Research Methods 2

Week 9: Document 1

Asking questions – an introduction to hypothesis tests

An introductory example

Takeuchi *et al*. (*Int. J. Oncology*, 2002, **20**, 53-58) report on the usefulness of SPECT scanning for predicting the response to chemotherapy of patients with breast cancer. They assessed 25 patients using both technetium and thallium. The retention of the isotope is measured by the retention index, which is specified as a percentage. Of the 25 patients, 11 responded and 14 did not.

For thallium the mean retention indices were:

	Mean retention index
Responders (n=11)	47.5
Non-responders (n=14)	55.8

Does this mean that those responding to chemotherapy have a different mean retention index for thallium?

Before answering it is necessary to consider more carefully what we mean by the question. Naively it is clear that the mean retention index does differ between these two particular groups, 47.5 is different from 55.8. However, even superficially this answer unsatisfying. The reason for this is that if the answer to our question is to be useful, it needs to go deeper than just these 25 patients. In some way we want to know if *all* patients who respond to chemotherapy will tend to have lower values for the retention index when compared with *all* patients who fail to respond.

This puts the problem into a statistical framework. We want to know what we can say about all patients of a particular type, *the population*, based on a subset of patients drawn from that group, *the sample*. We are therefore looking at a problem of statistical inference that is related to those we have considered over the last few weeks.

A more precise version of the above question is to ask if the distribution of the thallium retention index in the *population* of patients who fail to respond differs from that in the *population* who do respond. In particular it follows from this that the means of the populations would coincide[†]. Symbolically, the mean of the non-responders is written m_{NR} and that of the responders is m_R. The question is: is $m_{NR} = m_R$? This demonstrates the nature of the difficulty we face. We want to know about the equality of two quantities, neither of which we can ever know precisely.

If we are to make progress with this problem then it is natural to start from the information about the population means which is available in the sample means. If the two population means are the same, then the sample of responders and the sample of non-responders constitute two samples drawn from the same population. We know from Week 6 that the means of two samples from the same population will differ from one another. So if all we have are two sample means that are different, the challenge is to try to decide which of the following obtains.

- Is this difference just what can be expected when two samples are drawn from the same population, i.e. the difference *is due to chance*

[†] There is clearly more to assessing the equivalence of two distributions than simply asking if the means are the same. However, the equality of means is a very important assessment and is the one which will preoccupy us in this course.

or

- Is the difference telling us that there is difference between the underlying *population means*?

Sorting out which of these obtains is the challenge we face. The solution is to perform what is often called a *significance test* but which is better called a *hypothesis test*.

The nature of the answer

For those who have not met it before, the nature of the answer provided by a hypothesis test is rather surprising and it is probably worth outlining it now, before the details of the later explanations have a opportunity to cloud the key issues.

The key problem is that virtually any size difference between sample means *could* arise by chance. As such we can never look at a difference between two sample means and say "yes, that difference proves that the population means are different". Indeed, nor can you say "this difference proves that the population means are the same".

What we can do, to start with at least, is to note that if the samples do come from populations with the same mean (effectively they come from the same population[†]) then:

1. The difference in *population* means is zero
2. The difference in *sample* means will vary about zero
3. Large departures of the difference in the sample means from zero is unlikely.

If we see a **big difference** in samples means, then it is **unlikely** that this difference has arisen by chance and the observed difference provides us with evidence that the population means are different.

What constitutes a **big difference** needs to be pinned down more precisely. Also, we need to provide a quantification of what we mean by **unlikely** and this is what hypothesis tests do. The result is the ubiquitous *P-value* , many of which now decorate virtually every medical research paper.

We will now expand on some of these issues. Further reading can also be found in chapter 9 of Bland.

Definition of a P-value

This section will introduce the idea of a P-value and calculate one. However, you should bear in mind that this will be done in a way that is intended to illustrate what a P-value actually means. It is not (and could not be) the way that a P-value would be computed in practice.

In the paper by Takeuchi *et al*. mentioned above, the mean retention index for thallium was 47.5% among the 11 responders and 55.8% among the 14 non-responders..

Suppose that the retention index has a Normal distribution and that the *population* mean is the same among the responders and non-responders. We could use Minitab to generate a sample of size 11 from this population and compute its mean. We could then generate another sample, this time of size 14, compute its mean and then calculate the difference between the sample means. In this was we have mimicked what happened in the study. We could then repeat this process until we have 1000 differences between sample means and then look at the distribution of these differences.

In order to do this we need values for the population mean and SD. Of course these are unknown and this is one reason why the approach cannot be adopted in practice. However, for our present illustrative purposes we can make do with plausible values: the paper by Takeuchi *et al*. suggests that a

[†] For this course, we ignore the niceties of populations with the same mean and different spreads.

plausible mean is 50% and 50% is also a plausible SD[‡]. Below is a histogram of the 1000 differences in the sample means.

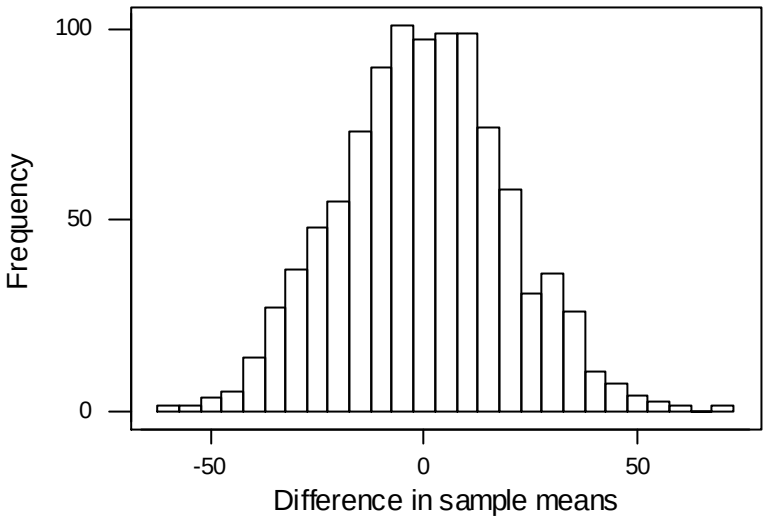

Figure 1: distribution of differences in sample means

This shows that *if the two samples come from the same population* then the observed differences in sample means are distributed as above. As you would expect, with both samples coming from populations with the same mean, the difference in sample means varies about 0. However, this histogram tells us that the amount of that variation is substantial. Some of the differences are above 50 or below –50. Of the 1000 mean differences that have been generated, 208 are outside the range –25 to 25, and 616 are outside the range –10 to 10.

The difference that Takeuchi *et al*. actually observed was 55.8 – 47.5 = 8.3. Such a value is clearly entirely typical of the difference that might arise *even if the population means* were the same for responders and non-responders. Of course, we still do not know if the population means *are* the same in this study but it is clear that a perfectly tenable explanation of the difference we have observed is that it is due to chance.

Can we be more quantitative? The answer is yes. We can compute how many of the 1000 mean differences we generated are further from 0 than the observed value. Doing this, we find that 676 values that we know have arisen from populations that have the same mean are further from zero than the observed difference. Therefore, *if the populations of responders and non-responders share the same mean* there is a chance of 676/1000 = 0.676 of seeing a difference at least as great as that we have observed. The value 0.676 is the *P-value*. As the P value measures a probability it is always a number between 0 and 1.

The P-value is a measure of how surprised we are to observe a particular difference, *if the population means are the same.* If the value is large, as it is above, then chance is clearly a plausible explanation of the difference. There is therefore no evidence to discredit the assumption of equal population means.

If, on the other hand, the P-value was small, say 0.02, then we are saying that there would be little chance of seeing the difference we did see, *if the population means are the same.* We are then faced with accepting that we have seen an unusual event, or we can believe that the observation casts doubt

[‡] the retention index can be negative, so this equality of mean and SD does not cast doubt on the assumption of Normality

on the assumption of equal population means. In other words, if the P-value is small, the data provide evidence that the population means are different.

Hypothesis tests can be applied in many circumstances, not just for comparing population means. However, in each application they start from an assumption similar to the assumption of equal means. This assumption is given the name *the null hypothesis*. It is essentially the assumption that the observation you have is due to chance and that there is no genuine difference, hence the use of *null*. The aim of the test is to assess whether the data provide evidence that discredits the null hypothesis or not.

Effect of sample size

This will be explored in the example sheet and will also be an important theme in Week 12. The main point to be made here is that the amount of variation seen in figure 1 depends on several items. First, the amount of variation depends on the variability of the underlying observations: so if the population SD were 20% rather than 50% the histogram would be less dispersed. Second, the variability of a sample mean varies inversely with the sample size. So, for example, if the populations were as described above but the sample were from 110 and 140 patients rather than 10 and 14, the histogram would again be less dispersed. If the observed difference was still 8.3 then you would expect there to be a smaller proportion of the 1000 generated samples which were more than 8.3 from zero. This would lead to a smaller P-value. A given observed difference may provide evidence against a null hypothesis, if the sample sizes have one set of values, but no such evidence if the difference is based on smaller samples. This is essentially because results from smaller samples would naturally be expected to be more variable.

In practice, the procedure used for calculating P-values takes the role of sample size into account.

Some terminology

The P-value provides evidence against the null hypothesis if P is small. This immediately begs the question of how small is small. There are no hard and fast rules here but there are some widely recognised conventions. The following points are pertinent

1. If $P < 0.05$ but $P>0.01$, then we say that there is evidence against the null hypothesis, or, alternatively, the test is significant at the 0.05 or 5% level.
2. If $P < 0.01$, then we say there is strong evidence against the null hypothesis, or that the result is significant at the 0.01 or 1% level.
3. Occasionally we might find $P < 0.001$, when we would say there was very strong evidence against the null hypothesis, or that the result is significant at the 0.1% level.

Different writers might use slightly different words, but the idea of strengthening evidence against the null hypothesis as P gets smaller is universal.

If $P > 0.05$ then the result is sometimes said to be *not significant*, or that there is no evidence against the null hypothesis. Whatever the value of P it is important that the exact value is reported in papers and articles etc. There is a tendency to see 'NS' in journal, indicating that the result was 'not significant'. However, the use of conventional cut-points, such as 5%, 1% and 0.1% should only be taken as a guide. It is clearly important to know if a 'NS' difference resulted from P=0.06 or P=0.91.

An important subtlety

If we obtain a small value for P, say $P = 0.03$ or 0.002, then we can claim that there is only a small probability that the observed difference is due to chance, and therefore there is evidence that the null hypothesis is false.

If we obtain a larger value for P, one that might be deemed 'non-significant', we cannot conclude that there is evidence that the null hypothesis is false. What can we conclude? It is very important to realise that the appropriate conclusion is:

'There is no evidence that the null hypothesis is false'

It is *not*

'There is evidence that the null hypothesis is true'

What our test has done is detect whether a difference could be due to chance. This is not the same as saying that it *is* due to chance.

Many studies are flawed by assertions that there are no differences between groups, when what is meant is that the investigator has not found evidence of a difference. It may be that the investigator has not looked hard enough. A common example of a reason why no evidence of a difference is found is that the sample sizes are too small. There is more on this issue in Week 12.

Summary

The key features of a hypothesis or significance test are the following

i) There is a null hypothesis. In this document this has been that the population means of the responders and non-responders are the same. The null hypothesis amounts to the assumption that the difference you have observed is due to chance.

ii) The test produces a P-value, a number between 0 and 1, which is the chance of obtaining a difference at least as large as that observed *if the null hypothesis is true*

iii) The interpretation is that the smaller the P value the stronger is the evidence against the null hypothesis. Conventionally, values less than 0.05 are often taken to discredit the null hypothesis.