Lecture 20 transcript

Slide 1

Today we will be talking about statistical tests that are designed to answer whether you should accept your experimental hypothesis. There is a huge amount of information available about inferential statistics. I'm just going to go over some of the key concepts and some of the most important ones that hopefully we will be using in our field studies. I'm just going to go through the theory. But if you check the mathematics for biologists sheet that I have made available there is more information.

Slide 2

So our learning outcomes for this lecture: At first we should know the difference between type one and type two errors.

We should know the difference between parametric and non-parametric tests.

You should understand when you need to transform data.

And have a basic understanding of some inferential statistics.

Slide 3

When we have done our experiment, carefully designed and replicated it, and collected our data, we have done our best to make sure that our results did not occur by chance. We've controlled for as many variables as possible.

The statistical tests that we use will calculate a p-value. This is a probability value. And the p value will help determine the significance of your results in relation to the null hypothesis. The null hypothesis states that there is no relationship between your independent and dependent variable. Everything you see is due to chance. The alternate hypothesis states that there is an effect of the independent variable on the dependent.

The graph here shows the probability distribution of your dataset. On the x axis is the range of all possible results that could be generated of your dependent variable. On the y axis is the probability of observations. We can see that the highest point of the bell curve is the most likely outcome for our experiment. It has the highest probability. If we are in one of these small tails on the left or the right, these are much more unlikely observations and therefore they have probably occurred due to our experimental treatment, rather than chance.

Statisticians have chosen 95% as the statistical significance threshold. If you get a value of 0.05 or less in your statistical test for your p value, you can be 95% certain that your result did not happen due to chance. If your variable is less than 0.05 which is a one in 20 chance, this is the very extreme end of your set of possible results as we see in this figure and if you get this value, you will reject the null hypothesis, your results are statistically significant.

If you get a P value of 0.01 or less, you can be 99% sure your result is not due to chance. So there is a big difference between your treatment and your control. We call this highly statistically significant.

Slide 4

We've done our test and we have a significant result. Our p value is less than 0.05. Our treatment has worked and we can say that our independent variable causes a change in the dependent variable. But wait. We can never have 100% certainty, a p value is based on probability. And so we could make a mistake. If we make a decision based on our statistical results, there are four possible outcomes that we see in this figure.

Two of these outcomes are correct, when we accept the null hypothesis and it is true. Also, if we reject the null hypothesis and it is false. The other two outcomes are wrong and they are named a type one or type two error.

A type one error is known as a false positive. You report your findings as true, but they are actually false. So the null hypothesis is true, nothing has happened in your experiment but you report an effect.

You can reduce your chances of committing a type one error by reducing your acceptable level of p. So instead of looking for p is less than 0.05 and 95% chance, you look for p is less than 0.01- a 1% or 99% chance. An example could be if you find a positive pregnancy test but the woman is not pregnant.

A type two error is known as a false negative. You accept the null hypothesis and say that nothing has happened in your experiment when there is in fact an effect. An example would be if you tested somebody for a disease and decided they did not suffer from it based on your statistical results. When actually they did have the disease.

The consequence of a type one error mean that unnecessary actions might be taken, while type two results in no action being taken when it is needed. Type one is generally considered more serious than type two.

Slide 5

We can make a decision on which statistical test to use based on our data structure. If you have interval or ratio data, you can use this flow chart to choose the appropriate test. So all your data is in the form of numbers. This could be counts, measurements, proportions or lots of other forms. I must also make clear that the data I will be talking about when you're making this choice will be independent, not paired. By that I mean that you were sampling from different populations. Remember when I told you about the Wessex BESS experiment and I put a rain shelter and one plot and had a roof control and an unroofed control in the same plot? These would have been paired. They would be non-independent because they were all in the same plot. Whereas if you compare two different plots, these are independent. Your measures must not be influenced by other samples as far as possible.

Slide 5

If you have one experimental treatment to compare with an unmanipulated control, you would conduct a t-test. Your experimental treatment would only have one level. So it would be something like one level of drought against a control. If you had two levels of drought and your control, you would require and one way analysis of variance or ANOVA for short. We can also have situations where we want to look at two continuous variables that change together. For example, we might want to see if body length and weight of a snub nosed monkey population are related. If they do not directly impact each other, but we think they might be related, we would choose Pearson's correlation tests. If one does impact the other, we would use linear regression. I'll go into these in a little more detail in this lecture.

Slide 6

So the tests I just introduced you to are called parametric tests. A parameter is a property of the data like it's mean and standard deviation. Now, if our data set displays central tendency, like in this graph on the left, with the mean near the centre of the data set and the data spread evenly around it, we can use a parametric test. We can trust that the parameters are good indicators of the properties of the data. If this is not the case and the data has a lot of skew like we see in this graph on the right, we can see it skewed to the left. So we would have lots of zeros or very low numbers. We need to try a different approach.

Before we move to non-parametric tests, we could try a data transformation, which is where we take the natural log or square root or square of each data point. The data in the graph on the right might benefit from such a transformation. If we try that, and we don't get a convincing bell curve, we will have to analyse our data using non-parametric tests. These are less powerful and so less desirable than parametric tests because they are based on the median and not the mean of the data.

Slide 7

The first test I will talk about is the Chi square test. This is a good way of analysing frequencies of counts such as how many black beetles, we have found compared with white. We are essentially asking if the frequencies we observe are different to those we would expect by chance.

We generate a test statistic from our data based on how many degrees of freedom we have. We compare our test statistic to a table and it will tell us if our results are significant at a p is less than 0.05 or p is less than 0.01 level.

As before, we carry out the calculation shown and sum this calculation for all of our data points. So the observed frequency is what we see in this table below. This is our data. So we've had 10 individuals that fell into category 0, 7 individuals that fell into category 1, 10 individuals that fall into category 2, and so on. So we have 10 categories we have 100 data points. But you may be wondering how we calculate the expected frequency. The frequency I have displayed here is the observed, which are our collected data. We need to know the expected frequency.

I created a data set of 100 random numbers.

We have the observed frequency. But if they were drawn randomly, we would expect each integer to occur a certain number of times.

Slide 8

The expected frequency is the probability of occurrence multiplied by the sample size. The probability of a sample from a data set being in one of 10 categories is 10%. Our data set is n = 100.

10% is otherwise represented as 0.1 so 0.1 multiplied by 100 is 10. 10 values will appear in each of the categories.

We put it into the equation as below. I've only presented the first three values. So you see, we have 10 - 102 over 10 = 0. We have 7 – 102 is 9, over 10 = 0.9 and so on.

You can see from my working that our chi square test statistic is 5.2. Degrees of freedom is the number of categories, minus 1. We have 10 categories. So we have 9 degrees of freedom.

Slide 9

We check the table. We have degrees of freedom down the left side under the column name r. And for our experiment we have 9 degrees of freedom. We look at the heading and in the column marked 0.950 which is chi squared at the 0.05 level we see 16.92. Our data are no different to if they had been allocated by chance, in order to get statistical significance, our test statistic would need to be higher than 16.92 but we are much lower. So it is not significant. We have seen no effect of the categories on our data- we accept the null hypothesis. And so our treatments have not affected the dependent.

Slide 10

Now we'll talk about the Student’s t-test. Don't be fooled by the name, the t-test was invented by a man named Student. So it's not designed for students. It is designed to compare the means of two continuous data sets. The sample size for this test would be less than 30 so that's what we mean by small samples.

The equation here looks a bit scary, but it isn't really. It’s things that we've come across before. If we look at the equation, everywhere you see a number and subscript that’s hanging below the line, that just refers to which data set it is talking about. So remember we have two datasets, we're comparing two samples. And so we have sample one and sample two and that's all that that means

On the top we have the mean of dataset 1 minus the mean of dataset 2. We do everything in the brackets first then solve the square root, and then the final division. Let's look at an example.

Slide 11

Our example here is male and female warblers, and we are asking if they have different average weights. I have provided some data parameters. As you can see, we managed to catch more females than males. So we have 6 males and 8 females, a higher n.

We have the two means which is x bar. So we have x bar 1 is 74.8 and the mean of x bar of the females are 72.99. We have the standard deviation for both which is s and then we have s2, which is the variance and we're going to do our calculation.

So we'll put the numbers into the equation. Now, you might want to pause the video here and just take a good look at how the algebra maps to the parameter values where I've inserted the numbers.

So first, we solve everything inside the brackets. Remember if you have brackets and then a number. You see this 6-1 in brackets and then 1.08 you would do 5 multiplied by 1.08 you see.

Feel free to get a calculator and follow along with this just to make sure that you're happy with how this works. So solving for our equation, we can see that our final test statistic t is 2.55. As for the chi square we check the table of values to see if it is significant. Bear in mind, every statistical test will have a different statistical table. So here it's the t-test has its own table for checking these values.

Our degrees of freedom are 12. You add the n values together. So that's 6+8, and then you would subtract one degree of freedom for the male and one for the female so 14 - 2 is 12.

The tabulated value when we look is 2.179 and our value is 2.55 so we accept the alternative hypothesis.

The bird weights are different between female and male.

\slide 12

We move on to one way ANOVA or analysis of variance. In principle it is similar to the t-test, but we have three or more groups. We could be comparing the weights of three different bird species, for example. Assumptions of one way ANOVA strict, you must have normally distributed data. With 68% falling within one standard deviation of the mean If not, you could either transform the data or try a non-parametric alternative like Kruskal Wallace.

So you would start by doing your data into your histogram to check for the bell shaped curve. So with one way analysis of variance, you'll get an F-statistic. This will be compared to the value in the table and you work out degrees of freedom in the same way as the t-test. The analysis is based on the variances of the three data sets or more however many datasets you have. It's comparing variances and it's based on sum of squares.

Now, we looked at sum of squares briefly in the previous lecture, and it is the squared value of the difference of the data point from the mean. So that's what ANOVA is looking at, whether the means and the variances of the data are different.

Now, the reason we do ANOVA and not lots of tests is because in statistics, if we were to carry out many tests to get one answer you run the risk of type one errors, finding a false positive. So let's say you had a dataset A, B and C. You could not run a t-test looking at A and B, then looking at A and C, and then looking at B and C. This will lead to inflated chances of finding a false positive. And it's very bad, it's a bad thing to do. ANOVA is there for us to look at the variance around all three means at the same time, and to give a trustworthy result.

So the graph here is from some work that we did and it shows the output of a two way ANOVA. We've got drought and we've got plant species. And this is from the Wessex BESS experiment. And we were looking at how drought affects quality of flowers for pollinators. It was a really nice piece of work.

And so the factors are flower species we have *Lathyrus* *pratensis* on the left, *Onobrychis viciifolia* in the middle and *Prunella vulgaris* on the right. So these are the three plant species. And then at the bottom we have control as C, RC as roofed control and then D for drought. And in this graph, we are looking at whether there is a difference between number of flowers in the flowering head. That's called a raceme. And what we see is, well, first let's look at the n. So we have the n, the number of each sample here. So in the control we have 48 flowers and the roof control we have 46 flowers and then the drought. We have 44 flower samples.

So each of these box and whisker plots is a combination of two levels. The stars in the middle panel are a good way of describing significance. So for this species in the middle *Onobrychis viciifolia*, we can see that there is a significant difference. And because there were two stars it’s highly significant. The p is less than the 0.01 level. One star would be 0.05. And we can see that there is a significant difference between the control and the drought. There are more flowers in the control than there are in the drought. That means that the loss of water significantly impacts on the ability of the *Onobrychis viciifolia* to flower and this might have implications for the availability of nectar for the pollinating species. The other two species do not have any significant stars. For *Lathyrus pratensis* on the left, this could be (so these are the medians these means), and this could be because there really isn't an effect. For *Prunella vulgaris* that doesn't look like there's quite a big difference in medians. But there's a lot of variance. Look at the size of this whisker. So there's so much variance so much difference that we simply cannot detect a signal of the drought. It's hidden the effect of the drought. If it is present, and this is why we really want quite low variance, if possible, and why you tried to control for as many factors as possible in an experiment because it could be any number of reasons why we have so much variation.

So the final thing is you might need to do what is known as a post-hoc test. We would have done a post-hoc test here and it's just to see which treatments are different from the other treatments. What we seem to have because the line is looking at this one on this one. It seems that these two are different from one another. But the one in the middle has no difference it's somewhere in between. So there's no difference between the roof control and either control or drought. It's just the other two treatments are different from one another.

Slide 13

Pearson's correlation is looking at whether two variables covary and whether they change together. Correlation is a technique that measures the strength of a relationship between two quantitative continuous variables. The relationship could be positive, where both increase together, or negative, where as one increases the other decreases. One example could be does blood pressure change with age? And you might expect to see a positive relationship like we see here on the left.

You first have a look at the data to see if it is worth testing by plotting a scatter plot. And all three of these are examples of scatter plot. If you have a perfect fit, as we do here, you see r=1. This is a perfect fit. There was probably something wrong.

If you use statistical software, it will try to plot a line of best fit through your data. This is a line that has the lowest distance from the points possible. Your data will always have variance, which means there will always be a scatter like we see here.

This is very unlikely, and you should suspect that something has gone wrong if you get an r of either 1 or -1. That's just not going to happen with ecology data.

Slide 14

In ecology there is so much variation with everything we measure that an r value of over about 0.3 or 30% fit is seen as good. This graph here is pretty strong. The fit is based on the distance of each point from the line of best fit and then squared. Again, we see the sum of squares, which we've met previously and ANOVA is based on it. It's just a useful measure of variability. Please be aware that just because you see a strong linear relationship does not mean that one causes the other.

Also with Pearson's correlation, because we are not saying that these directly impact these, it doesn't matter which axis you plot your data on.

Slide 15

This is the golden rule of statistics- the relationships that you see between variables must be supported by sound theoretical reasons why your variables are linked. Correlation does not mean that one variable causes the patterns in the other.

**So Correlation does not imply causation**.

Slide 16

Our final statistical test is linear regression. Like correlation, we expect the relationships between our variables to be linear. That is a fit to a straight line. But in this case, one variable is caused by the other. Here we have independent and dependent variables, again, for example, temperature is the independent variable and plant growth rate is the dependent variable. Again we plot the data in a scatter graph and make a line of best fit. This line is called a regression line and it is the result of a mathematical equation. That predicts the values of the dependent variable based on the independent variable. Please note that we cannot predict the dependent variable beyond the range of the independent. Something unexpected might happen at higher values. So for example, if you were to look at enzyme activity and you only looked between zero and 30 degrees centigrade. You could not predict activity at 50 degrees centigrade, because the model would not be able to predict that the enzyme will be killed at 40 degrees. So it would probably just give you a constant increase to 50 degrees. And we know that that can't possibly be true. So you must not try to extrapolate beyond the range of the independent variable.

We are looking to predict the value of plant growth rate at a given temperature

We again get a value of the fit and it is usually presented as r squared. So you see here that the r squared is 0.7952. Now, this is about 80% which is very, very good.

R squared is bounded between zero and one and can be considered as a percentage of the fit.

Every straight line on a graph has two main parameters and this is what we're trying to calculate with our linear regression. We are calculating the intercept, which is where the line meets the y axis. So on the graph, you see that it cuts just above 0.1. And then the other parameters, b is the slope which is the steepness or gradient of the line. I've placed the correct equation on the graph for your reference.

And you can see that the intercept is 0.1169. The equation is the other way around here, but it doesn't make any difference and the slope is 0.6334. If the independent variable increases by one point, so if we move from left to right by one, then we should have gone up from the intercept by 0.6334 and you can see if you want to make a triangle. And you measured it. You see that one here is just above point seven. You would subtract this from this and you would get 0.6334.

So linear regression is a very important predictive tool, but you must use it with some caution and you cannot go beyond the range of this data. So this data goes to one, we have no way of knowing what happens at 1.2 we can suggest that it might continue up and end up at about 0.85 possibly. But like with the enzymes. We just don't know. There might be a sudden drop in activities. So we cannot make that prediction.

Slide 17

So in summary, type one errors are false positives, type two errors are false negatives.

Parametric tests are based on the mean being the central point of the data set and the data being normally distributed. Parametric tests have more statistical power than non-parametric tests. So we should use those ideally, if we can.

And the structure of your data set would dictate the appropriate test to use.

Discussion

2 minute presentations of your posters.