[MUSIC PLAYING]

JOHN WILLOUGHBY: Hello. And welcome to a course on the use of statistics in natural resource management. This is module two of the course, descriptive statistics.

You'll see the other four modules listed here. Module one was the introduction to statistics, which I recommend you view before this module. And you can see the remaining modules in the series on inferential statistics.

So today, we're going to talk about descriptive statistics. And we're going to talk about types of numbers. There are three main types of numbers that I'll deal with.

The first are binary numbers, simply 0s and 1s, often used to denote true or false with 1 equals true and 0 equals false. And when a succession of what are usually called trials are conducted, you could also think of this as a sample size where n is some number of observations. Each of those observations, either the attribute of interest is present, in which case, it gets a 1, or it's not, in which case, it gets a 0.

And the set of data that is constructed from this is referred to as binomial data because the data will follow a binomial distribution. And we'll look at what that distribution looks like later. So here's an example.

We might randomly drop 200 pins in an area and record whether the pin hits bare ground. And we record a 1 when it does and a 0 when it doesn't. These data are binomial. And the number of 1s recorded divided by 100 or 200-- I'm sorry-- is an estimate of the cover of bare ground.

Another type of number is an integer. And an integer is a whole number with no decimal. And most commonly, you get integers when you count something. So for example, we might count the number of individuals of a particular animal or plant species in a given area or the number of species in a circumscribed area, which would be the species richness of that area, or the number of flowers on a plant, another example of a count, or the number of disturbances in a protected area.

All of these, they can also be used to rate the quality of something. For example, we might rate the habitat quality of a site from 1, meaning that site is in poor quality, to 10, where it would be excellent. The bottom line is an integer can't have any decimals associated with it.

Real numbers, on the other hand, can or they either have or at least they can have a decimal part. So some examples, well, the heights of plants, the weights of animals, above-ground biomass. Whether the measurements actually have decimals depends on the type and scale of measurement. So for example, if we measure height to the nearest centimeter, our values will have no decimals. But they could have decimals if we chose to measure more accurately, maybe to the nearest millimeter, for example. So those are real numbers.

In this course, we distinguish primarily between binary versus continuous value. So statistics books often make a distinction between discrete or integer values-- discrete is another name for an integer-- and continuous values, which are real numbers. And in some contexts, there are important differences, for example, in statistical distributions, which we'll talk about a bit later.

But many of the statistical procedures we're going to discuss in the two modules on inference are the same for both discrete and continuous values. For example, confidence interval calculations, statistical tests, it doesn't matter in those cases whether the values are integers or real numbers. And so we're going to lump those two types of values, integers and real numbers, under the heading continuous values. We're going to treat them both as a type of continuous value. So the most important distinction in terms of the types of tests and methods for calculating confidence intervals is between binary, or binomial data, and continuous data. And so that's the dichotomy we'll focus on in this course.

So summarizing data, you've got a bunch of data you've collected on something, some numbers, some measurements. And now your job is to summarize that data in some way. And there are three measures of central tendency that are often used.

It turns out two of them are much more common than the third. But the mean is one. The median is another. And the mode, the least common one, is the third measure of central tendency. Although, we'll see that the mode isn't necessarily in the center of the data set. In some cases, there isn't even a mode. But we'll talk about that.

So the most common, and a hugely important statistic, is the mean, also is the average. If you were doing an Excel spreadsheet, for example, they don't use the word mean there. You would use average. They have a command average to use there, a formula.

So let's say we've got seven plots. And we've counted something in each of those plots. And the counts are shown here on the right for each of the plots.

Then we calculate the mean. And this is you're probably all familiar with how this is done. You simply add up the counts and divide by the number of counts. And that gives you the mean, which is 3.3 in this case.

Here is shorthand for that. It may look intimidating at first. But really, this is just a way of telling us we want to do this. So the mean is the simple X bar here. And this symbol here is the summation symbol.

And what it's telling us is that we want to sum all of the xi's. We want to some x1, x2, all the way up to x7. This is saying that the i's are in increments of 1. And so we want to go from i equals 1 up to equals i equals n. In this case, that's 7. And then we simply divide that sum by n. So this formula here is essentially a shorthand for what we see up here.

So one thing here is that the mean is a hypothetical value. It doesn't have to be an actual value in the data set. So for example, if these counts were of banana slugs in a sample of plots in a coastal redwood forest, there's a mean of 3.3 slugs per plot. But counts are necessarily integers. You can't have 3.3 slugs, unless one had some unfortunate accident happen to it. But we still can have a mean of 3.3 even though the counts themselves are integers.

So that was the mean. Now, the median, here's our data set again of banana slugs per plot. And there are the counts. So we would sort these from small to large.

So we go from 0 to 7 in this case. And the median is simply that value, once the data has been sorted, that's in the middle. So in this case, the median is 3. Remember our mean was 3.3, rather, but the median is 3.

Now, no one thing, that this only works if the data set has an odd number of values. Because when that's the case, there's only a single middle value. Sometimes, though, like in this case, we have eight counts. And so we've got an even number here. And so the median in that case is the mean or average of the middle two values. So in this case those middle two values are 3 and 4. And so our median would be 3.5.

The mode is the third measure of central tendency that I'll discuss. It's the most frequently occurring value in the data set. So if this is a data set, we see that the only value that occurs more than once is 0. And so the mode in this case is 0. Notice that's not in the center. So even though the mode is often referred to as a measure of central tendency, in this case, it's not really the center of the data set.

Here's another example. In this case, we have all the numbers are single numbers except for 2 and 7. So there are two numbers that have two values each at that number. So in this case, we actually have two modes, one at 2 and one at 7. So we can talk about this data as being bimodal.

Here's another example. In this case, we have no numbers that have repeated themselves. And so there's no mode. So in this case, the data set doesn't have a mode at all. So as I mentioned, the mode may not be in the center of the data set and, therefore, not really a measure of central tendency.

So which should I use, mean, median, or mode? Well, it turns out that the mode is not used much in practice. Although, people will look at distributions and say, well, that looks like a bimodal distribution. But other than that, it's not used very much in terms of analyzing data, at least not by me.

The median is important, as we'll see, because it's a robust statistic. And we'll talk about what robust means here in a little bit and particularly when we talk about box plots. The mean is hugely important. And many statistical tests employ the mean.

And means can be easily converted to population totals. Oftentimes, we want to know or estimate the number of individuals of a particular species in an area that we're managing. And the median can't be used for that purpose. But the mean can and is.

So here's an example. So we have this area where we've sampled 25 quadrats. And our mean is 31.7 plants per quadrat. Should read quadrat over here rather than quadrants.

But because we know the size of this overall area and we know the size of the quadrats, we can calculate the total potential quadrats in this area that we're sampling. And there are 250 possible quadrat locations. And that's big N as we learned from the first module. Big N is the total number of possible sampling units that you can put in an area. Little n is the number that you've actually sampled.

So in this case, once we know big N, we can simply multiply big N by the quadrat mean. Because we know there's 250 possible quadrats, the mean number of plants per quadrat is 31.7. So if we multiply those together, we get an estimate of 7,925 plants. So the mean is the statistic that is used to determine a population total. And it's also hugely important, as I mentioned, for other reasons too, for statistical testing, for example.

So one thing to recognize here is that people hear the term statistical model, and they think that, all of a sudden, some complex mathematics is going to be foisted upon them. And certainly, statistical models can be pretty complex. But it's important to realize here that the mean, the average, just the simple average itself is a statistical model.

And so as with any statistical model, the mean allows you to predict. So for example, a record album by a recent artist has a 4.7 rating on Amazon. And you might know that Amazon rates there's five possible points. And I don't think Amazon allows 0s. But your rating has to be 1, 2, 3, 4, or 5. It can't be 4.5 or some decimal rating.

So based on that mean of 4.7 out of 5 possible points, what rating would you expect a new listener to give the album? Well, mean is an excellent starting point for that. It wouldn't surprise you if the new listener gave a rating of 5. That's the closest to 4.7.

But a rating of 4 would also make sense. Certainly, it's possible that they would give it a rating lower than that. But your prediction of 4 or 5 is probably a better prediction than a prediction of 1 based on the mean of 4.7.

So similarly, if we obtained an estimate of 31.7 plants per quadrant in a sample of 25 quadrants and then we decided that we wanted to sample another quadrant, how many plants would we anticipate having to count in that new quadrant? Well, a good place to start there again would be the mean of 31.7. And obviously, we can't count 0.7 plants. But we might count 32. We might count 30. We might count 33.

And so the question then becomes, how far away from 31.7 would be likely be? I mean, how likely would it be, for example, to have to count 60 plants or to get a 0 if the mean is 31.7? And that brings us to the question of, how do we quantify error associated with the mean?

With the mean, as with any model, there's some amount of error associated with the estimate. So how can we quantify it? Well, let's look at another example.

Let's say that we measure the clutch size in a sample of European starling nests. Maybe we've got a problem with starlings. And we want to see how successful the nests are.

And so we go out there, and we measure 5 nests. And in nest 1, we see that there's 2 eggs. Nest 2, there's 8 eggs. Nest 3, there's 7 eggs and so on.

So we measure those. And we calculate a mean. It turns out the mean is 4.8 eggs. Now obviously, no nest has 4.8 eggs. But that's the mean. And the mean can have decimals.

So if you have a mean of 4.8, what is the likelihood of, if we measured another nest, how many eggs might you expect to find in that sixth nest? Well, 5 certainly wouldn't surprise you. And 4 probably shouldn't surprise you. And maybe 6 and 3 wouldn't surprise you either.

But how can we quantify this? So here are the differences between the mean and the actual values that we got. So we can see that we got pretty close in nest 4 to that mean. But here in nest 2, there were 3.2 more eggs than the mean.

So these are errors. These are not errors in the sense that you didn't measure correctly. They're errors in the sense that they are this far, each one of these is this far from the mean value. And so what we're trying to do is quantify that error in our model.

And so one option might be to use the total error. So here, we have our 5 nests. Here are the clutch sizes. Here's the mean. And here is the deviation from that mean.

So we can see here that in nest 1, the clutch size was only 2. So we were 2.8 eggs lower than that mean. Whereas in this case where we had a clutch size of 8, we were 3.2 eggs higher than the mean. So in some cases, our error is negative. In some cases, it's positive.

And we could add all those up. But the problem is they add up to 0. They're always going to add up to because they cancel each other out. So what can we do instead?

Well, we use the squared error. So here's that same table. And here's our deviations from the mean. Here's the error. But we square that.

When we do that, all of the numbers are now positive. And so we add those up. And here is the total squared error for our particular model. The error values are now positive.

And the total of those error values is called either the sum of squared errors or, more simply, the sum of squares. And here's the formula. Again, it may look intimidating. But remember this is shorthand.

This is the summation sign. SS means sum of squares. Here's each of the individual X values shown up here, the nest numbers. Here are the values associated with each of those.

So you take these values. You subtract the mean, which is X bar here, square it. And you do that from i equals 1 up to n, so from 1 to 5. And so essentially, that's shorthand for that formula, which is much harder to write out, particularly when if you've got X's that go all the way up to 60 or 100 or something. So that's why we use this shorthand notation.

Now, a computer program is going to do this work for you. But this is what it's doing. So OK, so we have this squared error now, the sum of squares. But the problem with the sum of squares is that it gets bigger with the number of observations. It keeps going up. It's not going to ever level off.

And so we need a way to take the number of observations into account. And we do that by dividing the sum of squares by the number of observations. And it turns out that when you have a sample, instead of dividing by n, the number of observations, you divide by 1 less than the number of observations, or n minus 1.

I won't go into the reasons for that here. But if this were a population, if this were all of the nests in an area that exist, then you would divide by n rather than n minus 1. But here, we're going to divide by n minus 1.

And in the context of the mean, this statistic now that we've calculated by dividing the sum of squares by n minus 1 is called the variance. And the variance of a sample is denoted by s squared. And here again is the formula. Here is the sum of squares formula that we saw earlier.

And now we're dividing by n minus 1. And so our sum of squares was 26.8. We had 5 nests. So n minus 1 is 4. And so our variance is 6.7.

6.7 what? You might ask. Well, that's a problem. Turns out that the variance is in squared units, the number of eggs squared. Since that's not very meaningful, we take the square root of the variance. And that number becomes the standard deviation, a hugely important statistic along with the mean. Probably the most important statistic most those two, the mean and the standard deviation.

And again, here is our variance formula. But notice, this time, we've taken the square root. And so we end up with 2.6 as the standard deviation.

And that standard deviation is in the same units now as our mean. So we can say that our estimate of the clutch size in the 5 nests that we sampled is 4.8 eggs per nest with a standard deviation of 2.6 eggs per nest. And so here are those values.

Now if we sampled another nest, how many eggs would we expect to find? Well, our best guess is 5 eggs because that's the closest to the mean. But how likely would it be to find 6 eggs or 8 eggs, 10 eggs, 2 eggs? And that's where the standard deviation comes into play. It tells us that it probably wouldn't be particularly unusual to find 2 eggs, since 4.8 minus 2.6 is 2.2, or 7 eggs but less likely to find 1 eggs or 9 eggs.

So let's look at another example. So let's say I measure the height, in centimeters, of 50 plants. I calculate the following mean and standard deviation. The mean is 50 centimeters. The standard deviation is 10 centimeters.

If I decided to measure another plant, how tall is it likely to be? Or stated another way, how tall is the typical plant? Because heights typically follow a normal distribution, we can use that distribution to determine how tall our additional plant might be.

So here are those heights. Notice I've measured these actually to decimal points. But the mean is exactly 50. The standard deviation is 10.

And if the collection of heights follows a normal distribution, as height measurements often do, we can use that distribution to predict how likely it would be to get heights of a particular size. So the normal distribution is also called a bell curve or a Gaussian distribution. It's a probability distribution that can be used to describe many types of naturally occurring data.

But the normal distribution is completely described by two parameters, the mean and the standard deviation. So if you have those two parameters, you can draw the normal distribution. And so here's an example of using the mean and standard deviation of our plant height data.

So we can see that here's the typical bell-shaped curve. Here's our mean of 50 centimeters. The standard deviation in this case is 10 centimeters.

Here are so again the same figure, the same normal distribution, same mean and standard deviation. But now we can see that, since the standard deviation is 10, here's plus 1 standard deviation greater than the mean. Here's plus 2 standard deviations. Here's plus 3. Here's minus 1, minus 2, and minus 3 standard deviations.

So it turns out that 68% of the values are going to be within that plus or minus 1 standard deviation from the mean. 95% of the values are going to be within plus or minus 2 standard deviations from the mean. And then finally, 99.7% of the values are going to be plus or minus 3 standard deviations from the mean.

So notice that this changes depending on how variable the data are. So here, the standard deviation is now 5. And so the mean is still 50. But the standard deviation is 1/2 of the previous slide.

And so the distribution is still normal. But notice that the curve is much narrower. And now 95% of the heights are within a 40 to 60 centimeter range. So that turns out that that's plus or minus 2 standard deviations. So 68% of the heights are between 45, 55. 95% between 40 and 60.

Well, what if the standard deviation was larger? What if it's 15 instead of 10 or 5? Well, the curve is now flatter. And now 95% of the heights are now between 20 and 80 centimeters, so much more variability in terms of the number of height or the size of the plants here.

So back to predictions, if we were going to measure another plant, how likely would it be to be a certain height? Well, you can see it depends on the size of the standard deviation. If you have a larger standard deviation, you're going to have a bigger range of variability.

And so you wouldn't be, in this case, for example, greatly surprised to have a plant of 40 or 60 or even 30 or 70 centimeters. Whereas in this case, you would be surprised perhaps to have that additional plant be smaller than 40 centimeters or greater than 60. Because, again, 95% of the values are between 40 and 60 in this case.

So the coefficient of variation, well, it turns out the standard deviation is, as I mentioned, a very important value because it's expressed in the same units as the mean. But its magnitude can be understood only in the context of that particular mean. What do I mean by that?

Well, let's say that we measure the weights of 30 mice and the weights of 30 elephants and obtained the following means and standard deviation. So for the mouse, the mean weight is 40 grams with a standard deviation of 15 grams. For the sample of elephants, the mean is 5,444,000 grams with a standard deviation of 1,500,000 grams.

So if we just looked at the standard deviations, we might think that the elephant is more variable, right? It has a much bigger standard deviation than the mouse. But the problem is it's relative.

The coefficient of variation, on the other hand, puts the variability on the same scale. And it does that by dividing the standard deviation by the mean. So if we do that for the mice, we see that the coefficient of variation is 0.375. For the elephants, it is 0.276.

So despite the much larger standard deviation, the sample of elephants is actually less variable than the sample of mice. And you should be aware that coefficients of variation are often given as percentages instead of proportions. For example, in this case, we'd say that the coefficient of variation is 37.5% in the case of the mice and 27.6% in the case of the elephant.

So coefficients of variations for monitoring data, well, it turns out James Gibbs and associates analyzed 512 published studies that tracked the trend of plants or animals over at least a 5-year period. And they found that the mean coefficients of variation averaged by type of plant or animal ranged from a low of 0.14 to a high of 1.31. Individual study coefficients of variations ranged from a really low value of 0.02 up to a very large value of 8. So if we expressed this in percentages, that would be an 800%. So that's a very large coefficient of variation.

In my experience with monitoring plants, a coefficient of variation of 0.5 or less is close to the best you can expect. And higher coefficients of variations are common. For example, in a recent plant-monitoring study that I was involved in, coefficients of variation for the estimation of cover in 1 meter by 5 meter quadrats ranged from a low of 0.37 for estimating litter cover, which turns out to be litter was more evenly distributed throughout the area, to a high of 1.81 for woody debris cover. And the reason for that is the woody debris only occurred in a few quadrats. But where it did occur, it had high cover values. But many quadrats had values of 0. So a very high coefficient of variation in that cover. Here's the reference for the Gibbs article that I cited above.

So now we're going to segue into exploring your data using graphs. It's very important to know what your data set looks like. And you should be aware that means and standard deviations are great summary statistics. And you need to use them.

But you also need to understand that they may do a very poor job of characterizing the actual data distribution of your data set. So here's an example of four data sets. Each one has a mean and standard deviation or a mean of 100 and a standard deviation of 10.

So here, I am showing the means and an error bar around those means that is the standard deviation. So here are our four sets of data. So you would think, well, the data distribution is going to be pretty similar, maybe even the same, right? Because the mean and standard deviations are all the same.

Well, that's not the case. Here are the different data distributions associated with these means and standard deviations. So up here, we have a data set that is pretty uniform from the lower values to the higher values.

Whereas here, we have a bimodal data set. Here, we have most of the data is down here congested together, but there are two outliers up here. This data set is more regular but still is less regular than this one.

The bottom line is all of these four data sets are quite different in terms of the distribution of the data. And you wouldn't know that if you just looked at the means and standard deviations. So I just want to highlight the importance of looking at your data graphically.

So let's talk about some commonly used graphs to look at data sets. And one of the most common is a histogram. And here's a histogram of plant heights.

And when you've got it-- so essentially, to understand this, each of, these they're called bins. They look like bars. You might call them bars. But they're referred to as bins when you're talking about a histogram.

So in this first bin here, we have a count of two plants. And we see that those two plants, the height of those two plants is somewhere between 25 and 30 centimeters. We don't know whether they could be 26 both of them. They both could be 29. We don't know. But we know they're somewhere in that range.

Over here in this tallest bin, we see we have a count of 11 plants following somewhere between 50 and 55 centimeters. So it shows you the shape of the distribution. So here, we have a bin width of 5 centimeters.

But we can choose the bin width we want. So here, I've chosen a bin width of 2.5 centimeters. So now we can actually drill down to and figure out more closely where, for example, our counts of 2 fall in this first bin. We now know that they fall some where between 27.5 and 30 centimeters. Whereas before, we just knew they were somewhere between 25 and 30.

But we still can't distinguish actual data points. So here, for example, we wouldn't know whether both of them were 28, if one was 28, and one was 29, and so on. We can overlay the histogram with a normal distribution.

And I want you to notice now the axis is now density. We have to change the axis in order to draw this normal curve. And by density, it means if you added all of the area up here, it would add up to 1. So you can see the heights here approximate a normal distribution.

We could put something called a rug plot at the bottom of the histogram. A rug plot shows you the data point. So here now in this first bin, we knew that there were two somewhere in that bin before.

But now we can actually see where those two data points fall. So one of them is somewhere around maybe 27, maybe at 28 centimeters in height. One of them is probably 29 point something in height, close to 30.

And so now we can distinguish the individual data points, at least most of them. Recognize though that some of these, if they are here, for example, we may have a situation where there's over plotting. So we might have two or three or four values that we can't distinguish from each other because they're over plotted on one another because they're the same value. But it allows you at least to get a better feel for the actual data values themselves looking at this graph.

We could dispense with the bins entirely and just do a density plot to look at where most of the values and the shape of the distribution. This is sometimes called a kernel density plot where the kernel describes the method used to create the smooth line. In this case, the method is Gaussian, which means draw normal distribution or draw as close to a normal distribution as the data will actually allow you to. And that's what we've done here.

Here's a density plot overlaid by the theoretical normal distribution. And you can see that our data set pretty close to normal. We could combine all of those things. Here's a combination histogram, density plot, and rug plot.

Another type of plot is a dot plot, sometimes called a Wilkinson dot plot after Leland Wilkinson who invented it. They're also called dip plots in some instances. I think that's what Leland Wilkinson called them.

But here, we can see the individual data points. We can see the distribution of plant heights and their individual data points. So it's a good plot for that.

We can also do something called a normal probability plot to look at whether our data follow a normal distribution. So these are also called normal quantile-quantile plots or normal Q-Q plots. And essentially, what is happening here is that the data are sorted in order from smallest to largest.

Each of those is a quantile. So if there are 50 measurements-- and there are in this case, 50-- we have 50 quantiles from small to large. And so here are the sample quantiles, the actual quantiles on the left. Here are the quantiles we would expect, if this were a normal distribution, on the x-axis.

And if this follows a straight line or if it approximates a straight line, the data can be considered to be normal. And here, we've also drawn a 95% confidence band. We'll talk about confidence intervals later. But if the values fall within this band, we could be 95% confident that they're within the range of a normal distribution in this case is one way to interpret this.

So these are plant heights. If I look at plant areas instead, it turns out areas don't follow a normal distribution, oftentimes. And that's the case here. And we can see that many of these data points, both at the high end and the low end, aren't within this confidence band and, therefore, can be considered to not be following a normal distribution.

And here's a histogram of the plant areas that we just saw in the normal probability plot. And we can see, from the histogram, again, it's not normal in terms of its distribution. So that brings us to box plots.

And now we'll talk about why the median is important. So box plots were introduced by John Tukey in his classic 1977 book *Exploratory Data Analysis.* And it turns out they're a very robust way to visualize the data set.

And they make use of what's called the five-number summary, the minimum, the maximum, and the first, second, and third quartiles. So why robust? Why are they considered robust, the box plot, the mean versus the median?

When data are normal, the mean equals the median. They're both the same if the data are exactly normal. But when data are skewed, the mean is affected more than the median.

And here's some examples. So here are plant heights and plant areas. So the plant heights, we can see that the solid red line, in this case, is the mean. The dashed green line is the median. They're both about the same. They're not exactly the same. But they're close.

If we looked at areas, on the other hand, again, the green dashed line is the median. And the red line is the mean. So now we can see, because we've got these large values out here to the right, if this is a right-skewed distribution. These large values are drawing this mean to the right. Whereas they're not affecting, they haven't affected the median, at least not to any great degree.

And let's see why that might be the case with another example. So let's say we take a sample of five people and ask them their net worth. And here's what they told us.

Well, we had a range of $11,000. Here's an individual not doing too bad, $210,000, pretty good. Here's a person that seems to be doing really well. And his net worth is $192 billion.

Well, the mean here is $38,420,000,000. The median, on the other hand, is $99,000. Remember the median is simply the middle value when you arrange this data set in order. So it's $99,000.

And notice how the mean has been affected by this one individual who, unfortunately, the data included Jeff Bezos. So it skewed the data very markedly to the right. That's the difference between the mean and the median. So because of that, the mean is not considered to be a robust statistic. But whereas the median is.

So robust measures of dispersion, so just like the mean, the standard deviation is also sensitive to departures from normality. So if we can use the median instead of the mean or at least in addition to the mean, we're probably still going to use the mean in most cases. But if we use the median instead of the mean because it's more robust, what can we use in place of the standard deviation?

And the answer is we can use quartiles, which split the distribution of data into groups of equal numbers. Quartiles are used in the construction of box pots, which are a great way to explore your data set. So let's look at some box pots.

So here's a couple of box plots. Box plots are often also called box and whisker plots because we have this box. That's the box. And then we have these whiskers on the box.

And so here the box is kind of centered. And the whiskers are kind of evenly spread on both sides. In this case, these were plant heights. With plant areas, you can see the box is kind of more truncated down here. And now we have these dots out here, which are called outliers. So let's look at the anatomy of a box plot.

So here's our first box plot. And so this value in the center of the box is the median. So it's the center. So 50% of the values in this data set are to the left of that median, and 50% are to the right.

The edges of the box are called the lower quartile, or 25% quantile, and the upper quartile, the 75% quantile. So in this case, we can say that 25% of the values in this data set are below this 25% quantile or quartile. And 75% of the values are below the upper quartile.

And so this value here, the interquartile range consists of 50% of the values in the data set. But the ends of each of these whiskers, this is the lower inner fence, and it's no more than 1 and 1/2 interquartile ranges below the lower quartile. The upper inner fence is no more than 1 and 1/2 interquartiles ranges above the upper quartile.

So that's the anatomy of that box plot. But now let's bring in the plant area box plot. And we'll see now that these values that are beyond either the lower or sometimes they can be lower than the whisker, but, in this case, they're all above the upper fence, or these are above the upper inner fence. Here's the inner fence.

And so if they're more than 1 and 1/2 interquartile ranges above this upper quartile, then they're mapped as outliers. And those are values that, typically, you're going to want to examine because it could well be that maybe this value up here around 8 was a mistake. Maybe there was a transcription error. And instead of 8, this is supposed to be 1.8 or something. But you need to at least examine those to see if they're real values. It turns out they were real values in this particular data set.

So that's the anatomy of a box plot. And here's a box plot. We can actually overlay the box plot with the actual values. And so now we can actually see the individual data points as well as the summary that the box plot allows us to see.

So if we were assessing normality with a box plot, this box plot, the distribution is probably close to normal because we've got the median is almost in the center of the box. The box itself is almost in the center of the entire set of data. And so it's likely this is approximating a normal distribution.

Whereas this box plot clearly isn't. The median is way down here. It's not in the center of the box. The box itself is way over here on the left. And we've got these four outliers.

We could also plot box plots by category. So here, we've got measurements of cactus stem diameter. And these were made. This is by geologic type. These are geologic formations here.

And we've done box plots of the data within each of these geologic formations. And we can compare the distribution of the data between those formations. And it turns out, in this case, the width of the boxes is proportional to the square root of the sample size.

So we can look and see that, in the Cedar Mountain geologic formation, for example, our sample size was not nearly as large as it was in the Morrison Formation. The reason for taking the square root of the sample size, we'll examine that later. It turns out that that's a very important statistic is the standard error, which is the standard deviation divided by the square root of the sample size. But we'll talk about that in the next module. But that's the reason for the square root of the sample size.

So those are some graphs you can use to examine your data sets and should use. Now I'm going to segue into statistical distribution. We've already looked at the normal distribution.

It's defined by two parameters, the mean and the standard deviation. So if you have those two parameters or estimates of those two parameters, you can construct a normal curve. It's arguably the most important distribution used to model data.

But there are other important distributions that you should be aware of, the binomial distribution for binary data, 0s and 1s, which we talked about, the Poisson distribution for count data, the negative binomial distribution for overdispersed count data-- we'll talk about what overdispersed means-- and the data distribution for cover data. Those are all important distributions that you should be aware.

So the binomial distribution models data consisting of a number of trials in which the response is either 1 or 0. So what are some examples in everyday life? Well, if we flipped a coin 10 times, in other words, there were 10 trials, and we recorded how many flips were heads or tails, that result would follow a binomial distribution. How many free throws does LeBron James make out of 10 tries?

Some biology examples, let's say we sample 10 or 50 deer and keep track of whether each one has a particular antibody. Or we keep track of how many nests were successful out of a sample of 25 nests. Or we sample 200 quadrants and record how many contain plant species A. These are all examples of binomial distribution.

So note that binomial data involves probabilities. For example, if 15 out of 25 nests were successful, our estimate of the success rate is 15 divided by 25, 4.6. This estimate of a proportion is denoted p hat. And it's analogous to the mean of a continuous data set.

In fact, you can calculate p hat by taking the average of all the 0s and 1s in your binomial data set. You get the same answer. So the binomial distribution is described by p hat, so what the proportion is, the estimated proportion, and the number of trials. So let's look at some examples of this.

So here, we've got a binomial distribution with a probability of 0.5. And we had 100 trials or a sample size of 100. So this would be an example of the distribution we would get if we flipped a coin 100 times and kept track of the number of heads and the number of tails.

And we would see that, on average, the probability would be 0.5 of getting a head or a tail. But sometimes-- you know about getting 50 heads or tails, right? So but sometimes, you're going to get 40 heads and 60 tails. Sometimes we're going to get 60 heads and 40 tails and so on.

And so the probability ranges from a low of just over 0.3 to a high of 0.7 or less. And with the highest probability being around 0.5 in the center. So that's what the distribution looks like.

Notice that it's pretty close to normal. So it turns out when the probability is near 0.5, the normal distribution does a pretty good job of modeling the binomial distribution in this case. So what if the probability is now just 0.2 of something happening?

We saw 100 trials or a sample size of 100. And in most of the cases, 20 out of those 100 are going to have the item of interest, or they're going to be 1s. And 0.8 won't. But again, there's variability associated with it.

Notice again distribution is still rather normal as long as the number of trials or the sample size is 100. But if the sample size is lower, here, the probability is still 0.2. But our sample size is only 20. Now the distribution is skewing off to the right.

But what about when the probability is very close to 0 or 1? Here, the probability is only 0.01 of something. We still have 100 trials. Well, it's not even close to normal anymore even at n equals 100.

So it turns out that normal distributions can successfully model the binomial data set when probabilities are greater than about 0.2 and less than about 0.8. But it fails at probabilities closer to 0 or 1. And it turns out the beta distribution is actually a better model of binomial data because it works at all probabilities. And we'll talk about the beta distribution here in a couple of minutes.

So the Poisson distribution, Poisson distribution models discrete data like counts. And you can describe the Poisson distribution by a single parameter, the mean. And the reason for that is that the variance is equal to the mean in a Poisson distribution.

So if you know the mean, you know the variance. And remember the standard deviation is the square root of the variance. So for example, if we had a Poisson distribution with a mean of 4, our variance would also be 4. And the standard deviation in that case then would be the square root of 4, or 2.

So here's a couple of examples of a Poisson distribution. Here's one where the mean is 4, which means the variance is, therefore, 4.And the the standard deviation then is the square root of 4, or 2.

And here's what it looks like. Notice not too many 0s or values near 0. Here's a situation where now our mean is 16. Our variance, therefore, because it has to equal the mean in a Poisson distribution, is 16. And the standard deviation is 4. Notice no 0s in the data set in this particular case.

Although these distributions aren't normal, they're closer to normal than many count data sets in real life. And the coefficient of variation, remember we talked about coefficients or variation for monitoring data. And I said it was very unusual in my experience to have coefficients of variation less than 0.5. And often, they're worse than that.

And in this case, the Poisson distribution for this example on the right, the standard deviation was 4. And the mean is 16. So the coefficient of variation is 4 divided by 16 or 0.25, which is lower than most real data sets.

So it turns out the Poisson distribution is often not a good distribution to model counts with. And we'll see why here. Here's a histogram of what I think is a more typical count data set with, here, the coefficient of variation is 0.75.

So here's the histogram of the actual counts. Here's a theoretical Poisson distribution superimposed on that. And so the problem here is that the Poisson distribution expects a mean of 20 to have a standard deviation of the square root of 20. Remember the variance and the mean are equal. So the standard deviation should be 4.47. But the actual standard deviation here is 15.

So the Poisson distribution expects no 0s, for example, in this data set. Whereas we have quite a few 0s or at least numbers that are 0 or close to 0. So it doesn't do a good job of modeling this data set.

Instead, distribution that is often used and better at modeling count distribution is the negative binomial distribution. And here are the reasons just to reemphasize why the Poisson doesn't work is that standard deviations of most count data sets are higher and sometimes much higher than the Poisson model predicts. So when the actual standard deviation is larger than the model standard deviation, the data are referred to as overdispersed.

And in fact, it's not uncommon for count data, or at least the kind of count data used in natural resource management, to be overdispersed. And negative binomial is a better way to model that kind of data. Unlike the Poisson distribution, the negative binomial distribution allows for the standard deviation of the count data to vary. So here's a negative binomial curve fit to our count data. And you see it's almost a perfect fit, much better fit than the Poisson. Another example here with a different mean and standard deviation, but again a pretty good fit.

And finally, we'll talk about the beta distribution. The beta distribution has not been given the due it deserves until fairly recently. It turns out the beta distribution is a really good model of proportional data, in other words, data that ranges from 0 to 1. This includes estimated plant cover data, either basal cover or canopy cover, as long as you express the cover as a proportion and not a percentage. In other words, we would say the cover is 0.25 instead of 25%.

The beta distribution is defined by two shape parameters that are often called alpha and beta or simply shape 1 and shape 2. And these parameters, fortunately for us, can be estimated from the mean and standard deviation of the data set that you have. So here are some different data distributions.

All correspond to a mean cover of 0.3 but with different standard deviations. So here, we have a mean cover of 0.3 and a standard deviation equal to 0.3, which means the coefficient of variation here is 1. And we can see that the distribution is markedly skewed to the right.

Here, we have a data distribution of 0.4. So the standard deviation is even greater than the mean. And you can see here that we have 0 values or values that are close to 0. But we also have quite a few values close to 1, a distribution that is not uncommon when you're monitoring, for example, cover data.

You might have a plant that's not well-distributed throughout the area that you're sampling. Therefore, you get a bunch of 0 values. But where you do get it, you get really high cover values in some of the quadrants. And therefore, those are 1s or close to them.

And then here, you see standard deviation is smaller and getting more normal. When the standard deviation is really small compared to the mean, it's almost normal. And by the way, this is a really good paper on using the beta distribution to analyze plant cover data. And these graphs were inspired by an example given in that paper.

Some more examples of the beta distribution, here's a histogram of*Agrostis exarata* cover. The red curve is the theoretical data distribution. The histogram are the actual data. And you can see that that data set was modeled pretty well.

And down below is the histogram of rose species, actually, more than one rose species that were lumped together here in the analysis. Then you can see that the beta distribution, the theoretical distribution shown by the red curve models the histogram the actual data shown in the histogram pretty well.

So that's it for statistical distributions. And that's it for this module on descriptive statistics. And so I hope to see you in module three where we'll talk about the use of inferential statistics to estimate a population using a single sample.

[MUSIC PLAYING]