Research Methods 2

Week 12: Document 1

The Power of a Hypothesis Test.

Note that the discussion this week is in terms of comparing two independent groups of a variable which has a Normal distribution. The comparison is therefore made using an unpaired t-test and the null hypothesis is that the population means in the two groups are equal. This makes the narrative much more definite. However, the ideas of power and sample size can be applied much more widely.

What mistakes can you make with hypothesis tests?

You learnt in week 9 that at the conclusion of a hypothesis test you have two alternatives. You can conclude that the null hypothesis is false or, that it is true and that you have just seen data that occurs by chance with probability given by the P-value. So, if you see a small P-value, say P=0.02, you may well conclude that the observed difference is sufficiently large to discredit the null hypothesis.

So, if the null hypothesis is true there is a 2% chance that you will wrongly conclude that it is false. This is known as the Type I error rate of the test and is given by the P-value.

What happens in the other circumstance – namely when the null hypothesis *is* false but the P-value is not small, so you are not able to conclude that it is false? If the null hypothesis is false, then the chance that you *can* conclude it is false is known as the *Power* of the test. Alternatively, the chance that you cannot conclude that the null hypothesis is false is called the Type II error rate (and this is necessarily 1 minus the power). There are a couple of general points that ought to be made at this point.

- If the null hypothesis is true, it is true in only one way, namely the group means are the same. If it is false, it can be false in an infinite number of ways. The group means can differ by one unit, by two units, a millionth of a unit, etc. This asymmetrical aspect of the difference between true and false null hypotheses will need to enter into our discussion.
- The definition of power just given talks about '… being able to conclude that [the null hypothesis] is false…', so a precise discussion needs a precise definition of what would enable us to conclude that a null hypothesis is false. This is usually taken to mean $P < 0.05$. So, implicit in our definition of power is a threshold significance level that distinguishes null hypotheses we can conclude are false from those that we cannot conclude are false. Of course, if we wished to be more stringent, then we could have chosen another threshold, say $P < 0.01$. However, figures for power will change depending on the P-value we choose as our threshold.
- Notice that in the preceding bullet point we are careful to talk about null hypotheses we can conclude are false and those we cannot conclude are false. We continue to draw the important distinction between failing to draw the conclusion that a null hypothesis is false and drawing the conclusion that the null hypothesis is true. In statistics we are hardly ever in a position where we can conclude that a null hypothesis is true.

In fact, unlike the Type I error rate, the power of a test is not a single number, as it depends on several factors, which will now be discussed.

On what does the power of a test depend?

It is, perhaps, easiest to approach this issue by thinking about randomly generated data, as you have encountered in several of the exercises in previous weeks.

Suppose that you generate a sample of size 12 from a Normal population with 15 and SD 1 (you could think of this as the haemoglobin concentrations of healthy males. Then generate a sample of size 15 from the same population and perform a *t*-test to compare the two groups. Repeat this exercise many

times, say 2000 times. As the two populations are the same, then you should expect 5% of the 2000 P-values you have obtained to be less than 0.05. In other words, only 5% of the tests suggest that the populations are different, which would be expected when the null hypothesis is true.

If the exercise is repeated but now with a false null hypothesis, what happens? Suppose that the populations are now Normal with common SD of 1 but with means 15 and 13 (this is similar to the populations of haemoglobin concentration for males and females).

Changes in the difference between the population means

Of the 2000 *t*-tests with the new means, only one gave P>0.05. Thus if the difference in means is 2 units (where the SD is 1) then there is a greater than 99.9% chance of detecting that difference with samples of 12 and 15. If the difference in means had been only 1 unit, with the same sample sizes and SD, then the number of P-values less than 0.05 is 1423, that is 71% of the tests reveal a difference between the groups.

As might be expected, the *t*-test is more likely to indicate a difference between the groups (i.e. that P<0.05) if the difference between the *populations* is larger. In particular we have seen the *power*, the chance of picking up a difference when there is one, is over 99.9% when the difference in population means is 2 units. We have also seen that the power is 71% when the difference is 1 unit. So the power of a test depends on the difference in the population means.

Changes in sample size

If the data generation exercise is repeated for the case when the population means differ by just one unit, but the sample sizes are doubled to 24 and 30, then the number of tests that indicate a difference between the groups is 1889, so increase in sample size has led to an increase in power from 71% to 94%. If the sample size had dropped instead of increase, to say 6 and 8, then the number of tests out of 2000 with P<0.05 falls to 832 and the power falls to 42%.

It should not be surprising that an increase in sample size leads to an increase in power. With large samples the difference in sample means should be a better estimate of the difference in population means. Consequently if the difference in population means is not zero (i.e. the null hypothesis is false) then the difference in sample means has a better chance of indicating this if the sample is larger than if it is smaller.

It will also be shown (in the exercise sheet) that the population SD also has an effect on power. See if you can predict the direction in which this effect operates.

How can the power be controlled?

We have seen that the power of a test is not a single number but a quantity which depends on a collection of factors, namely the difference in the population means, the population SD and the sample size. When conducting a study only the sample size is under the control of the investigator – the others are population parameters which describe the distribution of the variable under study.

In fact, things might seem worse than this at first glance. Not only are the parameters not under the control of the investigator, they are not, and cannot be known to the investigator. How then can an investigator conduct a study with adequate power? The answer lies in the following points

- The investigator will certainly want to run a study with adequate power. If the study is under-powered, i.e. the power is too low, then there will be too high a chance that the study will fail to detect a real difference between the groups.

- Since the power depends on the difference in the population means, the investigator needs to decide at what difference he want to have adequate power. If there is a very small difference between the groups then very large samples would be needed to pick this up. If there is a larger difference then a more modest sample size would be needed. However, it is unlikely that a very small difference would be of practical importance, and the investigator may have no objections to having inadequate power to detect unimportant differences.

Consequently investigators must decide, before a study starts, what size of difference between the population means constitutes an important difference and design the study to have adequate power to detect a difference of this size. This will usually entail setting the sample size to make sure it is sufficiently large to detect such a difference. Formulae are available to do this. These will not be discussed here – they can be found in Bland, chapter 18. For the case of a variable with a Normal distribution the investigator will have to specify four things before a sample size can be calculated.

i) The smallest difference in means that it is important the study be able to detect. This is a quantity that is assessed on the basis of clinical judgment.
ii) Some estimate of the population SD. This will often have to be gleaned from the literature of related studies, or from a more limited pilot study.
iii) The significance level at which the null hypothesis will be discredited, usually 0.05 or 0.01.
iv) The power to be achieved at the difference specified in i).

It is of the utmost importance that studies have adequate power. If a study results in a P-value of, say, 0.45, then we cannot conclude that there is a difference between the groups. However, nor can we conclude that there is no difference between the groups. If the study had adequate power then it is unlikely that there will be an *important* difference between the groups. However, if the study has poor power, then there might be quite a good chance that an important difference exists but the study has failed to pick it up.

How do I assess the power of a study that has been completed?

In general the answer is that you don't. We have seen that power depends on, among other things, the difference in population means. This difference can never be known as it depends on unknown parameters. When a study is being designed a sample size is determined by selecting a specific difference in these parameters which represents something of a clinical importance. This difference is not based on data but on a judgment of what is important. At no stage should we rely on a sample estimate of the parameters to conduct sample size/power calculations.

Before a study is conducted all but one of the determinants of power are fixed - only the sample size can be changed. Once a study has been completed, all the determinants of sample size are fixed. There is, therefore, no purpose in attempting a power calculation at this stage. The temptation to do so often stems from the need to know if a non-significant P-value has arisen because the difference in means is genuinely unimportant, or because the study did not have the power to detect an important difference. This is a round about way to find out what differences there might be between the groups. The correct approach at this stage is to compute a confidence interval, and see whether or not the interval includes a clinically important difference.