Lecture 19: Descriptive statistics

Slide 1

Today we're going to talk about descriptive statistics and how to look at the patterns in your data. These provide simple summaries of the data and form the basis of qualitative data analysis. Here we are simply describing the data. In the next session we are going to look at how to use inferential statistics to test our hypotheses.

Slide 2

So in this lecture we're going to understand the different scales of data.

We are going to understand why central tendency is important, we will know what variability is, and we will be able to look at the data and understand the distribution.

Slide 3

There are four main scales of data, and your data when you collect it will fall into one of these categories. And so when you know which category your data is in, if it’s nominal, ordinal, interval or ratio, then you can start to make decisions about how to present it, to visualize it and how to analyse it.

So nominal data is simply categories. And these categories, importantly, don't have a set order. So it's things like colours, or true and false and so on. There's no right or wrong way to order these data.

The next is ordinal. And these categories, again, but there is a logical order. So it's things like small, medium and large, or very rare to very dominant.

Interval scale is numbers. But while you know how far apart the numbers on the scale are, you cannot add or subtract them. And examples are temperature or day interval. So there is no absolute zero point. With temperature, 0 ˚C is not an absence of all temperature. And what that means is we cannot make statements such as 10 ˚C is twice as hot as 5 ˚C, because that's just simply not true. You can't sort of double and that kind of thing. And we can't subtract or divide. All we know about these data is that the values are evenly spaced. This means that the difference between 1 ˚C and 2˚C is the same as the difference between 9˚C and 10˚C. Dates or times are also on the interval scale, there is no zero date or zero time so we cannot double or multiply one date by another.

A ratio scale is where we can multiply or double and so on. It does have an absolute zero. So it's continuous data. These numbers are continuous they have an absolute zero. And there is an intuitive rank order. It's things like length, height, weight, you can multiply, you can add them together, you can perform simple maps on these On these numbers they received if they're on a ratio scale.

Slide 4

When we collect data, we must distinguish between precision and accuracy. These are two very important concepts precision is when values in your data are close together. So if you have a drought treatment and you measure the soil nitrogen in response to the drought and all your values were very close to four milligrams per kilo. If all of your data were very close to this one number under drought, you could say that these data are precise.

Accuracy is when the values are close to the true value. Low accuracy means that you are not getting values that are close to reality. So what that means on this figure is that when you have high precision and accuracy. We have the top left picture. You can see that the data points are tightly clustered so they have high precision but they are nowhere near the bullseye. So they have low accuracy.

On the top right, this target, we can see that they are very tightly clustered and they are right in the bullseye. So we have both high accuracy and high precision.

If we look at the bottom left, we can see that they're very scattered data points and none of them are anywhere near the bullseye. So we have low accuracy and low precision.

In the bottom right, we have still quite scattered data points, they are quite far apart, but they are essentially all around the bullseye. And if you were to draw lines in towards the bullseye. You would probably be pretty close to the central point. So you see it's quite accurate, but we have very low precision because there's big scatter.

So obviously when we do our experiment we are aiming for both high precision and high accuracy in our data. So we know that we can have faith in the output in the results. Low Precision causes problems for statistics because if the points are very scattered it can be hard to detect a difference between your treatments. So for example, if we look at this bottom right, and let's say that all of these dots are the treatment. And then we had another cluster that was equally spaced apart that was the control treatment. They would probably overlap, to some extent, and it will be very hard to tell whether there was a true difference because there's so much scatter, so much uncertainty in this data. We can't be certain that there is a strong effect.

So low accuracy is also a big problem. If the measurements are always wrong by the same amount, so let's say your clock is always two minutes fast, and you can adjust for that, you can be sure that you know you just subtract two minutes. This is known as bias, but if the numbers are always all over the place, and there's no systematic bias, that's a big problem in your data, you can't have much faith in your results and this can mean very big problems for your analysis.

Slide 5

When we collect our data, and we have put it into the spreadsheet ,checked for missing or obviously wrong numbers or labels and stored the meta data, which is the information about dates and users and all sorts of information about how we collected the data, then we need to then look at its characteristics. We will do this individually for each dataset. So let's say we measured soil nitrogen, soil carbon, soil phosphorus, we will look at the nitrogen alone, and then the carbon alone, and then the phosphorus alone. So we look at them all separately. And of course here I'm talking about the response variable and we have to assume that our independent variables have some kind of control or we've controlled for that. So we're talking about the response. So there are three main characteristics of interest when we look at univariate statistics.

There are three main characteristics of interest. We look at the overall distribution of the data, the central tendency, and the dispersion or the spread of the values

Slide 6

The distribution of the data is where the individual values fall within the range. The simplest way to look at this is to list each value of the variable and the number of times this value occurred in the dataset. It is more sensible, though, to create a range of values. So small categories within your range,

and put your data into these categories. So it might be 0-5, 6-10, 11-15 and so on. So then you would say, right, my first data point is 11 it goes in the 11 to 15 category. So I'll show you what I mean by this. Here I've created a set of random numbers for this data set. There's 20 numbers here- 20 values. I've written n = 20 so number = 20. And it's a small n, it’s not capital N because our data here are 20 numbers are a subset of a bigger population. So we have sampled within the larger population, we've taken 20 values.

Slide 7

Here I have presented a frequency table of the data. I've put the raw data at the bottom, our 20 numbers. I've put them at the bottom of the table underneath for reference, and you can see here that I've created 10 ranges. I've made tally marks for each data point in its range. So, for example, our first number here is 11 on our list, and it would go into the 11 to 20 range and so it gets a check. So one tally remember 36 will go into the 31 to 40 so it gets a check. A tally mark in that box. So then I count how many data points fall in each of these ranges and the overall percentage. So 11 to 20 has three individuals. So we can see that there's 11, a 12 and a 20. Those three individuals 11, 12 and 20 form 15% of the overall 20 data points. Each of these ranges is called a frequency class and the classes of intervals of 10. Now, you might want to increase precision and reduce the intervals to five. So you would have more classes. If we had more data. This might have been useful.

This technique is useful for data of counts like numbers of bird eggs in a season and we used it for our lecture on life tables back in the population ecology week. So if you remember that was an example of this kind of data that presentation.

Slide 8

A very important part of descriptive statistics is finding out if there is central tendency. We estimate the centre of the data distribution. There are three ways to do this, we can find the mean, the median or the mode. The mean is the most commonly used method, you simply add up all the values and divide by how many values there are. I have presented two types of means, here they have different mathematical notations, but the calculation is the same. On the left, we have the population mean which is signified by the Greek letter mu (μ). This means we have measured every individual in our population. In practice, this is not really possible. And so we sample a representative set of the population. The sample mean is denoted by x bar. So the x with the line on the top is x bar.

We start at the top of the equation with the numerator. X is a standard notation for every value of the variable we have measured. So each one of our 20 values. This symbol that looks like a capital E is the Greek letter sigma and it means to sum or to add all together. So we add together every value of our data set. We then divide by n, which is the number of values we have in the data set. In our case, it is 20.

Below, you can see how I've added all the values to get 912 and then divided by 20 to get 45.6.

Slide 9

The median is the middle value of all the values. You would arrange them in rank order from the smallest to the largest and the middle value is our median. In this case, we have an even number of value. So the median would be the average of the 10th and 11th value. Here we can see that they are the same. So the median is 46. It is quite close to the arithmetic mean 45.6 but this is not always the case. The median is an ordinal statistic, which means we only need to know the ranks of the values. This is useful if we don't know all the values or if there are very big or very small values that are very distant from the other values. These are called outliers.

The mode is the most frequently occurring value in the set. In our dataset 36 and 46 both occur twice.

Our dataset is therefore bimodal meaning two modes. The mode is the only measure that can be applied to ordinal scales. So the categories that have no rank order.

Slide 10

When we are first describing our data, it is helpful to make a histogram like in this figure. The data are once again arranged into frequency classes and the frequency on the Y axis is how many data values appear in each class. This graph is not based on our data set, because I wanted to demonstrate the idea of the bell shaped curve. And as we saw before our test data set is bimodal. This graph shows a bell shaped distribution into the data ranges. This is also known as a normal distribution or Gaussian distribution, because it has based on the mathematician Gauss who had a theorem about the properties of the shape of these graphs. So if you have a perfectly symmetric distribution with the same number of values on each side in each category, the mean, median, and mode would all be the same value. In biological distributions, we often do not get this nice central tendency often most of the data fall on the left side, which means a lot of zeros or very low values. This is especially common if you are looking at presence or absence of species in an assemblage like birds or mammals.

Slide 11

If there was total precision and accuracy, we would not need statistics. One sample or sampling unit would tell us all we need to know about the whole population. However, in ecology, we are dealing with a great deal of diversity at the genetic, species and environmental levels and we must account for this. We will want to know how the values are distributed around the central tendency and this case we’ll say the mean.

We can look at the range of values. Our lowest value in our data set is 11 and the highest is 79 so the range between these values is 68. This doesn't tell us a great deal on its own. And this is partly because we could have outliers. So most of the data could be around the 40 mark and then 79 is just on its own as an outlier. So the range in reality is not terribly helpful.

Standard deviation of the mean is a very useful way of measuring dispersion and it is the basis for various other ways of estimating variation in our data. Here are the equations for calculating the standard deviation of the mean. Note that for the population and the sample mean everything is the same in these two equations, except for the denominator of the fraction. The underneath, which is the bottom. So for the population standard deviation, it divides by N which is the number of individuals in the population. For the sample standard deviation it divides by n -1.

Slide 12

And also, again you can see we have mu which is the population mean. And then we have x bar which is the sample mean. I'm going to walk you through this calculation, because at the beginning of this course I was asked to go into the maths in a bit of detail so apologies if you're already confident with this. Okay, so I've put the equation on the top right hand side and for each step I'll highlight which step I'm talking about with a red circle.

The first thing we do is X minus X bar. So each of our values, minus the mean. I put the data in order and you can see that we go from negative values when we subtract the mean to positive. So if you added together all of these values, you would get zero. That's not very helpful for anybody. So we square these new values. So the -34.6, -32.6 etc. And for each one we square them to make them all positive. So we multiply each value by itself.

We take the squares and we add them together, which is what this Greek letter sigma is telling us to do. This is called sum of squares and it is something that is used later in statistical tests.. I add all of these values together, we get 6,934.8

Next we divide the sum of squares by the number of values, minus one. Remember how I said this was for the sample standard deviation, but not the population. So our result here is that is called the variance.

So, n minus 1 is 19. We divide our 6934.8 by 19 and we get 364.9895 and that is the variance. Then we take the square root of the variance to get the standard deviation of the mean. So the standard deviation is 19.11 if we round up.

n-1 is called the degrees of freedom. It's just like a penalty for when we have not sampled the entire population. It means briefly that if we are told that a sample of a number of observations has a certain mean and we are asked to invent numbers for the sample, we can only invent n-1 values. The final value must take the sample mean to the value we were told. So we have no further freedom of choice. This is something that occurs in our statistical tests.

Slide 13

So here are some more assumptions. Standard deviation is important because if our data do conform to this bell shaped curve or normal distribution, we can predict with some accuracy the range of our data. This is a mathematical property of this equation. So if our data are normally distributed, 68% of the values will fall within one standard deviation of the mean. 95% of the data will fall within two standard deviations of the mean and 99% of the data will fall between within three standard deviations. So for our data this means that 68% of our data will be between 26.49 and 64.71 because that is the mean plus or minus one standard deviation

Slide 14

Variance and variation are very important properties of the data used in statistics. The variance, as we saw was the second to last step in the standard deviation calculation. Or in other words, if you multiply the standard deviation against itself. So you square it, you will get the variance of your data.

Another important property of the data is the coefficient of variation. And this is very important for comparing things that are very different scales or have very different means.

So for example, we could be asking if whales have more variability in their body weight than mice. Obviously, we cannot do a different a direct comparison as whales are orders of magnitude larger than mice. If we looked at the standard deviation that would not be useful, either because the units of standard deviation of the same as the units for the mean. So that would also be orders of magnitude different between the whale and the mouse but the coefficient of variation gives a percentage. It is the ratio of the standard deviation to the mean, which immediately makes the mice and the whales comparable.

I've made up some data here, the whales mean weight is 1000 kilos, and the standard deviation is 150 kilos. If you divide the mean by the standard deviation and multiply by 100 you get 15%. Quite low variability. So whales tend to be a similar size in my experiment.

Mice have a very large coefficient of variation. You can see it is 67.8%. So we can see that mice have much more variable body weights than whales.

Slide 15

I've put together all of our findings for the dummy data set into a summary here. This is the first step in describing your data.

Slide 16

Finally, I'm just going to show you another way to present your data that can give you more information. This works with an interval or ratio scale. Box and Whisker plots are useful because you can compare various sets of data on one graph. In publishing, scientists are starting to turn away from bar charts in favour of these box and whisker plots.

The plot is based on five properties of your data. One thing to note is that we're using the median and not the mean. We have Q1 and Q3. These are the medians of the data between below the whole dataset median and above the whole dataset median.

Sorry if this is confusing. Basically you split the data into two sets at the median, and then you find the new median of each of these two sets. That is where your first and third quartiles lie.

The whiskers are the maximum and minimum values. However, some people calculate the whiskers as 1.5 times the interquartile range (IQR).

So the interquartile range is the distance between the first and the third quartiles. So you'll need to read the legend carefully to know whether the author has presented the whiskers as 1.5 times the interquartile range like we have in this graph or if it is just the maximum and minimum.

Box and whisker plot is a really nice way of doing a quick graphical examination of one or more data sets or of your experimental treatments. And you can see how it maps on approximately to the standard deviation curve the not the normal distribution curve.

Reading

Maths for Biologists