus one standard deviation of the mean. And so this is one standard deviation. This is the standard error. It's the standard deviation of the sampling distribution. So 68% of the values fall between here and 18 on the other side. And so this is half of that, half of that 32% that fall outside. So 16% falls outside this way and 16% falls out that way. So still not unlikely to get a sample mean of 22 from that same population, the population with a mean of 20.

Well, what about the probability of getting a sample mean as large or larger than this one? So now it's 24. Well, now we're getting out near the tail of the distribution. And we'll see that this is two standard deviations away from the mean. And so the right tail probability here, the probability of getting a value as large or larger than 24, is only 0.023, so only a 2% probability about.

And note that our year two sample mean in our example is 25, it's here. So we know, even though we haven't looked at the actual probability associated with 25, we know that the probability of getting a value that large has to be less than 0.023 because the probability of getting a 24 is 0.023. So what about a sample mean of 26? What's the likelihood of getting that from this population? Well, not very likely. It's way out of the tail. And so now, the right tail probability is only 0.001. So very unlikely.

And so we could go even further. We could say what's the sample mean of 28? So now we're four standard deviations away from the mean, and that likelihood is very minuscule. So if we see values like this in the tail, it's quite likely that those values don't belong to this sampling distribution with a mean of 20. They likely belong to a different sampling distribution and, therefore, a different population.

So they may belong to sampling distribution with the means centered on 25, so these three squares here. These values that were two, three, and four standard deviations away from this mean probably, or may, belong to a distribution that centered on a mean of 25 somewhere in here.

And so that brings us to null hypothesis significance testing now that we've seen how we can compare sampling distributions. So what we know about the use of sampling distributions can now be applied to see if two samples come from the same or different populations. So we start with a null hypothesis. And the null hypothesis is typically shown with an H and a subscript of zero. So the null hypothesis is that the two samples come from the same population. And if we reject that null hypothesis, then we accept the alternative hypothesis. And that alternative hypothesis is that the two samples come from different populations.

In a monitoring context, if we fail to reject the null hypothesis, we conclude there's been no change in the population between time periods. On the other hand, if we reject the null hypothesis, then we accept the alternative hypothesis and conclude that the population has changed. So here, we see the null hypothesis for our example displayed graphically. So according to the null hypothesis, our year two sample mean of 25 shown here comes from the same sampling distribution as the year one sample mean of 20.

So the null hypothesis says that this population has a mean of 20. And the null, then, says that this 25 also comes from this population that's centered with a mean of 20. The alternative hypothesis says that the two populations are different. So if we have sufficient evidence to reject the null, then we accept the alternative hypothesis. And we conclude that the year two sample comes from a different population. And so in a monitoring context, this would mean that the population has increased, the true mean population of this plant species has increased.

So if we assume that the alternative hypothesis mean is a true population mean, then the probability that a mean of 25 could come from a sampling distribution centered on the year one mean of 20, that probability is very low. It's only 0.006. And so the probability of a true year two mean of 15, so five plants less per quadrant than the mean, is equally low. And so typically, our tests are going to be what are called two-tailed tests. And so we're interested in the two-tail probability of getting a true year two mean. Either five plants per quadrat or more greater or less than the year one sample mean.

So we need to add those to the left tail and the right tail probabilities together, and we get a two-tailed probability of 0.012. In statistical parlance, we've just performed a one sample hypothesis test. We compare the sample mean to a hypothesized value. We compared our sample mean of 20 to a hypothesized value of 25. But we need to take into account the fact that the year two sample mean is itself an estimate with variability around it.

And so the way we do this is by combining the difference between the two simple means and the standard errors of both of the two sample means into a single test statistic. And that test statistic is the t statistic. So the t statistic is used in a t test to determine whether to reject the null hypothesis. And so here are the results of a t test of our example data using the statistical program R. This is the print out from conducting a two sample t test in R.

And there are three things we want to take note of here. One is the calculated t value, another is the degrees of freedom, 48 in this case, and another is the p-value. This is the probability value. So if we concluded based on these data that these two populations are different, there's a 0.083 chance of us being wrong. In other words, there's a 0.083 chance of us committing a false-change error. So that's what the p-value here means.

Notice, that the p-value is always positive, but the t value you can be either negative or positive. And it just depends on which sample means you subtract from the others. So we subtracted, in this case, the year two mean, which was 25, from the year one mean. So we ended up with a change of minus 5. So when you have a minus difference, the t value is negative. We'll see why that is when we see the formula.

But it's the absolute value that matters in a two-tailed test. So it doesn't matter whether this is negative or positive, we take the absolute value of t. So another way to say it is the p-value is the probability of getting a t value in absolute terms as large as the one observed when the null hypothesis is true. So what does this all mean? How was the t value calculated? How was it used to calculate the p-value? And finally, what are degrees of freedom? Well, we'll talk about all of these things.

So an independent sample t test, which is what we just had R run for us, is for two sample means. And it examined-- so here's the t statistic, the calculated t statistic. And it examines the difference of sample means over the standard error of the difference of sample means. And here's the actual formula. So here's our means, here's the year one mean and here's the year two mean. And in this case, I subtracted the year two mean from the year one mean. It wouldn't make any difference if I did it the opposite way. The t value would simply be positive in the example we have. But again, it's the absolute value that matters.

So in this case, I subtracted the year two value from the year one, the year two mean from the year one mean. Here's the standard error. Remember, the standard error is the square root or it's the standard deviation over the square root of the sample size. Well, here, I actually have variances in the numerator, so that's the standard deviation squared and dividing by n. And this is the standard deviation of the first year divided by squared divided by the sample size of the first year, and here's the same thing for the second year. And we take the square root of that.

So that's how the t the calculated t value is derived. And so if we apply that to our example, and again we get a negative 1.77 as our calculated t value, again, because we subtracted the larger mean from the smaller one. And so now, we're going to see how we evaluate this t value of 1.77. So we need to look at where that calculated t value falls in the appropriate t distribution. And what do I mean by the appropriate t distribution?

Well, remember we talked in module three that t distribution curves differ depending on the number of degrees of freedom. So degrees of freedom, and we'll talk about how to calculate that when we're comparing two means here in a minute. But remember, that for lower degrees of freedom that you have thicker tails and a flatter curve than you do for higher ones, and the normal is the dashed line.

So here's how we calculate degrees of freedom for an independent sample t test. Degrees of freedom are the sample size from the first year minus 1 plus the sample size from the second year minus 1. And another way of saying that is you could say, well, it's the sample size the first year plus the sample size the second year minus 2, you get the same result. So for our example, our degrees of freedom are 25 minus 1 plus 25 minus 1 is 48, 48 degrees of freedom.

And so here is a t distribution curve for degrees of freedom equals 48. And so if this value, if our calculated t value were zero, well, we'd be right here. The t distributions standardizes these differences, and so a calculated t value of zero essentially says the populations are the same. So we clearly wouldn't conclude that there's any difference.

But our calculated t value is minus 1.77. And we can see that that's quite a ways from zero, is it far enough away for us to reject the null hypothesis? Well, that's where our knowledge of the t distribution comes into play because we can look at this calculated t value, minus 1.77, and we know how much area is to the left of the curve. And it turns out that the probability of getting a value of minus 1.77 or larger, in absolute terms, so the probability of getting a value here or further out is 0.04.

And if we look at the other side, which we need to because this is a two-tailed test, we're interested in detecting change in either direction. So the probability of getting a t value of 1.77 or greater is also 0.04. So we would add those two together and the total two-tail p-value is therefore 0.08. So that's our calculated p-value. And what that means is there's an 8% chance of obtaining an absolute calculated t value as large or larger than the one that was observed if the null hypothesis is true. In other words, the two samples come from the same population.

So there's an 8% chance of making a false-change error if we say these two populations are different. And then the question becomes whether that is that value, that calculated t value, that 8% chance, is small enough for us to reject that null hypothesis and accept the alternative hypothesis that the two samples came from different populations. And that depends on the p-value we designate as our threshold before we conduct this study.

So how small does that p-value need to be before we say that the two populations are different? Well, this is the threshold p-value, and that's the false-change error rate that we specify, and we'll talk about where we specify that later. It turns out we need to develop a sampling objective. But the thing to remember right now is that we need to specify the threshold p-value before we sample, before we go out and do the study. We need to specify our false-change error rate.

So the most common threshold p-value is 0.05. It's the one you commonly see in the scientific literature. But in many monitoring situations, a threshold p-value a 0.05 results in insufficient power to detect changes that may be biologically important. And for that reason, a threshold p-value of 0.1 may be more appropriate. And so in this study, this example study, we selected that level before we initiated the study. So our threshold p-value or false-change error rate that we're willing to live with is 0.1.

So now, we compare the calculated t value, which is the t value calculated from the formula I showed you, and from our actual data, from the means and standard errors of our actual data. We compare that to a critical value of t to determine if there's sufficient evidence to reject the null hypothesis and conclude the two populations are different or the two means are different. And therefore, come from two different population.

And the critical t value depends on the threshold p that you set prior to sampling. So a table of critical t values or a t table can be used for this purpose. So here's an example t table. And so let's look at-- this is similar to one you'll find in a textbook. And so these are critical values of the t distribution, and let's examine what all this means.

So as we mentioned in module three, is the Greek symbol nu, and it is the degrees of freedom. So we read degrees of freedom in this column. But we need 48 degrees of freedom. So we'll have to go to the next page to get there. We have 48 degrees of freedom that we'll get to in a minute. So here's the alpha value, remember alpha is the same as the false-change error rate we said which I've also called the threshold p-value. So all of those mean the same thing.

So this is a two-tailed test. See this 2 in parentheses up here, alpha 2, these are the alpha values for a two-tailed test. We'll talk about one-tailed test as well and that would be this row here. But for now we're interested in the two-tail alpha value of 0.1. All right.

So we then go down to our degrees of freedom, we have 48. And we see that the critical t value for 48 degrees of freedom and an alpha of 0.1 is 1.677. And so we can compare that t value to the calculated t value. And if the calculated t value is larger then this critical t value, then we reject the null hypothesis. And in this case, it is. So our calculated t value was 1.77, the critical t value is 1.68, and we therefore reject the null.

And let's look at this graphically. Here, we plotted rejection regions. Again, this is the t distribution for 48 degrees of freedom. And we have a sample size of 25 in each year. And so here's the critical t value that we just saw from the t table. So any t value that falls out in these pink areas on either side of this t distribution will reject the null. So these are called rejection regions.

So in our case, our t value was minus 1.77, and that falls outside. It falls in this rejection region and outside of the 1.67 value. And so we reject the null hypothesis. And here's our p-value associated with that value of t, 0.04 on this side. Remember, we're also worried about this side, so there would be a 0.04 on this side, too, for this rejection region.

But what if our acceptable false-change error rate was 0.05 instead of 0.1? So now alpha is 0.05. Well, if that were the case, now the critical t value is larger. It's 2.01 instead of 1.68. Right? So now are calculated t value has to be larger. The absolute p-value has to be larger. So let's look at this in our t table.

So here, instead of looking in this column, the alpha 0.1 column, now we look in the alpha two-tail 0.05 column, and that's where we get our 2.01. And now, the calculated t value of 1.77 is less than the critical value. So we fail to reject the null in this case. And we see that here. So again, this is what we've set a false-change error rate of 0.05, alpha 0.05. Now the critical value is larger and our calculated t value no longer falls in the rejection region. So we would fail to reject the null hypothesis in this case.

So one of the things I want you to recognize is that there's this relationship between the test statistic, which is t in this case, and the p-value. And it's an inverse relationship. So as the t statistic gets larger, the p-value decreases. Here's the t table with just 48 degrees of freedom. And you can see that the t value gets larger here on this row, the p-value gets smaller shown up here. And this relationship follows for any test statistic. We'll talk about the Chi-Square statistic a little bit later.

And as it gets larger, the p-value gets smaller. So that relation holds regardless of the test statistic. As the test statistic gets larger, the p-value gets smaller. And just a graphic illustrating that, so here's our t distribution. And as that when the t value is 1, the p-value is 0.162. When the p-value is 2, its 0.026 and so on. So as the absolute t value goes up, the p-value goes down.

So let's talk about two-tailed versus one-tailed t test. We've already talked about the two-tailed test. So here are null and alternative hypotheses. The null is the population mean, it's not changed, and the alternative hypothesis is that it has changed. So if we fail to reject the null hypothesis, we can conclude there's been no change in either direction in the population mean. In other words, the population has remained unchanged.

But we could reformulate that hypothesis. We could have a one-tailed hypothesis instead. So we could say that our null hypothesis, in this case, is that the population mean has not increased. And the alternative to that would be that it has increased. But recognize that here, if we fail to reject the null hypothesis, we can conclude only that there's been no increase in the population mean. We can't make any conclusions as to whether the population mean has stayed the same or decreased.

Well, why would we want to do a one-tailed test in the first place? Well, if you're truly only interested in detecting change in one direction, a one-tailed test is more powerful to detect that change. And why? Well, it turns out the t statistic doesn't have to be as large. So here's our t table again. Again, just for 48 degrees of freedom, but remember, we use the two-tailed alpha value up here. So we used an alpha value a 0.1. And so here was our t value associated with that alpha value.

But if we're interested in detecting change in just one direction a one-tailed test, here is 0.01 for that. And you can see now that the t value, the critical value of t is much smaller. So now, the t value to reject the null hypothesis if our threshold p is 0.1, it only has to be larger than 1.299 for a one-tailed test as opposed to being larger than 1.677 for the two-tailed test.

And if we look at this graphically, we see that here we've got a false-change error rate of 0.05. We determined that before we did our sampling. But if that were the case, the critical value is now only 1.68. Remember before, it was 2.01. And the entire alpha now is apportioned to the pertinent side of the distribution, because we're only interested in detecting change in one direction, in this direction. We don't have to have a rejection region up in the upper tail because we're not interested in detecting change in that direction.

And so here's the rejection region again. And so now the probability of getting an absolute p-value as large is 1.77 is only 0.04. If our false-change error rate, if we'd set that at 0.05, this is smaller than that, and so we would reject the null hypothesis in this case.

So the disadvantage of one- tailed test, I mean, it's easier to detect change, why shouldn't we use them all the time? Well, in most monitoring situations, you're really interested in detecting change in either direction. So even though you might take management action only if the population of a sensitive species has decreased, we also want to know if it's increased. And several recent authors, here's a couple of citations, believe that one-tailed test should be the exception rather than the rule. We'll give an example next when a one-tailed test is entirely appropriate, in fact, desirable. And here are those references.

So here's a situation where a one-tailed test makes sense. So here, this is a common situation in restoration. We have a treatment area, an area that's been affected by some management action or needs restored for some reason. And we want it to look like an area that's with similar soils and similar aspect, an area that we think this treatment area should look like this reference area.

And so we take some action. We maybe fertilizer, plant some plants, or we do something out there to try to make this area look like the reference area. So traditional approach of let's say we're looking at a vegetation cover, we'd like the cover of a particular species or maybe it's total cover, whatever it is, we would like the cover of this treatment area look like the cover of the reference area.

And so the traditional approach to monitoring this might be the null hypothesis, basically, is that the cover of the treatment area is equal to the cover of the reference area. Right? So we sample, and we compare the means cover for both of these areas. And we determine whether those means are different enough for us to reject this null hypothesis and accept the alternative hypothesis that the treatment cover is not equal to the cover in the reference area. But there are problems with that approach.

A low-power monitoring design would fail to reject this null hypothesis, even when the populations are quite different. And we'll talk about power more a little bit later. But just think about maybe we just put two or three sampling units out there. Well, it's going to be really difficult to detect change between two areas if your sample size is really low. And yet, if we fail to reject the null, we would assume that we've got some standard we'd like to meet. And the standard is we want the treatment area cover to be the same as the reference area cover. Well, we have to conclude that was the case even if it weren't.

So the other issue is a high-powered design might reject the null hypothesis, even when the populations are sufficiently similar. Now, the one thing to recognize is that the null hypothesis is sort of a straw man. You never expect the cover to be exactly the same. So what we're saying is we want the cover to be close enough. But if our sample sizes were really large, we might say, well, this area is 2% different than this area in cover. Well, is that really different biologically? Probably not.

And the other issue is the approach doesn't explicitly state what level of difference between the treatment area cover and the reference area cover is acceptable. So here's the alternative, here's a better approach. Here, we change the burden of proof, essentially. And so now, our null hypothesis is that the cover of the treatment area is less than or equal to 0.8 times the cover of the reference area. And the alternative hypothesis is that the cover of the treatment area is greater than 80% of the cover of the reference area.

Well, what does this do? Well it explicitly states what the success criterion is now. So we want to cover the treatment area to be greater than 80% of the cover of the reference area, and that's shown in this alternative hypothesis. So now, a low-power design won't result in our rejecting the null hypothesis. So we'll have to conclude that we haven't met that success criteria. The treatment area has not, in fact, more than 80% of the cover of the reference area. So now only a high-powered design will allow the rejection of the null hypothesis and the true treatment value needs to be more than 80% of the reference area covered.

And the analysis here is a one-tailed test because, again, we're interested in detecting change in only one direction. So it would be a t test if the data are continuous and a Chi-Square test if the data are binomial. And we'll talk about a Chi-Square test here later. So this is a situation where a one-tailed test makes perfect sense, and in fact, is desirable. And so this is also termed a one-tailed bioequivalence test or a non-inferiority test. And a good reference for this is a chapter in this book by Manly.

So we've talked about one and two-tailed test for the independent sample t test. Now, we're going to talk about a paired sample t test. So the independent sample t tests are appropriate to compare to samples collected using temporary sampling units. In other words, with separate random samples taken in each year. But what if we use permanent sampling units? So we randomly locate some number of sampling units of quadrats, belt for belt transects at time one. And then we permanently mark those sampling units of measure the same set of sampling units at time two.

So here, we show this graphically. So a time one, we take a random sample of 25 quadrats, and we mark these. We mark the corners, for example. And then we go back and we measure the exact same set of quadrats at time two. So the two samples are no longer independent. Right? Because the sample at time two is the same as a sample at time one. So they're said to be paired, the samples. You'll also see the terms correlated and dependent. They're correlated samples, their dependent samples.

Because we measure each sampling unit twice, we now have a data set consisting of pairs of observations. And in our example, since we have an n of 25, we have 25 such pairs of observation. It's not appropriate to apply an independent sample t test to those data. Instead, we use a paired sample t test. And paired tests, such as the paired t test, are often much more powerful than an independent sample t tests in detecting change.

So if you're setting up a monitoring study, it always behooves you to think about at least whether you should have permanent sampling units or not. And in the next slide shows how, in some cases, the paired sample t test would be much more powerful than an independent sample t test. So here we have two samples of 10 cover values, each taken using line intercept transects. And the common values of both samples, you'll range from below 10% up to 80% or higher.

Without even doing an analysis, we can see there's no way an independent sample t test would show a significant difference in coverage between these two time periods. But these samples aren't independent. Instead, we took a random sample at time one, and then we permanently marked the transects both the beginning and end point and intermediate points. And so then the next time, we measured the cover along those same line intercept transects. And here were the results.

And we can see here now that the cover of every one of these transects went down except for this one. And so because it's now pretty clear that there may well have been a decrease in cover between time one and time two. It wouldn't have been at all obvious from this data set. But now that we look at the data paired, it's pretty clear there's a decrease. And in fact, if we run an independent sample t test, which we shouldn't because those data are paired. But if we did, our calculated t value would be 0.6, and the p-value would be 0.55, not even close to significant.

So threshold p, we might have threshold p-value of 0.1. Well, the calculated p-value would have to be less than 0.1, and it's much greater than 0.1, in this case. So we would say there's no significant difference. But because the data were paired, we're going to do a paired t test. And the paired t test ignores the between transect variability and looks at the differences between time one and time two for each of the transects. And then it determines whether the mean of that set of differences is in 0, in which case we wouldn't reject the null. So let's look at the table to make this clearer.

So here are values at time one and time two. And again, we're not going to use either one of these data sets. We're not going to calculate the mean and the standard deviation of either of these. Instead, we take the difference in cover between time one and time two, and we take the mean of that set of differences. And then, we take the standard deviation of that set of differences and divide the standard deviation by the square root of n, which is 10 in this case, and we come up with a standard error of 1.98.

And here is the formula that we use for a paired t test. The calculated t value is the mean of the differences here divided by the standard error of the differences. And so our t value, in this case, is minus 3.34 and the calculated p-value associated with that t value is only 0.009. So we reject the null hypothesis, and we accept the alternative hypothesis that the population has changed.

So we were able to do that because we used a paired t test, and used permanent sampling units. Took a random sample the first year, marked them, and came back and measured the exact same sampling as the second year.

Here's an R printout of that example, paired t test it tells us, and here's the t value, here's the p-value. And remember, nine degrees of freedom, because it's the number of differences. Even though we had sample size of 10 the first year and a sample size of 10 the second year, we measured 20 line intercepts over the two year period. It's the set of differences that goes into the degrees of freedom. So it's n differences, there were 10 of those, minus 1.

We could also look at a 90% confidence interval. And notice that the confidence interval doesn't include zero. So if it doesn't, then that's another way of concluding that we should reject the null hypothesis. So it turns out that the higher the degree of positive correlation between the pairs of sampling units, the more powerful the test. And you can measure this with a correlation coefficient. The closer that coefficient is to one, which is perfect correlation, the more powerful the test.

You can also apply a paired sample t test even when there's no correlation. But recognize that it won't be as powerful as an independent sample t test. And the reason is that independent t tests will have more degrees of freedom. So if it were independent, we did an n of 10 the first year and an n of 10 the second year, we'd have a sample size of 20 and 18 degrees of freedom. But because the paired t test looks at only the differences, our degrees of freedom there are only nine.

So automating the t test. Well, every statistical package will calculate a t test. It's one of the most commonly used statistical tests. And we've already seen output from R for both independent and paired sample t test. But Excel will do these. There's a free data analysis tool pack add-on in Excel that will calculate both independent and paired sample t tests. And they'll also do one-tail and two-tail t tests. And there's instructions available on the web to help you conduct t tests in R or Excel, so you should be able to find those via a web search.

One further thing to mention here is that statistical packages, including R and the tool pack in Excel, allow you to do an independent sample t test assuming unequal variances. This is also called the Welch test or Welch approximation. And you should always choose that option. The regular independent t test assumes that both populations have equal variances. If the sample variances differ, the Welsh procedure lowers the degrees of freedom to account for the unequal variance. And that increases the p-value, so it's going to make it more difficult to detect change, but it's more reliable in situations where variances differ.

And in natural resource data, variances often differ between population. So you should always use that correction. If it turns out the variances are the same, the Welsh test returns the same p-value as the regular test. But note that this assumption of equal variance doesn't apply to the paired t tests. The paired t test only looks at the differences between the pairs and, therefore, there's only one variance is the variance of the set of differences.

So in module three, I introduced this finite population correction factor. So if you've sampled more than 5% of the population, you should apply the finite population correction to the results of the t test. So here's an example. We got a t statistic from a t test, and this could be either an independent sample or a paired sample t test. And our calculated statistic was 1.645. And in each of two years, you sampled n equals 26 out of a total of big N equals 100 possible quadrats. So you sampled 26 out of the total possible number of quadrats of 100 possible quadrats.

So you adjust the t value that you got from the Excel or R or whatever statistical package you use by the finite population correction. And so here is the corrected t prime. And that's t divided by the square root of 1 minus small n over big N. So in our case, our t value was 1.645, and then we divide that by the square root of 1 minus small n 26 divided by big N 100. And it increases the t value.

So now, our calculated t value is 1.9. And remember, the higher the t value, the more likely it is we're going to reject the null hypothesis. So the more Power we're going to have to reject the null if, in fact, the null is not true. So the problem is the R or Excel gave you a p-value associated with the original t statistic of 1.645. You now need to adjust that p-value to account for the fact that your t value is now larger. And we provide an Excel workbook-- here's the title of it-- on the National Training Center website that allows you to plug-in your newly calculated t value and get the p-value associated with that.

So far, we've looked at t tests which are the tool of choice to compare the means of two samples involving continuous data. And continuous data, for our purposes, are both real numbers, numbers that have decimals, and integers, whole numbers. And that's fine, t tests are great for samples involving continuous data. But t tests aren't appropriate for binomial data.

So recall binomial data consists only of zeros and ones. So we examine a set of sampling units, for example, for a particular trait. Each sampling unit either possesses the trait, it's a hit, or it doesn't, it's a miss. And it turns out that for a binomial data we're going to use something called the Chi-Square test. And again, this is for two independent proportions. So here are some examples of where we might use a Chi-Square test, we might have binomial data.

Is the proportion of deer infected with chronic wasting disease different between two management areas? Or has the cover of bare ground measured using randomly positioned point intercepts changed between time one and time two? Has the frequency of the weed Medusahead measured in randomly positioned quadrats declined following herbicide treatment? Note the samples must be independent. So the samples would consist of a different set of random points, quadrats, or individuals. We wouldn't use permanent sampling units.

So here's an example of Chi-Square analysis. We randomly position 400 quadrats in the same area at two time periods. And we use a different random sample at each period. We determine whether each quadrat contains the plant species of interest. And we want to know if the frequency or proportion of quadrats containing the plant is increasing or decreasing over time.

Prior to sampling, we decide to accept a false-change error rate of 0.1 as the threshold for significance. And remember, this is the same as saying alpha equals 0.1 or that are threshold p-value is 0.1. So we put our results into this two by two contingency table where the columns are the two years and the rows represent whether the plant was present or absent.

So in year one, we can see that 123 of our 400 quadrats contain the plant of interest. 207 didn't. In year two, 157 quadrats had the species of interest, so that increased. And the number of quadrats without the species decreased to 243. It's important to realize that the numbers in parentheses are frequency of occurrence in year one and year two and in the last year for both years combined.

The Chi-Square test is conducted on these actual numbers of quadrats or points or individuals, whatever the sampling units are, and not on the proportions. But as we'll see in the next slide, the proportions are used to determine the number of quadrats or points to be expected under the null hypothesis. So this is the two by two contingency table. So now, we need to calculate the values that would be expected if the null hypothesis were true.

So if the null hypothesis of no change is the same as saying the true proportion is the same in both years, the true proportion. So if the true proportion is the same, then the values obtained in each year in our samples can be considered estimates of that same proportion. Therefore, we can use the total frequencies over here on the right hand column here, under the total, to determine the expected values under the null hypothesis.

So we can see that 35% of our 400 quadrats would be expected to have the species present, and 65%, again, the average of these two values here would be expected not to. So that's what we see here. So we multiply that 35% times 400 and get 140. And then we multiply the 65% times 400 to get 260 quadrats. So these are the expected values. We would expect 140 plants in year one, to have the plant, in the same in year two, and 260 of those quadrats to not have the plant.

So these are the expected values under the null hypothesis. And here's our formula. So the Chi-Square value, the Chi-Square statistic, is the sum of the observed minus the expected values squared over the number expected. So let's look at how that is applied to our example. So here is the sum of the observed and expected. So 123 were observed in the first year, but we expect 140. We square that, divided by 140. That's the number of expected. The same with the absent. 277 actually absent, but we expected 260. We square that value, that difference, and divide that by 260, and we do that for the other two situations.

We end up, after doing this math, with the Chi-Square statistic of 6.352. And we compare that calculated Chi-Square statistic of 6.35 to a table of critical values of the Chi-Square statistic. And you can find one of those in appendix five of the Measuring and Monitoring Plant Populations book. And we see if the value is sufficiently large to be significant. An important thing to realize here is that degrees of freedom are calculated in a different way from the t test.

So degrees of freedom, or nu, for a contingency table, or calculated by taking the rows minus 1 times the columns minus 1. So we had two rows and two columns. So these values are 1 and the degrees of freedom are simply 1 for a two by two table, and that will always be the case. So now, we go to a Chi-Square table, very similar in many respects to the t table. And we go over to our threshold p-value, which we said was 0.1, our alpha value. We have one degree of freedom, so we use this row. And we see that the critical Chi-Square value is 2.7.

And we then compare our calculated Chi-Square value, which was 6.35, to that critical value, and we see that it's larger. And because it's larger, we reject the null hypothesis and conclude that the populations are different. We can look at this graphically. Here's a Chi-Square distribution with one degree of freedom. Notice that because we square these observed and expected values, the Chi-Square distribution can only take positive values. And so the distribution really has only one tail to evaluate significance unlike the t distribution, which had two two-tails.

So just like we showed with the t distribution that we can plot a rejection region here, and we see that here's our critical Chi-Square value of 2.7. So any calculated Chi-Square value that is larger than that, that falls out in this rejection region here, the result will be that we'll reject the null hypothesis. So this is a two-sided test. Despite the fact there's only one tail in a Chi-Square distribution, you can still perform one-tailed and two-tailed tests. Although, it's probably better to say two-sided and one-sided in the case of the Chi-Square distribution since there's only one tail.

But in this case, this was a two-sided test with a threshold p-value of 0.1, so this is the rejection region for 0.1. If we had an objective false change error rate of 0.05, then our Chi-Square statistic, the critical statistic, would be higher, 3.84. And now, we'd have to have a Chi-Square value greater than 3.84, again, for a two-sided test. For a one-sided test, again, we'd be back at the critical Chi-Square value for a one-sided test at alpha equals 0.05, which is our objective here, is the same as for a two-sided test at alpha equals 0.1.

Again, if you have good reason to use a one-sided test, you'll have more Power to detect change in that one direction. And here, we just show that if you upped your objective false-change error rate to 0.1, then the critical Chi-Square statistic is even smaller. So you're more likely to have a Chi-Square value larger than that and end up in the rejection region here where you reject the null hypothesis.

So the Chi-Square test can be used only for independent samples, again, involving binomial data. But what if you had quadrats, for example, that you randomly located in an area of interest, and you permanently marked them and went back to the same quadrats? Well, you'd have permanent sampling units in that case, and the Chi-Square test would not be appropriate.

And in some cases, you might be able to measure individual organisms at two different times. In that case, McNemar's test would be used. It uses a two by two contingency table, like this Chi-Square test, but the data that you enter into the contingency table differs, and so let's take a look.

So let's say that we had 100 quadrats in a macroplot in each of two years. Our threshold p-value is 0.1. And in year one, 60 quadrats have species X in them. But when we went out in year two, only 50 quadrats had species X in them. So if these quadrats were temporary, in other words, we took a new random sample in each year, then we'd use this contingency table. And here would be the results. We'd have a Chi-Square value of 2, a calculated Chi-Square value of 2, and a calculated p-value of 0.15.

So this p-value is higher than the threshold. And so we, in other words, there's a 15% chance now of making a false-change error if we say these two populations are different. And we were only willing to accept a false-change error of 0.1, and so we would not reject the null hypothesis. We would not say there's been a change between these two years. But we didn't.

Let's say these 100 quadrats were randomly located the first year, and then we permanently marked them. So then the second year we came back, and we measured the exact same quadrats. So then, we would enter the data in a different way. We would instead make note of which quadrats had the species in both years, and there were 50 of those. They went from present to present, in other words. Here's year one, here's year two. Zero species went from absent to present between the two years. But 10 went from present to absent. And then 40 were absent in both years.

So notice, we've put this information down in a different way. It's still a two by two table, but it's got different information in it. And so here's the formula for McNemar's test. And it turns out that only the quadrats that changed between the two years, the ones that went from present to absent, which is what PA means down here, or the ones that went from absent to present are part of the formula. And so we end up here's our calculation. We end up with a Chi-Square, a McNemar's Chi-Square, of 10, and as we saw here, that gave us a p-value of 0.0016.

And so we would, in fact, reject the null hypothesis and say that the species has changed between the years. So McNemar's test for permanent sampling units when the data are binomial, zeros and ones. One more thing before we leave this discussion of a Chi-Square test and McNemar's test. There's something called the Yates' correction for continuity that several textbooks, or some at least, advocate for the use of the Yates' correction. And it's only for a two by two contingency, too. And it's both for the Chi-Square and the McNemar's test. And it's based on a paper by Yates in 1934.

So essentially, this is the correction. So remember, the regular Chi-Square did the observed minus the expected square over the expected. But Yates said, well, you should subtract one half from that before you square it. And the default of some software programs, including R, is to apply the Yates' correction to any Chi-Square test or McNemar's test involving a two by two table. It's supposed to result in more accurate p-values, particularly when sample sizes are small.

But it turns out it's too conservative. Through simulations, many authors now have shown it should not be used even for small sample sizes. And here are some references for your perusal. So don't use it is the bottom line.

There's also something called a Fisher's exact test that you'll see in the literature and software programs will run for you. It's another test commonly performed on a two by two table instead of the Chi-Square test. But it has a problem, too. It's exact, but it's the exact only under a certain set of assumptions. And it turns out the p-values are similar to the ones you'll get with Yates' correction, and here's a paper that shows you shouldn't use it either. It's too conservative. So don't use the Fisher's exact test and don't use the Yates' correction. And here is a reference to that paper.

So how can we automate the Chi-Square and McNemar's test? Well, every statistical package will calculate both of those, just need to ensure that the Yates' correction is not applied. Here's our default output. So I just ask it to do a Chi-Square test on a data set, and it outputs the Chi-Square test with Yates' continuity correction, at least it tells you that, and we don't want that. So instead, I have to tell R not to apply it. So I put this command correct equals false. And then it just outputs the regular Chi-Square.

And notice that the Chi-Square values and the p-values are different. They're smaller when you don't apply Yates' correction. And again, Yates' correction is too conservative, so don't use it. So what about applying the finite population correction to the results of a Chi-Square test or a McNemar's test? Well, if you've sampled more than 5% of the population, you should apply the FPC to the results of a Chi-Square McNemar's test. Now remember, this only applies when you have a finite population. Right?

So if you were sampling with point intercepts and recording a hit or a miss on vegetation, point intercepts are not finite. You can have an infinite number of points out there. And so you would use the Chi-Square test or the McNemar's test-- well, you wouldn't use McNemar's because it's unlikely you'd have a permanent points. But you'd use a Chi-Square test, but you wouldn't apply a finite population correction because it wouldn't be appropriate.

So if you're sampling with quadrats or something with area, then you have a finite population. And then the question is, have you sampled more than 5% of the population? And if you have, you should apply the finite population correction. And so here's an example. We have a Chi-Square statistic from either a Chi-Square or McNemar's test, and we use quadrats. And in each of two years, we sample 77 out of a total big N 300 possible quadrats.

And here is the way we adjust that. We take the Chi-Square the value we got from our computer program, and we'd adjust it by dividing by the finite population correction. And I don't know if you remember, but for the t test, this finite population correction, we took the square root in the denominator. But because the Chi-Square statistic is itself squared, we don't do that with the adjusting the Chi-Square.

So by doing that, we've increased our Chi-Square from 2.7 to 3.6. And now, the only thing we need to do is calculate a p-value associated with that new Chi-Square. And again, there's this Excel workbook that we provide on the National Training Center's website that allows you to do that. Because Excel's data analysis tool pack doesn't do a Chi-Square test, it does do, as we mentioned earlier, a t test, both paired and independent sample, but there isn't a Chi-Square test option.

So we've provided an Excel workbook on the National Training Center website that will automate the Chi-Square test for a two by two contingency table. It provides the Chi-Square test statistic and a corresponding p-value and also applies the finite population correction if you're sampling from a population that's finite.

So far, we've only discussed one of the two kinds of sampling errors that are possible when conducting a significance test, the false-change error. And that's the only error considered when conducting a test, either the t test or a Chi-Square test. We set the false-change error rate we're willing to live with, like 0.1 for example, and then we compare the calculated p-value from the test.

And so if the calculated p-value is less than our threshold p-value, then we reject the null hypothesis and conclude the population has changed. On the other hand, if the calculated p-value is greater than our threshold p, we accept the null and conclude the population has not changed. But if we accept the null, there's another type of sampling error we may have committed, a missed-change error. And the missed-change error is also known as a Type II error or beta.

And the Power of a test is 1 minus beta. So the missed-change error, remember this is the situation where our monitoring system has said no change has taken place, so we fail to reject the null. But there has been a real change, and we don't know this, but there was a real change, we just missed it. So that's a missed-change error. And the missed-change error rate is a function of the level of change you want to detect, the acceptable false-change error rate and the sample size.

So let's look at Power and missed-change error rates. So Power can be defined in a couple of ways. One, it's just the probability of avoiding a missed-change error. Another way of looking at it is probability that a significance test will detect a true change. And it's the complement of the missed-change error rate, so it's 1 minus beta. So if the missed-change error rate is 0.3, then your Power is 1 minus 0.3 or 0.7, sometimes expressed as a percent. So in this case, our example would be we'd have a Power of 70% to detect the change that we want to detect.

Until fairly recently, like for ecology the 1990s, Power has been given insufficient attention. Some introductory statistics courses don't discuss Power at all. And that's likely because significance tests only concern themselves with the false-change error. And no Power or missed-change errors, you don't need to specify those in order to conduct a significance test. So it kind of flew under the radar.

And so the result of this is that the false-change here rate is controlled for, we control for that by setting a threshold p-value, by setting a false-change error rate. But the missed-change error rate may be very high and Power, therefore, low. This paper by Peterman pointed out that very few published studies in which a null hypothesis was not rejected reported the Power of the test to detect a true difference. They just assumed that the null hypothesis was correct. That, in fact, there was no difference.

But when the Power is low, this could be a dangerous assumption. And it's unfortunately not uncommon to have Power less than 0.5. And if the Power is less than 0.5, you're better off flipping a coin than actually doing the study. Because if you flip a coin to determine if there's been a change or not, at least you have a 50% chance. If your Power is less than 50%, then you're more likely to make an error doing the actual study. And so you need to pay attention to Power, and you need to adequately plan your study in order to have sufficient Power to detect the level of change that you want to detect. And we'll talk about how you do that.

So let's look at Power in a little more detail. So Power is a function of these four things-- the standard deviation, the sample size, the Minimum Detectable Change, this is the change that you want to detect, the minimum amount of change that you want to detect, and the false-change error rate or alpha. So how do you increase Power? Well, there's four ways you can improve Power, make it larger.

So you can decrease the standard deviation. So if there's a way to do this and make your standard deviation smaller, you should do it. And we won't have time here to talk about it, but the size and shape of quadrats have a big effect on the standard deviation. And it turns out that long and skinny quadrats often have a lower standard deviation relative to the mean or so lower coefficient of variation then square quadrats do.

And so there are ways, at least in some cases, of decreasing the standard deviation. Well, that's one way to increase Power. Another is to increase the sample size, which makes sense. The more you sample, the better handle you're going to have on the population you're making inferences to. Or you can increase the Minimum Detectable Change. Obviously, you want to start out your Minimum Detectable Change would be something that's biologically significant to you. But you might have to increase that.

So your Power to detect a larger amount of change is higher than it is to detect a smaller amount of change, so that's another way to increase Power. And the fourth way is you can increase the false-change error rate. Remember, we talked about the inverse relationship between the missed-change error, which is the complement of Power, and the false-change error. As one gets smaller, the other gets larger. And so if you make the false-change error rate larger, your Power will be larger.

So how do we balance these two kinds of errors, false-change and missed-change errors? Well, in the medical field, if you're screening patients for some lethal disease, the null hypothesis is that the person doesn't have that disease. But you're less concerned, in this case, about making a false diagnosis, which would be a Type I error analogous to a false-change error, and you're more concerned about failing to diagnose the disease, which would be a Type II error analogous to a missed-change error. Because if you fail to diagnose the disease, the patient's going to die. So there, the Type II error is more important to you.

What about a court of law? Well, the null hypothesis is the person is innocent, but if it's a criminal case, the proof has to be beyond a reasonable doubt. So you have a greater chance of a guilty person going free, which is a Type II error analogous to a missed-changed error. But that's all right because that's the way criminal courts are set up. But a civil case, on the other hand, the proof is based upon the balance of probabilities. So the two types of errors would be closer to equality. You'd be equally concerned about making either one of those errors.

What about a potential industry pollution source? Well, the null hypothesis is there's no pollution impact. And so industry would like to have a very low false-change error rate. So maybe, not even 0.05, maybe 0.01 because they're concerned with finding a problem. They're less concerned with Power and occasional missed-change errors. But if you're an environmental group, you're more concerned about making missed-change errors than false-change errors. You're concerned about missing impacts that are actually occurring.

And missed-change errors are often more serious than false-change errors in monitoring, particularly if something's going downhill. And there's many examples where fisheries have collapsed because the monitoring failed to reject the null hypothesis of no change when, in fact, the fisheries was going downhill. And some people have proposed something called the precautionary principle where you take action even if you can't statistically show things are not doing well, a particular resource is going downhill, you take action anyway.

So how do we use Power Analyses? There are two ways. The best way is prior to doing your study, so we'll call that Prior Power Analysis. And it's used to determine the sample size needed to achieve a specified level of Power. But it's also important, can be used in a post hoc manner-- it's called Post-hoc Power Analysis-- for interpreting non-significant results. And it's used to determine the power of an already completed study to detect a biologically significant effect.

So here's this equation we showed. And realize, these aren't all directly related, but these are a function of these four items-- the standard deviation, the sample size, Minimum Detectable Change and alpha. And I just want to show that you can rearrange these. Instead of solving for Power, we can solve for sample size. and that's what we would do in a higher power analysis. So before we did the study, we would have an estimate of the standard deviation. We come up with the sample size required to achieve false-change error or certain false-change error rate and Power.

We could also rearrange this and solve for the Minimum Detectable Change. So we can say, OK, we want to power of 90%, our missed-change error rate's 0.1. We know the standard deviation. And let's say we only have the resources to sample 50. We have 50 sampling units. So we could solve for a minimum detectable change, in that case, and see what that is, what's the minimum level of change we'd be able to detect with that sampling effort.

So what we need to do is we need to specify a sampling objective, just like we did for the case where we were estimating a single population parameter. We need to do the same thing now for detecting trend. And the things we need to specify are the Minimum Detectable Change, the Power, and the false-change error rate. And we want to specify that in a sampling objective. And this should be done prior to the actual sampling. And then that sampling objective can be used to determine the necessary sample size.

So here's an example sampling objective for trend detection. Be 90% certain of detecting a 25% change in the mean density of caribou with a false-change error rate no greater than 10%. Well, this is our Power the 90% certain, we want 90% Power. Here's our Minimum Detectable Change 25%. And then here is the false-change error rate 10%. All right. So this is expressed as a percentage here, we could express these as a decimal as well.

So our threshold p-value is 0.1 or the alpha is 0.1, all those mean the same thing. So there's one example. Here's another example. I want to be 95% certain of detecting an absolute increase of 10% canopy cover of weed species X, and I'm willing to accept a 5% chance of committing a false-change error. And there are those, Power is the 95%. So in this case, I want 95% Power. I want to be able to detect an increase of 10% in canopy cover, and I'm willing to accept a 5% chance of committing a false-change error.

So how to choose the Minimum Detectable Change, the false-change error rate, and Power. So the Minimum Detectable Change should be selected based on the level of biological change you deemed to be significant. That may be that based on sample size calculations, you later have to increase that Minimum Detectable Change, but at least use that as your starting point.

So the level of biological change that you think will be significant. So in terms of the false-change error rate and Power, because of work done by Cohen in the behavioral sciences, there's a tendency these days in the scientific literature for people to set the false-change error rate at 0.05 and the Power at 0.8. Which, remember, because Power is a complement of the missed-change error rate, Power of 0.8 is the same as a missed-change error rate of 0.2.

The problem with this is that it establishes that a false-change error is four times more important to you than a missed-change error because we've said we're willing to accept a 5% chance of making a false-change error, but a 20% chance of making a missed-change error. And in monitoring, it's often, maybe always, the case that the missed-change error is at least as important as the false-change error.

So for that reason, I like to set them equal to one another. Accordingly, if I set a false-change error rate of 0.05, that would correspond to a power 0.95. Since 1 minus the missed-change error is Power. Unfortunately, using a Sample Science program, you'll quickly find that under most circumstances, setting alpha 0.05 and Power at 0.95 will require a larger sample size than you have the capability of achieving. The only time this wouldn't be the case is if the standard deviation relative to the mean was pretty small, the coefficient of variation was pretty small.

Most of the time, this isn't going to work for you. And so it seems to me that levels of alpha of 0.1 and a power of 0.9 are a reasonable compromise in many situations. So you set your false-change error rate at 0.1 and your power at 0.9. So how about, then, determining sample size for detecting a difference between two means. Well, you need the following four pieces of information in order to calculate the sample size. This is for a t test, so a difference between two means.

You need the acceptable false-change error rate, you need the acceptable missed-change error rate or power, one or the other, they're essentially the same. One is the complement of the other. And the desired minimum detectable change is the third thing. All three of these are in your sampling objective, but then you need an estimate of the standard deviation. If you're using temporary sampling units, this is the standard deviation of the first year sample. If you're using permanent sampling units, this is the standard deviation of the set of differences between the sampling units between two years. So you need two years' worth of data.

So again, the first three of these are specified in your sampling objective, but the last one, an estimate of the standard deviation, requires a pilot study. If you're determining sample size for detecting a difference between two proportions, you're going to use a Chi-Square test for that. And what you need, in that case, is the acceptable false-change error rate again, the acceptable missed-change error rate or its complement power, the desired minimum detectable change, and an estimate of the proportion at time one.

The first three of these you specify, again, in your sampling objective, but the last one you need a pilot study or, as we mentioned in module three, if you assume the proportion is 0.5, you can actually calculate a sample size without a pilot study. Because doing that results in a conservative sample size if the actual proportion is closer to 0 or 1 because for proportion data, for binomial data, if the proportion is 0.5, you'll always need more sampling units. The sample size will have to be bigger than if it's closer to 0 or 1.

So pilot sampling. So again, you need an estimate of the standard deviation to detect the difference between two means, if you're going to do a t test. And unless you have an estimate from a similar study, which is possible, but unlikely, you need to do some sort of pilot sampling. And the pilot sampling should be conducted in the same manner as the planned sampling with a sufficient number of sampling units to ensure a stable estimate of the standard deviation.

And if you're using permanent sampling units, the pilot study needs to last two years. You need two years of pilot sampling because you need the standard deviation of the differences between the sampling units with that set of differences. So pilot sampling is also useful in calculating the sample size for detecting the difference in two proportions. Although, as we mentioned, the nature of binomial data is such that you can conservatively estimate the sample size by assuming the initial proportion to be 0.5.

So sample size calculations. The methods for calculating the sample size required to detect given levels of change for continuous and binomial data are beyond the scope of this course, but Measuring and Monitoring Plant Populations will give you the equations to calculate sample sizes for these three situations. But we also provide an Excel workbook on the National Training Center website that automates these sample size calculations, and we provide instructions as well. And so this is what I recommend is that you use the automated Excel workbook.

So we've talked about comparing the difference between two means or two proportions in the case of binomial data. What if we have three or more? So three or more sample means could be analyzed using an analysis of variance or ANOVA. In the typical ANOVA, the samples must be independent. In other words, they they're not paired. So you've got a random samples at each time point.

There is a repeated measures ANOVA that you could use on permanent sampling, unit but it has drawbacks, assumptions that have to be made. So you can use ANOVA for means if you have three or more sample means. You can use a Chi-Square test for three or more sample proportion. Again, they have to be independent. But instead of a two by two contingency table where you have two rows representing presence or absence and two columns representing two years, you have a 2 by n where the number of columns is equal to the number of years.

So for three sample proportions, for example, if we have three years, we have a two by three table. Again, samples must be independent for a Chi-Square test, not paired. The problem though with both the ANOVA and the Chi-Square procedures to compare three or more populations is that a significant p-value from the test only indicates that one or more of the populations are different from one or more of the others. You don't know which one. It doesn't tell you which specific populations differ.

To determine which populations are different, you need to conduct additional tests. And various post hoc tests have been developed to determine which population parameters differ. But the easiest procedure is simply to perform pairwise t or Chi-Squared test, whichever is appropriate for your type of data, to determine which populations are different. So for example, we have three years of monitoring data. We'd compare year one to year two, year one to year three, and year two to year three. So if your data are continuous, you conduct three separate t tests. They'd be independent t tests if you employed temporary sampling units or paired t tests if you used permanent sampling units.

If your data are binomial, you'd conduct three separate Chi-Square tests, if you're sampling units are temporary, or three separate McNemar's test, if you're sampling units are permanent. Because this results, though, in what are called multiple comparisons, some argue there's some type of correction for multiple comparisons is required. The rationale for this correction is that with a single test, the chance of making a false-change error is equal to the threshold p-level you've set.

So if you set a threshold p-level, an alpha, of 0.05 and you do a single test, then we can expect that we'll make a false-change error by chance alone about 5% of the time. But if we do three independent tests, each of them with a threshold p-value of 0.05, the chance of one of them being significant by chance alone increases, and it increases to 1 minus the probability that none of the test yields an erroneous rejection of the null, which is equal to 1 minus 0.95 by 0.95 times 0.95 equals 0.14. So now the probability that one of them will be significant by chance alone is 0.14, no longer 0.05.

So to correct for this increased false-change error rate, several corrections have been proposed, and that one you'll see most commonly is called the Bonferroni correction. And it controls what's called the familywise error rate, FWER, associated with this set of tests that you performing. And the objective is to ensure that the overall false-change or Type I error rate does not exceed the objective error rate.

And to ensure that, the threshold p-value, which is the overall objective error rate, is divided by the number of significance tests, for example t tests, to derive a corrective threshold p value that's then applied to the results of each test. And this is more easily seen by an example.

So for example, if our original threshold p was 0.05, and we sample of the population in three years, then we'd have three t tests. Right? We'd have year one versus year two, year one versus year three, and year two versus year three. And so we adjust the original threshold p-value of 0.05 by dividing by 3. So now, it's 0.05 divided by 3, and it's 0.0167. So now, our calculated p-value must be less than, for any of these tests, must be less than 0.0167 in order to declare any of the differences significant.

Well, what's wrong with that? And believe me, there is something wrong with it. It punishes you for collecting additional information. So let's say we collect data in only two years, year one and year three. And we decide to use a threshold p-value of 0.05. So the mean of year one is seven plants per quadrat, and the mean of year two is three plants per quadrat. And so we only need one t test because we just have two years. So your false-change error rate is still 0.05. We don't need any kind of correction.

So the t test yields a calculated p-value of 0.031. So this is the p-value is less than our threshold, so we conclude that a change has taken place between the year. OK. But now suppose we also sampled in year two, and we came up with a mean of four plants per quadrat. So now we have the following data. So we have seven plants per quadrat in year one, four in year two, and three in year three. So a steady decline in the mean.

But now, because we have three tests we're going to do, we've got to now, according to the Bonferroni correction, divide the 0.05 by 3, so now our threshold p-value is 0.0167. So now, because our p-value of 0.03 is greater than the threshold p-value, now we have to conclude there's no difference between year one and year three. When we just had year one and year three data, we had no problem declaring a difference. But now, because we've had to apply this correction, we say there's no difference. So we're penalized for having more information even though that information seems to show that, indeed, there's been a decline.

So the other thing to realize is that by adjusting this, correcting this p-value, it's going to make it less likely we'll make a false-change error, that's for sure. But it makes it much more likely it will make a missed-change error. Here's another example. So here's an experiment looking at the effects of grazing on 10 plant species. We have five grasses and five forbs. And the G stands for grazing or grazed and the U stands for ungrazed. So we had grazed plots and ungrazed plots.

And in some cases, the grazed plots had-- and I'm not sure what they were looking at, let's say, this is cover. So that the grazed plots had less cover of grasses in some cases under grazing and higher cover grasses in others under grazing, but none of those were significant. For the forbs, all five forbs decreased under grazing, and they significantly decreased. Here are our threshold p-value is 0.05.

So these tests, so let's say these were t tests, all showed significant differences. But we made 10 tests. Right? So because we had 10 separate independent t tests, we needed to make an adjustment. So if we took the Bonferroni adjustment, we would divide this 0.05 by 10. And now, all of these would have to be less than 0.005 in order to reject the null hypothesis of no difference, and none of them are. So we'd have to conclude there's no difference for any of these species, and yet, the probability that 5 out of 10 tests being significant at 0.05 is only 0.000006. So clearly, the Bonferroni correction is not appropriate here. This is the paper from which this example came.

So what should I do if I make multiple comparisons? Well, one option is to don't use the Bonferroni or any other correction. Because of the problems with the Bonferroni and most other corrections, and there are others that are less conservative than the Bonferroni but are still problematic, there is one that seems to work pretty well that we'll discuss in a minute. But we recommend just not using them.

Because comparisons between years are essentially planned, as opposed to post hoc unplanned comparisons, you're justified in using the uncorrected threshold p-value or alpha value. But if you write a report, be sure to state that you did not apply a correction for multiple comparisons to your calculated p-value. And the next slide gives you some references that you can cite to justify this, and there they are.

So the second option is that you can use the correlated Bonferroni procedure of Dresner and Dresner 2016. So the rationale behind using the Bonferroni correction was to correct for the increased likelihood of making a false-change error with multiple tests. So as we mentioned, an example, if we do three independent tests, each with a threshold p of 0.05, the probability that one of them might be significant rises to 0.14. But this assumes that the tests are independent, and that's clearly not the case when comparing data from different pairs of years.

If we compare year one to year two and year one to year three, for example, the year one data appears in both comparison, thus the comparisons are correlated and not independent. So the Bonferroni and similar familywise error rate correction procedures assume that ordered statistics are independent when, in fact, they're not. And the Dresners show using simulations that with 10 significant results are sorted, there's a 0.942 correlation between the largest and second largest statistics even when data are generated from independent standardized normal distribution.

So there's a correlation that's not taken into account by Bonferroni and most of the other correction procedures. Whereas the Dresner procedure takes that correlation into account. In their paper and the website associated with the paper, they provide an Excel workbook that implements it. So both the paper and the Excel workbook are available for free, and here is how to access it. And so if you're going to use a multiple comparison procedure, then we recommend this Dresner and Dresner procedure be applied.

So big data and multiple comparison corrections. I just want to mention this because our recommendation not to use post hoc comparisons applies to the situation where you're comparing just a few means. These days, science-like genomics may gather 10,000 variables. These are gene expressions on 100 individuals. And so you could have 10,000 separate tests on means and clearly, in those cases, you need some control on false positives. But you don't use the Bonferroni or other familywise error rate corrections, you use something called the false discovery rate, which we won't get into here, but I just wanted to make it clear that I'm not saying you shouldn't use post hoc corrections when you're comparing many means.

So what if we have a larger number of years then, say, three or four? Well, we have more than four years of data, we should no longer be using two sample tests to test the differences between pairs of years. Instead, we should use linear models. And the objective here is to determine if the slope of a regression line through the estimated means or other statistic for each year is significantly greater or less than 1. Here's an example.

So we have 10 years' worth of data, and we draw a regression line through the data points. And in this case, here's our regression formula or estimate. And our slope here is minus 0.163, so it's going down. And now we test to see whether that slope is significantly different than zero. And if it is, we can conclude there's been a population decline. And analyzing counts or other measurements over periods of five or more years is geared toward determining trends in population. And this trend analysis is beyond the scope of this module.

A future model may address this type of analysis, but a good introduction to the use of regression techniques for monitoring is chapter 10 in this book by Elzinga et al, *Monitoring Plant and Animal Populations.* So that's the conclusion of module four. Some potential future modules that we may produce include the analysis of trends using linear models, effective presentation of the results of statistical analysis using graphs and tables, and using R to analyze monitoring data.

Well, thank you for participating, and maybe, we'll see you in a future model.

[MUSIC PLAYING]