JOHN WILLOUGHBY: Hello and welcome. This is John Willoughby, and this is a course in the Basics of Statistical Analysis for Use in Natural Resource Management. This is Module 3 of the course, where we're going to talk about how you can use statistics to estimate a single population parameter. And this is the first module in which we'll talk about inferential statistics, a branch of statistics in which we make inferences from a sample to a population.

This is the third of four modules. Hopefully, you've already looked at the introduction to statistics and descriptive statistics module. In this module on inferential statistics, we'll talk about estimating a single population parameter. In the following module on inferential statistics, we'll look at comparing two or more population parameters to see if the populations are different.

And I'd like to thank my colleagues Dan Salzer and Marie Denn for providing some of the slides I use in this presentation. Dan and Marie and I instruct an annual BLM National Training Center course on measuring and monitoring plant populations and vegetation. We've been doing that for many years now on an annual basis.

Dan and I have been teaching my class now for more than 25 years, and Marie has been part of the class for some 15 of those years. So let's first distinguish between what we mean by populations and what we mean by samples. Often, we wish to know something about a particular species or habitat. This is common in natural resource management.

For example, we might want to know the number of dead conifers in the Tahoe National Forest, or the average chest height of a male moose in Nova Scotia and New Brunswick-- an example I talked about in a previous module-- or the canopy cover of coast live oaks in the Hastings Natural History Reservation, or the presence or absence of salamanders in designated stream regions in the forests of Coos County, Oregon.

Another example might be the plant species richness of vernal pools at the Jepson Prairie Preserve, or the average height of deerbrush plants following the King fire, or the average number of seeds produced per plant for Peirson's milk-vetch, a federally threatened species. What we're interested in is the true population parameter.

For example, we want to really know what the true number of dead conifers is in the Tahoe National Forrest, or the true canopy cover of coast live oak the Hastings Reserve, or the true average chest height of male moose in Nova Scotia and New Brunswick. Although we could theoretically measure the entire population, take a census, this is usually not possible.

Instead, we take a sample and use that sample to make inferences about the population as a whole, and this is the domain of inferential statistics. So let's get some definitions under our belts here. So we've got the term "population," and the population is the complete set of all objects for which we want to make inferences. So in the moose chest height example, that would be the chest height of all the male moose in Nova Scotia and New Brunswick.

So our population is all the male moose. We obviously can't measure all the moose, so we're going to take a sample. When we take a sample, that's part of the population. It's a subset of the total number of sampling units. So in the moose example, we took a sample of 94 male moose, and we would want to randomly select those from the population.

In the moose example, these were moose that were examined at hunter check stations. These were moose that had been shot by hunters, and there might be some concern about whether that was truly a random sample or not. But the author in that study assumed that it was. So when we take a sample, the individual objects that make up the population would be all of the sampling units in that population.

And in the case of the moose, we're not really going to know that number. That number is big N, and we'll talk about symbols for populations and samples here in a minute. But the entire population would be all of the moose in the area that we're interested in. So the sampling unit is each individual male moose. So we've got a sample of 94 of those, and there's some number that comprised the entire population that is an unknown number.

So population parameters versus sample statistics. A parameter describes a characteristic of the entire population. So it could be the population mean. It could be a proportion, or it could be the population standard deviation. All of those are population parameters a statistic is a characteristic of a sample. So we could have the sample mean, the sample proportion or the sample standard deviation.

And it's our hope that that statistic well represents the actual parameters, so that for example, the sample mean is a good representative of the true population. Sample statistics are used to estimate the corresponding population parameters. Greek letters are often used to signify population parameters while Roman letters are used for sample statistics.

So let's look at some of these designations. So we see the population parameter, mu, this Greek symbol that looks a little bit like a u but it is called mu, is the population mean, whereas X-bar is a sample mean. Sigma, the Greek letter sigma, is used to represent the population standard deviation, which would be the true standard deviation that we're trying to estimate, whereas s is the sample standard deviation.

Sigma squared represents the population variance whereas s squared represents the sample variance. The letter p, lowercase in this case, but sometimes you'll see it in upper case, represents the population proportion, whereas p-hat represents the sample proportion. And then finally, big N, uppercase N, represents the number of members of the population, or the number of sampling units that you could possibly sample in that entire population, whereas small n, lowercase n, designates the number of sampling units.

Sometimes, big N is known to you, particularly if your sampling units involve area like plots or quadrats, because then you can calculate. You know the area of the quadrat. You can calculate the area within which you're sampling, and you know how many possible quadrats you could put in that population. But other times, big N might be infinite. Like it might be the number of points.

Maybe you're sampling with point intercepts, and there's an infinite number of points that you could theoretically put in that population. So in that case, big N is infinite. In other cases, it might as well be infinite because it's unknown to us, because the number of male moose, for example, in the example that we gave of trying to estimate chest heights in a very large area, we have no clue how many moose occur in that entire area. So big N is unknown in that.

So we need to make a distinction here between a biological and a statistical population. What do I mean by that? Well, when I use the term "population" in this module, I'm going to be referring to the statistical definition of the term. Sometimes, like in the moose example, the statistical population is the same as the biological population. The sampling unit in that case is a moose and the statistical population is all of the moose in the area of interest.

So the biological and statistical populations are essentially the same in that case, but in many other situations, the statistical population is different, even when the purpose of the sampling is to estimate something about a biological population. For example, if we use quadrats to estimate the population size of a rare plant, the statistical population is the complete set of quadrats we could place in the area of interest, not the population of plants in there. The statistical population is the complete set of quadrats.

Statistical populations may also consist of point intercepts, stream reaches, line transects, routes, traps and so on. Target versus sample population. Typically, we've got a target population that we'd like to make inferences to, and it's this large area that's represented here in the outside of the circle. And although we'd like to sample all of the target population, oftentimes, we can't because of logistical problems.

Like maybe this large area includes public and private lands, and we can't get permission to sample all the private lands. And so we only sample the public lands and those private lands we can get permission for. Or maybe the topography as such-- you know, because of the steepness of the terrain, we just can't get to it, and so we leave that out. We don't sample it.

So there's typically a difference between the target population, this large area, and what you can actually sample, the sample population, which is this orange area inside of the target population. So basically, we sample what we can, and when we make statistical inferences, it can only be to the sample population, not to the entire target population.

But does that mean we can't make management decisions based on what we find in the sample population? No, we do it all the time in natural resource management. And just to complete this diagram, so here's our sample population, the orange area. We take a sample that's much smaller than the sample population and we make inferences to the sample population based on that sample.

And that sample is comprised of several sampling units. Each sampling unit, of course, is smaller than the total sample. So here's an example where we have a macroplot. We've delineated a square area around a population of plants that we're interested in, in this case. These square areas are called macroplots, typically, and they provide a convenient mechanism for sampling.

And here I've shown a macroplot that's 20 meters by 20 meters, and typically, macroplots are going to be larger than that. But in this case, I've just made it smaller so it fits on this slide, and we can look at things more clearly. So here we have a population of 397 plants, and I know that because I generated this using the computer.

If the plant is the sampling unit, like let's say we want to estimate the mean height, of plants in this population-- so if the plan is the sampling unit, then the sample is a randomly selected subset of plants. And we see that these red circles here are the sample of plants. The sampling unit is each individual plant, as I mentioned, and then the population in this case is all of the plants in the macroplot. That's the population we're trying to make inferences to.

And so here's a case where the biological population and the statistical population are essentially the same. They're the number of plants. But here's an example. Typically, if we're going to estimate total population size as opposed to plant height, some attribute of the plant itself, In this case, we're more interested in the number of plants. We're going to use a plot of some sort, a quadrat that has area associated with it.

So here, we're going to sample with 2 meter by 2 meter quadrats, and you can see the total number of 2 by 2 meter quadrats that we could put in this 20 by 20 meter macroplot. And there are 100 of those shown here. So the population is now the total number of quadrats we could put in this area and not the total number of plants.

So now the statistical population is this set of quadrats. So if we were to count the number of plants in all of these quadrats, something that with just 100, you could do, we come out with the following true population parameters. So the true mean is 3.97 plants per quadrat. The true population standard deviation is 3.754 plants per quadrat.

But in our example here, we won't examine all 100 quadrats. Instead, we'll take a random sample of 10. But before we do, let's look at the distribution of the data if we did take all of those 100 quadrat values. And you can see here that the distribution, number one, is not normal. There's no bell-shaped curve here.

Many of the values have relatively small counts, so 17 of the quadrats are 0s. 14 are 1s. There's a single plant. But now we have a few quadrats with large counts. So 5 counts are greater than 13, and one of the quadrats has 17 plants in it. This is not atypical of count data, particularly when using square quadrats.

This data set, if we were going to model it by some distribution, would likely be best modeled by the negative binomial distribution, which I talked about in the last module. So here's our 20 by 20 meter macroplot again. Because, again, the population is relatively small, it wouldn't be out of the question to count the plants in every one of the quadants.

So if we did that, that would be a complete census. But again, I made this population small to fit it on this slide, and in most cases, populations are larger, and therefore censuses are not practical. So we're going to use sampling instead. So here, we've taken a random sample of 10 quadrats from the population of 100 quadrats.

Notice that just by chance, when you take a random sample, we've got some quadrats that are right next to each other, and there's areas that aren't sampled at all. And that's the nature of simple random sampling, which we've done in this case. We won't have time to go into other types of sampling. I'll mention them later, but I just wanted to point that out, that just by chance, you can have quadrats close to one another.

So here, the sampling unit is a quadrant where there are 10 of those. The 10 quadrats comprise the sample, and the population is all of the quadrats. There are 100 in this case that could be put in that population, in that macroplot, without overlap. And so here are the results from that random sample of 10 quadrats.

Notice that the way we sampled these was by taking random coordinates. So we pick a random even number between 0 and 18 on the X-axis and a random even number between 0 and 18 on the y-axis. And we put those random numbers together and determine the coordinates. So in this case, for example this quadrat here, we have the coordinate 0 on the X-axis and 8 on the y-axis.

And so we then went to that quadrat and sampled it. At least, that's the way we would do it if we were truly doing this in the field. So here are the coordinates. Here the number of plants associated with each of the quadrats that we sampled at those coordinates. And over here are the sample statistics. So little n is 10, the sample size. And the mean from that sample was 5.2 plants, and the standard deviation was 4.02.

We can convert that mean value and standard deviation to a population estimate. So here we have 5.2 plants per quadrat. We have 100 quadrats, and therefore the estimated population size is big N, 100, times the mean of 5.2, or 520 plants. We do the same with the standard deviation. So here's the standard deviation for the quadrat values, and here's the standard deviation for the population. It's big N times the standard deviation, the per quadrat standard deviation, and we get 402.22 plants.

We can calculate a confidence interval, which I'll talk about in a minute how we go about doing that. But if we did a confidence interval, we'd see that the 90% confidence interval would be plus or minus 233 plants. So we could take a different random sample. Here's a different one. Another random set of coordinates we're selected, and in this case, the sample size is still 10.

But now, our mean estimate 4.2 plants. Remember before, the first sample. It was 5.2. So just by chance, we got a different mean and a different standard deviation. That's the nature of sampling. We're never going to get the same number when we do sampling. The question is, how far apart are those samples from one another, and that's what the confidence interval will calibrate.

So a population estimate in this case, again, we multiply 4.2 by 100 because that's how many possible quadrats there are. We got 420 plants. Standard deviation, 533 when we multiply big N times the quadrat standard deviation. And now our confidence interval is plus or minus 309 plants.

So here's yet another sample. So in this case, again, a new random set of coordinates, different quadrat sample, and we end up with different values here for the means and total, and a different confidence interval. So what if we took 1,000 samples of N equals 10? So 1,000 samples, but with a sample size of 10. Well, we could recruit 1,000 interns to do this work, and we'll send them out there.

And we send them out to the field to sample, and notice they're smiling now, but maybe not so much after they actually do the work. So here are the results of the sampling. We send them out there, and we took 1,000 samples of N equals 10. The graph here is a histogram showing a very important distribution, the sampling distribution, of the means of each of the 1,000 samples.

So recognize that this is a data set of means. So we have 1,000 samples. In each of those samples, we took the mean, and then we graphed the means here. And so you can see that some of the samples, the mean number of plants was fairly low. In some, it was fairly high, but most of them are somewhere between 2.5 and 5 plants per quadrat.

The actual mean of these 1,000 samples is 3.99, and that's very close to the true mean of 3.97. And this will typically be the case, that the mean of the sampling distribution of means will be very similar, sometimes the same, as the true mean, the mean that we're trying to estimate. The standard deviation of these 1,000 sample means is a hugely important statistic. It's called the standard error of the mean, or typically just referred to as the standard error.

And the standard error is simply the standard deviation of the collection of sample means. In this case, that value is 1.1. And here's the formula again for the standard deviation, but in this case, we're not taking the standard deviation of the original data set. We're instead taking the standard deviation of the set of means, the sampling distribution of the means. And that standard deviation is the standard error.

And note that despite the fact-- remember the original data set wasn't even close to normal. It was skewed to the right, but the sampling distribution, the distribution of the sample means from that population, are in fact, almost normal. So what if we have those interns take samples of 5 instead of samples of 10? So 1,000 samples again, but now this time, each sample consists only of 5 quadrats.

So we take a sample of 5. We calculate the mean. We take another sample of 5, calculate that mean. We do that 1,000 times, and then we plot that set of means here. So this is our sampling distribution. And in this case, the mean of the 1,000 samples is 4.01, still very close to the true mean of 3.97. But now, the standard deviation of the sample means is the standard error. In other words, it's larger.

Before, it was 1.1. Now, it's 1.649. That's the result of a smaller sample size, n equals 5 as opposed to n equals 10. Also, because we've taken a smaller sample-- the sample size is smaller-- the sampling distribution now is not as close to normal as it was with when the sample size was 10. What if we increased the sample size to 25, where these interns are really getting tired now? Now we're having them do yet a third sample size out there. They're definitely not smiling anymore.

But when we have them do that, we get 1,000 samples now, and notice that the sampling distribution is narrower. That mean of the sampling distribution here is 3.95, still very close to the true mean, but now the standard deviation of the sample means, which is the standard error, is smaller. Now it's 0.657. So when we had an n of 10, it was 1.1, and now it's smaller. It's 0.65.

And note now, the sampling distribution is essentially normal. So the sampling distribution can also be used to calculate a confidence interval, an empirical confidence interval around the estimate of the mean. So we do that by determining-- you know, these values are in order. Again, this is the sample distribution of mean, so this is a means of n equals 10 in this case.

Here's the estimate of the mean from the sampling distribution. But if we look at those values at which 5% of the values are outside of on both tails, both sides of the distribution-- so here is the value at which, if the value here, this is at 2.3. And so all of the values, 5% of the values are below this. 90% of the values are between these two lines, and 5% of the values are above this value here. This is at 6, 6.0.

So if we take those values, we can say that 90% of the values here, the means, fall between the value 2.3 and 6.0. So that's our estimate of the confidence interval. 90% of the means in the sampling distribution fall within these two lines, and so that's a 90% confidence interval. This is called a percentile confidence interval, usually. Percentile, quintile, those are synonymous terms. So 90% percentile confidence interval.

So here's the same calculation, but now we've got 1,000 samples of n equals 25. And the thing to be aware of here is that means of 25 are estimating the true value better than a smaller sample size, and therefore, the distribution of means is narrower. And now our confidence interval is also narrower. It goes from 2.9 to 5.0 in this case.

So we never want a sample with 1,000 interns. I mean, it's obviously not practical, but if we had 1,000 interns, it would make more sense for them to census the entire population rather than sampling 1,000 times. And so you'd simply count all of the quadrats here. There's only 100, and so one person or two people could do that.

The other issue, of course, with sampling with 1,000 interns, is that the habitat might not fare too well. Fortunately for us in real life, we only have to take one sample. We don't have to take 1,000 samples. But if we only take one sample, how do we determine the standard error, which is the standard deviation of the sampling distribution of means?

I mean, we could calculate that when we had 100 means or 1,000 means-- you know, a 1,000 samples-- but if we take one sample, how do we calculate that? Or how do we calculate a confidence interval, which we also calculated from the sampling distribution? But fortunately, there's a relatively simple way of estimating both the standard error and the confidence interval from a single sample.

And you do that by calculating the standard error, estimating the standard error of the sampling distribution, from the standard deviation and sample size of the sample, the single sample. And then the confidence interval is calculated from that standard error. So here's our formula for the standard error. It's relatively simple. It's the standard deviation of the sample divided by the square root of the sample size.

So that's how we estimate the sampling distribution from a single sample. So we use the standard error, then, to calculate something called a confidence interval, which I introduced a bit ago, but we're going to talk in more detail about now. So when we report estimated values based on sampling, we need to quantify the uncertainty associated with the estimate.

So all we have is the estimated parameter from a single sample, but what we're interested in is the population parameter-- so for example, the true population mean, or the true population total, or the true population proportion. But all we have is the sample mean, or the sample statistic. In the case earlier, we had a sample mean number of plants per quadrant, but those means vary by sample, and so we can't just use that sample mean by itself to estimate the true population parameter.

We need a confidence interval, and a confidence interval gives us a range within which the true population parameter is expected to fall with a specified level of confidence. So for example, we might have an estimate of the mean number of banana slugs per meter squared of 0.05, and we might then calculate our 90% confidence interval that says that we're 90% confident that the true number of banana slugs per square meter is somewhere near or somewhere between 0.04 and 0.06.

And we could show this graphically. You often see bar charts used for this purpose. And so here, we show the mean, the estimated mean at the top of the bar. So this is the [? point, ?] our estimated 0.05 slugs per meter squared, but here is the error bar here. It displays the range within which the true value may reside with 90% confidence.

And notice that I've told people that the error bar is the 90% confidence interval in this case. That's one way to display an estimated confidence interval. Another way would be in a table. So here, we show the estimated density in slugs per meter squared of 0.05, and here's the 90% confidence interval. And it's often displayed this way within brackets.

So let's talk about how we calculate confidence intervals. The standard error is used, as noted before. And I'm going to show you the wrong formula for a confidence interval, simply because you often see this in books. And it turns out it works OK if the sample size is large enough, but the sample size often isn't. And so we're not going to use this, and we'll talk about why.

But here, if we use this incorrect formula, we would multiply the standard error by a z value, a critical value from a z-table. And I'll talk about what that means in a minute. So it would be the mean plus or minus the standard error times z. And don't let these subscripts intimidate you here. This simply says that alpha-- we specify the confidence interval or the confidence level that we desire-- and the 2 just means we want a two-sided confidence interval.

In other words, the confidence interval is going to be the sample mean plus or minus some value. And so the z value comes from something called the standard normal distribution. We talked about normal distributions already, but the standard normal distribution is that distribution, that normal distribution, that's described by a mean of 0 and a standard deviation of 1.

And remember, we talked about earlier that most of the data are going to fall-- 95% of the data are going to fall-- within plus or minus 2 standard deviations from the mean. And here, this shows that. So it turns out that it's not exactly two standard deviations. The z value is actually 1.96, and so 95% of the values fall within plus or minus 1.96 standard deviations from the mean.

But in most cases, people just round this up to 2. So they'll say the 95% confidence interval would be plus or minus 2 standard deviations. So in this case, to do the confidence interval, we would multiply 1.96 times the standard error that we had calculated previously. If we were doing a 90% confidence interval, though, the z value is going to be smaller.

The z value here in this case is 1.645, so here we would multiply the standard error times 1.645. And that would be our confidence interval. Our confidence interval would be the estimated mean plus or minus 1.645 times the standard error. And here's a table of z-values that you could use for different confidence levels. So again, for the 90%, it was 1.645.

If we wanted an 80%, it would be 1.282, and so on. But these, again, are not the values we want to use, and that's because confidence intervals calculated using z-values are unreliable. Why is that? So, well, there's two reasons. They assume that the population standard deviation is known, and that's seldom the case. And instead, we're going to estimate the population standard deviation from a sample.

The second reason they're unreliable is that they assume that the sample size is large. And what do we mean by large? Well, that's relative, clearly, but the sample size should be at least 30 before you consider using a z-value. Smaller sample sizes than 30 are likely to do a poor job at estimating means and standard deviations, but the z values don't take that into account.

So what do we use instead? Well, we use something called the Student t Distribution, or simply a t distribution, and that was introduced by a statistician who worked for the Guinness Brewing Company. That statistician's name was William Sealy Gosset, and it turns out that he did a bunch of experiments to improve the efficiency and quality of the products. And thankfully, that's the case, because it remains to this day a very good product.

Some of these experiments that he conducted involved things like testing for different strains of barley, but it was very expensive to test that because you had to devote test fields to it. And so because of that, sample sizes were often small, for example, just three fields. Gosset quickly realized that z-values from the standard normal distribution weren't going to work for him, and so he came up with the t distribution instead.

He invented this distribution, and he wanted to publish it, but his employer wouldn't allow him to publish under his own name. So it's not clear why Guinness wouldn't allow him to publish under his own name. It may have been that-- one theory has it that they were worried that others might learn trade secrets if they knew that he worked for the Guinness Brewing Company.

But they did allow him to publish, and he published under the pseudonym Student. And in so doing, this well-natured and relatively modest man made one of the most important contributions of 20th century statistics. Here was his paper. He was a contemporary of two Giants in the statistical field, Ronald Fisher and Karl Pearson.

Those two individuals never got along. They had a long-running feud and pretty much hated each other. William Sealy Gosset, on the other hand, got along famously with both of them. So why the Student t distribution? Well, unlike the standard normal distribution, the z-value, the z distribution, the Student t distribution accounts for sample size, and it does this by using a concept called degrees of freedom.

So degrees of freedom are calculated by subtracting 1 from the size, n, of the sample. So a sample of 10, for example, would then have 9 degrees of freedom. n minus 1 or 10 minus 1 equals 9 degrees of freedom. A sample size of only 2 would have 2 minus 1 or 1 degree of freedom. A sample size of 1 would have 1 minus 1 or 0 degrees of freedom, and it's not acceptable. You can't deal with a sample size of 1.

So the distribution itself is similar to the normal distribution, but it has thicker tails with smaller sample sizes. So here's a comparison of t distributions for comparison. So the normal distribution is this dashed line, and you can see the t distributions at different degrees of freedom. So degrees of freedom of just 1, which would be a sample size of 2, would be this blue line.

You can see the tails are much fatter. The distribution is more squat and the tails are fatter, and you can see the other distribution, the other degrees of freedom as well for the distribution. So here's the correct formula for a confidence interval. It's the mean, the estimated mean, plus or minus the standard error times t, a critical value, a t critical value.

In this case, again, this subscript here now shows the confidence level that we want. We want a 2-sided interval, and now we've got this symbol that looks like a v. It's actually the Greek nu, which is often used to designate the degrees of freedom. So the t value comes from a table of critical t values, a t table, and corresponds to the desired confidence level, this alpha, and degrees of freedom, nu.

So let's say we want a 90% confidence interval. Let's look at the worst case scenario where the sample size n is only 2. And we're going to compare the normal z-table confidence interval with the t confidence interval. So degrees of freedom in this case would be n minus 1 or 2 minus 1 equals 1. So here's the standard normal distribution, and you know, that says that we would construct our confidence interval using a z-value of 1.645.

We would multiply that value by the standard error. But here's what the t distribution says we should use. Instead of using z, 1.645, we would use t equals 6.3. So essentially, if we calculated a confidence interval using the t distribution, It would be 3.84 times wider than the confidence interval we would get if we use z.

And it turns out that that wide confidence interval is much more likely to be correct. I mean, number one, it's going to tell us we need a larger sample size because this width is going to be really large, and so our estimate is not going to be very precise. But that's why we want to use t instead of z. We want to use the critical value of the t distribution instead of the one associated with the z distribution.

And so here's an example, a t table. These are critical to values for several levels of confidence that we might desire. And so the t table often specifies confidence levels in terms of alpha. So alpha, in order to calculate your confidence level that you want, you subtract alpha from 1. And so a 90% confidence interval would be indicated by alpha equals 0.1, because 1 minus 0.1 would be 0.9.

And that's the confidence level expressed as a proportion. We often then multiply it by 100 and express it as a percent instead. So we had 9 degrees of freedom, so for a confidence level of 90% and a sample size of 10-- remember, degrees of freedom is n minus 1, so the degrees of freedom are now 9. So we would use 1.833 as our multiplier. We'd then multiply that by the standard error.

So here's what the t table from a statistical textbook often looks like. Up here, here's our 2-sided 90% confidence level, which is 0.1. So in other words, for a 90% confidence level, it's 1 minus 0.9, which is 0.1. And here's the critical value again for 9 of freedom for a 2-sided test and a 90% confidence interval.

The table here only shows degrees of freedom up to 20. Actual tables continue to much larger value. Another thing I wanted to highlight here is that these t tables also have values for one tial, and we'll talk more about when you might use the one-tail value when we talk about statistical tests in Module 4.

So here's the calculations for our first sample of n equals 10 from that population of 100 2 by 2 meter quadrats. So our mean was 5.2. Standard deviation was 4.02, and we had a sample size of 10 quadrats. So the standard error in this case, remember, is the standard deviation divided by the square root of the sample size. So here, our standard deviation was 4.02.

We divide that by the sample is the square root of the sample size of 10, and we get a value of 1.272. We then calculate the confidence interval, a 90% confidence interval, by multiplying that standard error times the appropriate t value. And remember, that was 1.833. So here is that operation, and we end up with a mean and a confidence interval of 5.2 plus or minus 2.332.

And remember, this subscript here just shows the alpha value associated with our confidence interval. So it's 0.1. We wanted a 2-sided one, and here's our degrees of freedom, nu, of 9. So now we can say that, well, our confidence interval, the 90% confidence interval, is 5.2 plus or minus 2.3 plants per quadrat. Or we could say that we're 90% confidence that the true population mean is somewhere within the interval 2.97 to 7.5 plants per quadrat.

These were derived simply by subtracting the 2.3 from 5.2 and adding the 2.3 to 5.2. And so here's our range now of 2.9 to 7.5 plants per quadrat. . So that's our confidence interval. And when we express a confidence interval like one above, the plus or minus 2.3 plants per quadrat part, this part, is often referred to as the margin of error.

And it's very common to see this in statistical surveys of likely voters in elections, where a pollster might say, for example, that 54% of voters favor candidate A with a margin of error of 5%. So that margin of error is a confidence interval, but be careful with this terminology, because the confidence level used in those surveys is usually 95%.

So unless you're actually using a confidence level of 95%, I would avoid using the term "margin of error" and simply say 'confidence interval" instead. So we've talked about this before, but let's talk about it again, this time with confidence intervals. So converting density estimates into population estimates. We have an estimate now of mean density for our sample population of 5.2 plants per quadrat, and an estimate of the 90% confidence interval around that mean density of plus or minus 2.3 plants per quadrat.

But now we'd like to estimate the total number of plants in the population. We'd like to use the mean value and convert that to an estimate for the total number of plants. So it's easily done. Simply multiply the mean and the confidence interval estimates by the total population size, n. There are big N possible quadrats that we could put out there, and that's 100.

And so we get our estimated total by multiplying the estimated mean of 5.2 by 100, and our estimated total is now 520 plants. But for the confidence interval, we do the same thing. We just take the 2.332 plants per quadrat, which is the confidence interval around the density estimate, and we multiply that value times 100. And now our confidence interval is plus or minus 233 plants.

And note that the population size, N, can be calculated by dividing the total area of the macroplot which in this case is 400 meters square, by the size of an individual quadrat. A quadrat is 4 square meters in size. And this could be done regardless of the size of this area. This area might be several 1,000 or even 100,000 square meters, but you can still calculate big N by dividing the size of the quadrant into that value.

So big N would be very large in that case, but so then would be your estimate of the population size, because you would use that value of big N in the calculation. So here's a graphic example, a bar graph showing our estimate now of the population total. In this case, it's 520 plus or minus 233 plants, so over 90% confident that the true value falls somewhere between 520 plus or minus 233 plants.

So there's your estimated total, and it turns out that this lower value is 287 plants and the upper value is 753 plants. So just a graphic to display differences in confidence level. As we go lower in terms of the confidence, the confidence intervals get narrower. So they look better. They look like we have a more precise estimate, but in fact, these estimates all have exactly the same precision.

In this case, there's only a 50% chance that the true value's within this interval, whereas here, there's a 95% chance that it's within this interval. So recognize that you're not really changing the level of precision at all if you change the confidence level. These days, I tend to gravitate toward 90% confidence intervals, for reasons that will become more clear in Module 4. The most commonly used confidence level is 95%.

So if we have 90% confidence intervals and we did 100 independent samples, so 100 different random samples of n equals 10, we might get the intervals shown here in this graph. How many of those intervals would you expect not to include the true value? So here's the true value, the true mean, which we know in this case because this was computer-generated data.

How many of these confidence intervals, these 90% confidence intervals, would you expect to miss, to not include that true mean? You could see there are some. There's one here that doesn't include it, one here. Well, it turns out that you would expect in the long run 10 of those 100 confidence intervals not to include the true mean, because you're only 90% confident that the true mean is in there.

And it turns out in this particular run that 8 of the 100 samples failed to include the true value. Remember, in the long run, 90% of the time, 10 of those intervals, you are not going to include the value in the long run. But in this case, only 8 of them didn't. If we took samples of n equals 30, well, the confidence intervals are much narrower now, but we'd still expect 10 out of 100 not to include the true value.

And in fact, that's the case here. So here, we had exactly 10 not including the true value. Another thing to notice is that oftentimes, the confidence intervals will be parted with the mean, or estimated total in this case, in the center. So there'd be a little dot right here in the center. I didn't do that, and in fact, some statisticians think that you shouldn't, that the confidence interval should be plotted by themselves to emphasize the fact that the true population value may lie anywhere within the interval.

So that was calculating a confidence interval around an estimated mean, and mean values, means are used to summarize data that's continuous. And remember, we defined continuous data for the purpose of this course as including both integers, account data, for example, numbers that don't have decimals, as well as real numbers, numbers that do have decimals, like heights and weights and those kinds of things.

But now we're going to talk about what do we do when we have just 0s and 1s, so we have binomial data. And we summarize this data by calculating a proportion. So p-hat here, the estimated proportion, is the number of sampling units with the attributes of interest-- in other words, the number of hits, so it might be, for example, the number of quadrants that had a particular plant species of interest-- divided by the total number of quadrats or the total number of sampling units. So in other words, both the hits and the misses.

It turns out we get the same p-hat if we just took a mean. If we just added up all the 0s and 1s and divided by the sample size, we get exactly the same value and estimated proportion. So for example, we might take a random sample of 200 quadrats and keep track of whether species A is present or absent. If 50 quadrats contain that species, then the estimated proportion of quadrats containing the species is 50 divided by 200 or 0.25.

So we'd expect 0.25. Of all of the quadrats that we could possibly put out in the population, we'd expect 0.25 of those to have the species. So that's our estimate of the population, 0.25. But again, this is a sample. And so if we put another sample out there, we might get 0.2, or another sample, we might get 0.3. So we need to quantify the uncertainty around our estimate.

So how do we do that? How do we do a confidence interval around an estimated proportion? Well, one way is to use the normal distribution to approximate the binomial distribution, and we saw in Module 2 that the normal distribution is similar to the binomial as long as the proportion isn't close to 0 or 1 and if the sample size is large enough.

So here's the example we used there. So here's a binomial distribution with a probability of 0.5 and the sample size of 100. Here's a binomial distribution of 0.2 with a sample size of 100, and you can see both of these distributions are pretty close to normal. So if we wanted to use the normal approximation to construct a confidence interval, we first calculate the standard error of the binomial sample, just like we calculated a standard error when we were dealing with means.

It turns out the standard error of a sample proportion uses a different formula. Here's the formula. So standard error of an estimated proportion is the square root of that estimated proportion times 1 minus that estimated proportion over the sample size. So remember, it's the square root of all of that.

And then to calculate a confidence interval using that standard error, you would use the z-value. Now, remember, we said don't use z-value when we were dealing with means. And it's going to turn out that we're not going to really want to do it this way either for proportions, but I want to point this out because you'll see this in many, many textbooks.

So you would use the z-value. The confidence interval, then, is the estimated proportion plus or minus this z-value times the standard error that we just showed you for the proportion. And this is called the asymptotic binomial confidence interval, and asymptotic because it works well for large samples as long as the estimated proportion is not too close to 0 or 1.

And large samples here, remember, we said you could use the z distribution in the case of a confidence interval for a mean if the sample was large enough. Well, the same is true here. The sample is large enough. But it's a big caveat, and we'll talk about that. So if we were to do that for our sample-- remember we had our sample proportion was 0.25, and the confidence interval, therefore, is plus or minus the value of z times the standard error.

So in this case, it would be 1.64, which is the z value, associated with a 90% confidence level. And then here's our proportion, 0.25 times 1 minus 0.25 over a sample size of 200, square root. And if you do that math, you end up with a confidence interval of 0.25 plus or minus 0.05. But this asymptotic binomial confidence interval has some serious limitations.

So the only real advantage of it is that the formula is easy to understand, but it performs poorly when either the sample size is small or the probability, the estimated probability, p-hat, is close to 0 or 1. And because of that, another binomial confidence interval known as the Wilson confidence interval is recommended, and here are a couple of papers on that.

And we provide an Excel workbook on the BLM National Training Center website that allows you to calculate Wilson intervals. We're not going to cover the math behind that here, but just be aware that a workbook is available for you to do that yourself. The R package binom will also calculate the Wilson confidence interval as well as others. And so here are the papers that are cited above if you're interested in exploring this more.

I did want to show you a comparison of the two types of 90% binomial confidence intervals for various numbers of hits, so a number of hits out of a sampling of 50 sampling units. Our sample size is 50, and here we got 0 hits. Down here, every one of those sampling units had the attribute of interest associated with it. So if only 25 of those sampling units had the attribute of interest, then our estimated proportion is 0.5.

Here, our confidence interval, so the asymptotic confidence level, is on the left, the one you calculated using the normal distribution. The Wilson confidence interval is on the right. So it turns out if you're somewhere around 0.5, or even like between 0.3 and 0.7 or so, you know, the confidence intervals are pretty similar, calculated using either method.

But if you get up here, close to either 0 close to 1, the first thing you notice is that the asymptotic confidence interval goes below 0. Here, it goes above 1. Well, you can't have a probability greater than 1 or a probability less than 0, so clearly, that's wrong, whereas the Wilson intervals appropriately stayed either right at 1 or at least right at 0.

But notice another thing that's important here, that the Wilson intervals are asymmetric. Here, the asymptotic ones, you know, you've got the same amount below the estimated values you do above they're symmetric confidence interval, whereas the Wilson ones are not. And it turns out for binomial data, confidence intervals are asymmetric and should be for values that are either close to 0 or close to 1. And these are appropriately asymmetric.

So you want to use the Wilson interval. Here's the bottom line, and again, the Excel workbook will allow you to do that or if you know how to use R, the package Binom will do that for you as well. So I want to talk about finite versus infinite populations. So a population is finite if it's possible to count all of the objects that comprise the population.

So for example, if you have sampling units with area associated with them like quadrats, plots, belt transects, that population is always finite because we can calculate how many of those sampling units could fit into the area from which the sample is taken. A population comprised of individuals, for example, all the trees in a large area or all the moose in a large area, is theoretically finite, but in practice is finite only for a relatively small area.

So unless the area is small enough that we could, for example, map all the trees, and we knew the actual number of trees, we're not going to know in most cases how many trees there really are, how many moose are out there. So even though it's theoretically finite, in practice, it's really infinite because I mean, you just don't know. So a population is infinite if it's not possible to count all the objects within it.

So the trees and the moose, can't count them all. And there are sampling units that are by definition infinite, because if you're dealing with sampling units without dimension, such as point intercepts or line intercepts, then that population is infinite as well. Point intercept, the points are theoretically dimensionless, and so they're by definition infinite. And the y-intercept only has one dimension, so it's theoretically infinite as well.

And as I mentioned, if you can't count all the individuals, it is essentially infinite as well. So if you're sampling from a finite population and you're sampling more than about 5% of the total population, you should apply something called the finite population correction, or finite population correction factor, to your estimate of the standard error. And it will result in confidence intervals that are narrower.

And to see why this would be the case, let's look at our example again. So here's our 20 by 20 macroplot, and we're sampling with quadrats. And in this case, we've got 10 quadrats, and so we've sampled 10% of all the possible quadrats out there. Again, it's largely because this is a small area, but we've sampled more than 5% of the population, so we need to apply the finite population correction factor.

And as just kind of an extreme example of why we would appropriately apply a finite population correction, just be aware that it rewards you for knowing more about the population. So let's say we've sampled 98% of our population. We've sampled 98 quadrats out of 100, and certainly, by sampling that many quadrats, we should a lot more about the population, right.

And that's what the finite population correction factor takes into account. Now, in this case, it's kind of a ridiculous example because it wouldn't make sense if we'd sampled 98 quadrats. We just sample two more and we would have done a complete census, and we wouldn't have to worry about a confidence interval at all. We'd have the actual true value in hand, but this just emphasizes why it's appropriate to apply a finite population correction.

So here it is. Here's the formula. So it's simply big N, the population size, minus n, small n, the sample size, over big N. And it's the square root of that value. So here's an example calculation. So here's big N is how we would calculate it from the area. So let's say in this case, we've got a 20 by 50 meter macroplot. It's 1,000 meter squared.

Each individual quadrat is 10 meters squared, so 1 by 10 meter quadrat, for example. Our sample size is 30, 30 quadrats. So we first calculate big N by dividing the macroplot size by the size of the quadrat, and it turns out there are 100, 100 possible quadrat positions out there. And now we use that. We calculate the finite population correction.

Big N is 100, and then we subtract 30, divide that by 100, take the square root, and our finite population correction is 0.83. So how do we use that? Well, we take the confidence interval, half-width-- remember, plus or minus, so we take whatever that plus or minus value is, and we calculate the confidence interval by taking the standard error and multiplying it by the appropriate t value.

And then we multiply that by the finite population correction, which is a decimal, so it's 0.83. So these were our values. The standard error, this 3 in this case, our 90% confidence of t value is 1.69. So here's the confidence interval half-width calculated without the finite population correction, simply the standard error times that t value, and with the finite population correction, because that finite population correction is less than 1, is 0.83 in this case.

It reduces the size of the confidence interval, and so now, our confidence interval instead of being plus or minus 5 is plus or minus 4. So what about a binomial confidence interval? How would we apply a finite population correction factor to it? The same formula can be used. Big N minus small n over big N, the square root of that value. But because the formula for the Wilson confidence interval that we recommend for binomial data is more complex, we're not going to show here how you apply this to those intervals.

But just be aware that the Excel workbook that we provide on the National Training Center web page allows you to adjust that Wilson confidence interval with the finite population correction. So I want to talk a little bit about the importance of random sampling. All of these procedures require random sampling, and here's an example of why.

So this is taken from a web page that used to-- well, it's a former web page on the United States Geological Survey, that web page that was developed by statistician Paul Geisler. And basically, what he's asking you to do is from this population of 100 numbers, pick five numbers that best represent this population. So what I want you to do is I want you to pause the video here, and I want you to pick five of those values.

See if you can do better than a random sample. I want you to pick five values that you think best represent this population, and then I want you to calculate a mean. Just use a calculator and add up those five values that you've chosen. Divide by 5, and see what mean you come up with. And we're going to see how you stack up against a random sample. So pause the video. Select five numbers you think best represent this data set and calculate the mean.

So how'd you do? Well, it turns out the true mean is 3,725.32. Paul had-- people can actually input their means or input their numbers, and so he had 358 people do that. And so basically, these are called judgment samples, when you don't use a random procedure. You just go out there and you just select five sampling units that you think best reflect the population. These are judgment samples.

So most people's mean was 6,000, whereas a random mean-- so if you took a random sample of 358, that random mean would be 3,416. Notice it misses the true mean, but the difference is much less than everyone's judgment sample. And the reason is people were choosing too many large values.

Because the values differ so much from one another-- there were some really small values and really large values-- it's really difficult to select a representative sample. So most people, so here are quartiles, four quartiles of the data. And the random samples have about 25% of the values, a little bit more, a little bit less in each of those quartiles, which is what you should have.

But the judgment samples from these 358 people, too many people overestimated or picked numbers that were too large, and so their samples were not representative. So that's why you need to use random sampling in your studies, not judgment samples. You can't just go out and throw some quadrats down, for example, in areas that you think are representative, because they won't be.

You need random, some sort of random procedure. So random sampling is a must. All of the statistical procedures discussed in this course depend upon random sampling. Many types of random sampling that can be employed, and here they are. So simple random sampling, which is what we showed earlier when we were sampling the 20 meter by 20 meter macroplot-- it turns out it's not the most efficient sampling design, and so there's all these other types of sampling. Systematic random sampling, stratified random sampling, restricted random sampling, spatially balanced sampling, distance sampling. These different sampling techniques are beyond the scope of this course, but here are some references.

So most of the types of sampling that we listed on the previous slide are discussed in the two books that I have co-authored with Carol Elzinga and Dan Salzer, and in the case of the 2001 book, James Gibbs as well. A recent summary of spatially balanced sampling techniques along with many important references is in this citation I give here.

Distance sampling, which is a major sampling technique used for wildlife sampling in particular, is covered in detail in books by Buckland et al. Adaptive cluster sampling, a technique that is sometimes used for rare plants and other rare organisms, is described in this book by Thompson. And there are all of those references for you in case you want to look at them.

Finally, I want to talk about the importance of setting a sampling objective. It's important to specify a sampling objective prior to initiating sampling, and the sampling objective should specify the amount of uncertainty we're willing to accept when we estimate a mean or a proportion. So when we're estimating a single population parameter, like we're talking about in this module, the uncertainty is expressed as the maximum confidence interval we're willing to live with.

So we specify that objective in terms of a confidence level and a confidence interval width. So here are some example sampling objectives for estimating a single mean or a proportion. And again, we would set this before we go out and sample. So here's an example. Be 90% confident that the estimate of the density of banana slug is within plus or minus 20% of the estimated mean density.

So here's the target confidence level, 90%. We want a 90% confidence interval. And this is the target confidence interval width. We want that interval to be plus or minus 20% of the estimated mean density. Now, we don't know what that estimated mean is yet because we haven't sampled. But when we do sample, we want the confidence interval to be plus or minus 20% of whatever that means density is.

So if the density turns out to be 10 plants per quadrat, we would want our estimate to be plus or minus 2 plants per quadrat. What about a proportion? So here, we want to be, in this case, 95% confident that the estimate of the frequency of the proportion of occupied quadrats of medusahead is within plus or minus 10% of the estimated frequency value.

Well, the 95% confident here means that we want our confidence level, we want a 95% confidence level, and we want our estimate to be plus or minus, within plus or minus 10% of that estimated frequency value or proportion. Now, there's something important to talk about here. In this case, when we're dealing with means and we're saying plus or minus 20%, we're dealing with a relative value, right.

So again, my example here, if our estimated mean density was 10 plants, we would want our confidence interval to be 20% times 10, or 2 plants, plus or minus 2 plants. If our estimated mean density was 20 plants, then we would want this confidence interval to be within 20% of 20, which would be plus or minus 4 plants. So it's a relative value.

When you're dealing with binomial data, on the other hand, like we are in this example where we're dealing with a proportion, we recommend that you do not have a confidence interval that specifies a relative plus or minus 10%. Rather, we want this plus or minus 10% to be an absolute value. So for example, if our estimated proportion, we're saying we want our estimated proportion to be plus or minus 10%, which is the same as saying we want it to be plus or minus 0.1--

If our estimated proportion is 0.4, that would mean that we would want our confidence interval to be 0.4 plus or minus 0.1, right. So it's an absolute value, not a relative value.

So how about determining sample size for estimating a single mean or a proportion? So for a mean, you need three things. You need the target confidence level, which we specified in our sampling objective. You need to target confidence interval width-- for example, plus or minus 30%, which we also specified in our sampling objective-- but we also need an estimate of the standard deviation.

So the first two things, we specified in our sampling objective. The estimate of the standard deviation, we need to get through a pilot study, which we'll talk about here shortly. So for a proportion, so binomial data, and the proportion is the number of 1s out of all sampling units, you need the target confidence level, again, specified in the sampling objective, the target confidence interval width, again, in the sampling objective, and an estimate of the proportion of 1s or hits.

It turns out with binomial data, though, you can get by without that estimate. So you could get this estimate using a pilot sample, but it turns out that if you use 0.5, if you assume that the proportion is going to be 0.5, you can plug that 0.5 into a sample size equation and get a sample size, and it's a conservative estimate.

It turns out that the sample size to estimate a confidence interval around 0.5 is higher than the sample size needed for a confidence interval around, say, 0.1 or 0.9, values closer to either or 0 or 1. So pilot sampling. Calculate the sample size required to estimate a mean. You need an estimate of the standard deviation. So unless you have an estimate from a similar study, which is unlikely in most cases, you need to do some pilot sampling in order to come up with that estimate.

And the pilot sampling should be conducted in the same manner as the planned sampling with a sufficient number of sampling units to ensure that you have a stable estimate of the standard deviation. So at least 10 is what I recommend. It's also useful in calculating the sample size required to estimate a proportion, but again, as I mentioned, the nature of binomial data allows you to conservatively estimate the sample size by assuming the number of 1s in the population, the number of hits, is 0.5.

What about how to calculate the sample size? Well, that's beyond the scope of the course, but there are resources available to you. The book *Measuring and Monitoring Plant Populations* provides equations and instructions on how to determine the necessary sample size to estimate a single population mean or population total with a specified level of precision, or to determine the necessary sample size to estimate a single population proportion with a specified level of precision.

Those, however, they provide the equations, but we also provide an Excel workbook on the National Training Center website that automates these sample size calculations, and instructions are provided there as well. So that is my recommendation, that you go download that workbook, and you can calculate all of these sample sizes yourself.

So I want to end up here with some example data and sampling distributions, just to show you what sampling distributions do for you. So some of the most common data distributions are a normal distribution for many types of data. Height data, I mentioned previously in a different module. Data distribution is often appropriate for proportional data, in other words, data that ranges from 0 to 1. Poisson distribution for count data that's not overdispersed where the variance is not larger than the mean. Or more commonly for count data, negative binomial distribution is appropriate.

And the following graphs are going to show you these data distributions and some sampling distributions taken from them. And as you'll see, many of the sampling distributions approach normality. So here's the original data set for cover of rose species in a particular area. So here's the cover expressed as a proportion. Note that the distribution is not normal.

It's right-skewed, but here are sampling distributions of the means of samples of 25 on the left and samples of n equal 50 on the right. Notice that the distributions are pretty normal. Here's an original data set for a beta distribution with a mean cover of 0.3 and standard deviation of 0.4.

Here are sampling distributions, in this case of sample sizes of 25 on the left and 50 on the right. Again, pretty close to normal. Here's a histogram of agrostis exarata cover. Lots of 0s or values close to 0 in this data set. Very right-skewed. Here are sampling distributions of n equals 25 on the left and n equals 50 on the right.

Although they're not normal, they're a lot closer to normal than this distribution looked. And here is histogram. Here's the original data set of some actual counts, numbers of plants, up here. And here, this follows a negative binomial distribution. And down here, we see sampling distributions of sample sizes of n equals 25 and n equals 50. And again, the sampling distributions are relatively normal.

So the takeaway message here is that the central limit theorem bails you out. Even when original data is far from normal, the distribution of samples drawn from the original data often approximates normality. Sample size plays an important role in this. The further the original data set is from normality, the larger the sample size needed to ensure the sampling distribution approaches normality.

But we can often feel comfortable using statistical procedures that require the assumption of normality, something called parametric statistics, because the sampling distribution follows an approximately normal distribution even if the original distribution of the original data doesn't. So parametric statistics include confidence intervals constructed using the normal distribution, or its close relative, the t distribution.

And as we'll see in module 4, the central limit theorem allows us to use parametric tests. So that's it for Module 3, inferential statistics in estimating a single parameter, a single population parameter. I hope to see you in Module 4, where we'll talk about comparing sample means or sample statistics to determine whether two populations, two or more populations, are different.