T tests and hypothesis testing

Duncan Golicher

11/13/2018

## Hypothesis testing

What you must remember

* The difference between a null hypothesis and the working hypothesis
* The interpretation of a p-value
* The basic assumptions underlying parametric tests

What we will look at in greater depth on the course

* The strengths and weaknesses of hypothesis testing
* A range of methods for hypothesis testing
* The relationship between test based approaches and model based approaches
* Choosing appropriate statistical analyses
* Running diagnostic tests of assumptions

## Statistical hypothesis testing

Statistical hypothesis testing works like this.

Given that the measurements that we obtain from nature tend to vary in a random, unexplainable manner it may be very easy to be fooled into thinking that patterns exist that are really just down to **chance**.

We look for patterns in our data that would indicate something interesting, but we mistrust what we have found until we have ruled out the null hypothesis that the pattern could have occurred anyway simply as a result of random variation.

### The unpaired t-test

For example, when using an unpaired t-test we are interested in the difference between the means for two groups. If you take a set of numbers with random variation and split them into two arbitrary groups the means will never be exactly the same. If you just calculate the two means and then report them in your dissertation, or paper, as representing a real difference between the groups (one that would be reproduced by another researcher following your methods) you would quite rightly be asked by any scientific referee to provide some justification that the difference you have found is not simply an artefact of the intrinsic variability in your data.

So, in order to convince the sceptics, we set up a null hypothesis. The null hypothesis expresses the position that the difference could have arisen if the measurements were taken from the same population (an effect attributable to “chance”“). We then calculate a statistic that takes into account both the size of the difference and the random variability in the measurements. We calculate how likely it would have been to have got this statistic (or one more extreme) if the null hypothesis were in fact true. We want to be able to say that it is not very likely in order to convince the sceptics. If this statistic falls in the tail of its probability distribution under the null hypothesis we know that is unlikely that we have got the data we have (or data even more extreme) if the null hypothesis were true. We therefore now are justified in suspecting that the null hypothesis is false and we can reject it. The probability of getting a statistic (or one even more extreme) is called a p-value. So in order to reject the null we need large positive or negative values for the statistic (t), and a **small p-value**. The critical p-value for rejecting the null is typically set at 0.05 (one chance in twenty of getting the result (or one more extreme) if nothing were really happening (H0 is true). There may be a lot more work to do in order to understand how and why the difference occurred and what it implies, but at least we can rule out chance.

A t-test is one form of null hypothesis test that looks to establish whether there is a real difference between the means of two groups. The null hypothesis is that there is no difference between the means (the data could have been drawn from a single population).

### The t-statistic

Our test of a null hypothesis is going to take into account all the following aspects.

* The differences between the means: Large differences are more likely to be detected easily.
* The size of the sample: Large samples are more likely to produce more reliable results.
* The intrinsic variability in the data: High variability will make it difficult to reproduce the results.

Now, we have already seen that the standard error takes into account the last two elements that we want to include in our test Size of sample and intrinsic variability). So, the test statistic can be based on the number of standard errors by which the two sample means are separated. This will take into account the first element (differences between the means) as well.

$t=\frac{DifferenceBetweenMeans}{StandardErrorOfDifference}=\frac{\overline{x\_{1}}−\overline{x\_{2}}}{SE\_{diff}}$

For two independent (i.e. non-correlated) variables, the variance of a difference is the sum of the separate variances. This important result allows us to write down the formula for the standard error of the difference between two sample means.

$SE\_{diff}=\sqrt{\frac{s\_{1}^{2}}{n\_{1}}+}\frac{s\_{2}^{2}}{n\_{2}}$

Let’s run through this step by step in R

Again, we’ll load the data again and attach it.

d<-read.csv("https://tinyurl.com/aqm-data/mussels.csv")

attach(d)

To extract the values of Lshell for two sites we use square brackets. This feature of the R language will be explained in more detail next term.

Site1<-Lshell[Site=="Site\_1"]
Site2<-Lshell[Site=="Site\_2"]

Now we can get the means as we did before.

Mean1<-mean(Site1)
Mean1

## [1] 100.7692

Mean2<-mean(Site2)
Mean2

## [1] 109.236

The lengths of the vectors give us the n values. R calculates the variances *The formula above uses the population variance as calculated with a denominator of n rather than n-1. Here we have ignored this in order to use R’s defaults and keep it as simple as possible in order to demonstrate the principle behind the test. It will make a very minor difference when we compare the result to that produced by the software.*

N1<-length(Site1)
N2<-length(Site2)
Var1<-var(Site1)
Var1

## [1] 153.6646

Var2<-var(Site2)
Var2

## [1] 196.0907

We have what we need to calculate the standard error of the difference.

SEdiff<-sqrt(Var1/N1+Var2/N2)
SEdiff

## [1] 3.708613

And we can now find the value for t

t<-(Mean1-Mean2)/SEdiff
t

## [1] -2.283002

Finally we look up the p value.

To do this we need the value of t and the degrees of freedom. The term degrees of freedom refers to the amount of independent replication in the study design. In this case the degrees of freedom are calculated by adding the two values of n. However we calculated a mean from each group. As a non technical rule of thumb we subtract a degree of freedom each time we calculate a mean. So, we have n1 plus n2 minus 2 degrees of freedom. We multiply the p value by two as we are carrying out a two tailed test. A one tailed test assumes that we are only interested in one side of the question (eg. H0 The mean for site 1 is not larger than the mean for site2). We rarely work with such hypotheses in ecology. In most practical cases we will choose a two tailed test. You should always use a two tailed test by default unless you can provide a very good reason not to.

df<-N1+N2-2
pval<-2\*pt(t,df)
pval

## [1] 0.02680621

Of course we do never need to do all these calculations step by step unless we are trying to understand the logic as here. R does it for us with one command.

t.test(Site1,Site2)

##
## Welch Two Sample t-test
##
## data: Site1 and Site2
## t = -2.283, df = 47.762, p-value = 0.02692
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -15.924396 -1.009143
## sample estimates:
## mean of x mean of y
## 100.7692 109.2360

Notice that the p-value and degrees of freedom are slightly different to the ones that we got when working through the steps. This is because R has made a correction for the fact that the two variances were different. This is known as Welch’s test.

As this is the safest test to use, R uses it by default. Welch’s correction is a good one. It makes the test more robust than the classical method we have just gone through step by step.

### Interpretation of a p-value

So how do we interpret the result?

We have a p-value of 0.0269. This is below 0.05, which is the traditional cut off point for statistical significance. So there is a less than one in twenty chance of having got this result (or one more extreme) if there really were no difference between the mean values of the two sites. In this case we can reject the null hypothesis and conclude that there is a significant difference in mean shell length(p<0.05) between the two sites.

But what about this case?

Site3<-Lshell[Site=="Site\_3"]
t.test(Site1,Site3)

##
## Welch Two Sample t-test
##
## data: Site1 and Site3
## t = -0.68235, df = 8.1748, p-value = 0.5139
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -26.36357 14.28953
## sample estimates:
## mean of x mean of y
## 100.7692 106.8063

If a p-value is larger that the conventional cut-off (0.05) we can’t reject the null hypothesis. You have to be careful here. The test does not really say that the mean values for the two sites are the same. It simply suggests that as things stand there is not enough evidence to claim that there is a real difference. If we increased the sample size in order to reduce the standard error we might still be able to show a significant difference. ``Absence of evidence is not evidence of absence’’.

Notice that the variance for Site 3 is large compared to Sites 1 and 2.

var(Site1)

## [1] 153.6646

var(Site3)

## [1] 578.936

The larger the overall variability in measurements, the harder it will be to find significant differences.

#### Testing the assumption of normality

One of the assumptions of the t-test is that the observations within each group could have been taken from an approximately normal distribution. When working with large datasets minor violations of this assumption are not necessarily important. However major violations are almost always serious when we run simple tests on relatively small data sets (n<30).

As you gain more experience in data analysis it becomes possible to detect the gravity of a violation of normality easily using histograms and QQ-plots. However visual inspection works best with large data sets. When you are working with small data sets the assumption that the data could have been drawn from a normal distribution can be treated as a null hypothesis. A test of this hypothesis can therefore be used for diagnosis. If we have to reject the null hypothesis (p<0.05) there is good evidence that one of the assumptions of the t test has not been met. We would therefore usually have to use an alternative statistical procedure. We can run tests of normality for each site using the Shapiro Wilks test.

shapiro.test(Site1)

##
## Shapiro-Wilk normality test
##
## data: Site1
## W = 0.98072, p-value = 0.8888

shapiro.test(Site2)

##
## Shapiro-Wilk normality test
##
## data: Site2
## W = 0.92194, p-value = 0.05673

shapiro.test(Site3)

##
## Shapiro-Wilk normality test
##
## data: Site3
## W = 0.90365, p-value = 0.3115

So, in all these cases the p-value is greater than 0.05. We cannot reject the null hypothesis, therefore the tests provides no evidence to suggest that the assumption of normality is not met. There are some quite important caveats to this procedure that will be explained later on in the course. Used carefully these tests help to justify the choice of statistical method.

#### Testing for differences in variances

The traditional t test made the assumption that the variances in each group are the same. The contemporary procedure that R and other software adopts by default is known as Welch’s test. This slight modification does not require variances to be the same. We can use a t-test safely even if the assumption of homogeneity of variance is not met providing Welch’s procedure is followed. Reducing the number of assumptions we make when running a test is usually a good thing to do, providing that this does not affect the power of the test nor its interpret-ability. In this case it does not.

We might still like to test whether the variances are significantly different. In R we can use a simple test based on the F ratio. SPSS uses another related procedure called the Levene’s test which is included by default in the output% *Levene’s test can be run in R using the car package, or the R Commander GUI*

var.test(Site1,Site3)

##
## F test to compare two variances
##
## data: Site1 and Site3
## F = 0.26543, num df = 25, denom df = 7, p-value = 0.01276
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.06026177 0.75587865
## sample estimates:
## ratio of variances
## 0.2654259

var.test(Site1,Site2)

##
## F test to compare two variances
##
## data: Site1 and Site2
## F = 0.78364, num df = 25, denom df = 24, p-value = 0.5487
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.3471411 1.7570958
## sample estimates:
## ratio of variances
## 0.7836404

Testing for differences in variance can be interesting in itself. There is no reason why we are only looking at differences between mean. We may also want to know if the shell lengths are more or less variable at one site or another. This may have important ecological implications. Here is appears that there is a significant difference in the variability of the shell lengths between sites 1 and 3, but not between sites 1 and 2.

### Tests and confidence intervals

If you looked at the output from tests you should have seen that in addition to the p-value from the test R also provides confidence intervals for the difference between the means. The confidence interval is almost always more useful than the p-value alone. The test for significance simply tells us that there is a difference that is not attributable simply to chance variation (if all the assumptions are met). The confidence interval shows the possible size of the difference together with a measure of our uncertainty. If a 95% confidence interval includes zero then the p-value will not be significant. If the limits of the confidence interval approach zero then the p-value will be just below 0.05. If the confidence interval is far from zero then the p-value will be vanishingly small. So the confidence interval includes information regarding statistical significance.

If sample sizes are very large we can find significant differences even when these differences are in fact very small, or even meaningless within the context of the study. In many cases we expect to find a difference, so the p-value tells us nothing we didn’t already know. However, we may not know how large the difference is going to be. So you should always report confidence intervals where possible. In fact you should do a lot more than this when analysing your data fully. Good visualisation and diagnostics are essential. We will come on to these aspects as the course progresses.

### Using hypothesis tests wisely

Hypothesis testing is a powerful tool that allows you to produce formally defensible statements regarding your data. A good dissertation must include formal tests of all the hypotheses of interest. At the same time, statistical hypothesis testing can be one of the most misunderstood aspects of statistical analysis. The difficulty seems to arise when the limited goals of a statistical hypothesis testing are equated as a direct test of the broader scientific hypothesis that is really of interest. Well designed manipulative experiments can turn a statistical hypothesis test into a test of a scientific hypothesis. Unfortunately in ecological situations we are often analysing data derived from observations and surveys rather than carefully controlled experiments. In these situations the connection between the statistical hypothesis and the scientific hypothesis is usually less direct. Answering ecological questions using the available data often require running a series of statistical hypothesis tests together with more detailed quantitative analysis that extracts effect sizes and functional forms of relationships from the data. You will learn more about this next term. However these quotations from Dytham and Zuur set the scene and reinforce the point that ecological data analysis is always challenging.

Dytham on ``The art of choosing a test’’ *It may be a surprising revelation, but choosing a statistical test is not an exact science. There is nearly always scope for considerable choice and many decisions will be made based on personal judgements, experience with similar problems or just a simple hunch. In several years of teaching statistics to biology students it is clear to me that most students don’t really care how or why the test works. They do care a great deal that they are using an appropriate test and interpreting the results properly.*(Dytham 2011)

Zuur on “Which test should I apply?}” *During the many years of working with ecologists, biologists and other environmental scientists, this is probably the question that the authors of this book hear the most often. The answer is always the same and along the lines of ‘What are your underlying questions?’, ‘What do you want to show?’. The answers to these questions provide the starting point for a detailed discussion on the ecological background and purpose of the study. This then gives the basis for deciding on the most appropriate analytical approach. Therefore, a better starting point for an ecologist is to avoid the phrase ‘test’ and think in terms of ‘analysis’. A test refers to something simple and unified that gives a clear answer in the form of a p-value: something rarely appropriate for ecological data. In practise, one has to apply a data exploration, check assumptions, validate the models, perhaps apply a series of methods, and most importantly, interpret the results in terms of the underlying ecology and the ecological questions being investigated.*(Zuur, Ieno and Smith, 2007).

Dytham is correct in stating that most students don’t care how a test works. Many useful procedures involve quite advanced mathematics. Most ecological researchers do not fully understand how complex multivariate methods “work” either, although they should always understand the concepts underlying them. It is quite wrong to assume that no statistical understanding is required to choose an analysis. In my experience it is impossible to select any analysis without having at least an intuitive notion of the underlying statistical mechanisms. Zuur’s emphasis on **analysis** rather than **testing** alone is a result of experience with large data sets and complex ecological questions.

The simple “classical”" hypothesis tests that are emphasised in Dytham’s book are most useful for answering questions with relatively small data sets. They are a good starting point for developing skills in quantitative analysis. In some cases they may be all you need. However note this warning from Dytham **“A common tendency is to force the data from your experiment into a test you are familiar with even if it is not the best method”.** In other words, if all you have is a hammer every problem looks like a nail. This will be the theme in AQM.