Research Methods 2

Week 11: Document 1

Binary data, including the χ^2 test.

What are binary data?

So far in this course all the variables have been continuous. That is they can, in principle, take any value within their range. Examples are height, blood pressure, sodium concentration.

However, in many areas of medicine variables are not of this form. Sometimes variables can only take one of several discrete values, such as tumour stage, performance status, eye colour. Some of these discrete, or *categorical*, variables may be analysed by methods for continuous variables, if the categories represent some form of order and there are many categories. Many categorical variables are not of this form and special statistical methods are needed. We will not consider the general case but will concentrate on an important extreme case, namely when the variable can take one of two values. Such variables are known as binary or dichotomous variables. An example would be whether or not a patient responded to chemotherapy, or whether or not a tumour has spread to the liver or whether or not the subject's job exposed them to asbestos.

Occasionally it is sensible to derive a binary variable from a continuous one. For example, whether or not a patient survived more than five years from diagnosis is a binary variable, albeit defined in terms of a continuous variable, namely length of survival. While there may be a reason for this manoeuvre in a given application, information is being lost by doing this and in general it is not something which should be done lightly.

What are the parameters for binary data?

An example of a binary variable is whether or not a patient with breast cancer responds to chemotherapy. This variable defined the groups in the example used last week to illustrate the unpaired *t*-test, but may well be of interest in its own right. Of the 25 patients in that study, 14 had one value of this variable, namely 'did not respond' and 11 had the other value, namely 'did respond'.

The logic of statistical analyses for continuous data was clarified by making the distinction between a population, with its unknowable parameters, and a sample of data drawn from it, with its sample statistics. A Normal population is defined in terms of the population mean and SD. What are the parameters for a population of a binary variable?

Consider a population of patients with breast cancer who have been treated with a given chemotherapy protocol. Each patient is classified as a responder or a non-responder. The only parameter available is p[†], the proportion of patients in the population who respond.

Summarising binary data

If a sample of n patients is drawn from the population, then r of them will be responders and n-r will be non-responders. The *sample* proportion of responders is $p = r/n$ and is the best estimate of p; often this is multiplied by 100 so as to be expressed as a percentage.

So, for example, in the sample of 25 patients in the paper by Takeuchi *et al*. 11 are responders, so r=11 and p = 11/25 = 0.44 = 44%. So our estimate of the proportion of responders in the population is 0.44.

In any analysis of binary data, the proportion or percentage is a useful summary to present.

[†] In keeping with the convention on symbols for parameters p is used because it is the Greek letter corresponding to p, the first letter of proportion. It has nothing to do with the 3.14159…, the ratio of the circumference to the diameter of a circle.

Just as with a Normal population it is reasonable to ask how good is this as an estimate of p? In other words, what is the standard error of p? The answer is both similar to that for a Normal variable and also different. The answer is given by the formula in the table below

Binary variable	Normal variable
Sample proportion p	Sample mean m
Standard error $\sqrt{\frac{p(1-p)}{n}}$	Standard error $\frac{s}{\sqrt{n}}$

Similarity

Both standard errors get smaller as the sample size, n gets larger.

Difference

The SE for the Normal case involves the parameter s which measures the spread of the population. There is no separate parameter measuring the spread of the binary variable. Instead the spread is related to the quantity $p(1-p)$, which is determined by the population proportion p.

A full derivation of the formula $\sqrt{[p(1-p)/n]}$ is not needed but the following brief explanation may help to make this expression seem more plausible. Suppose that $p = 0$, i.e. the proportion of responders in the whole population is zero, i.e. there are no responders in the population. In this case no sample can contain any responders, so $r = 0$ and there is no error in this as an estimate of p, so the SE is zero. This is achieved by the formula, as $p(1-p)$ is 0 when $p=0$. Similarly, if $p=1$, then all member of the population are responders, so all n members of all samples must be responders, so $p = 1$, which again is an estimate of p that has no error. Again this is achieved by the formula, as $p(1-p)$ is 0 when $p=1$.

Comparing two groups: the c^2 test

Just as there may be interest in testing whether the population means of two groups of Normally distributed data are the same, there may also be interest in testing the equality of population proportions in two groups.

Although we have not discussed it in detail, the mechanics of a *t*-test depend intimately on estimation of s. As this parameter does not even feature in the formulation of binary data it is clear that a different form of test is needed. There are several ways of constructing a test of the equality of two proportions. Methods building on the difference between the sample proportion and using the formula for the SE of a proportion are discussed in Bland, chapter 9, section 8. We will use an equivalent alternative, the c^2 test (also covered in Bland, chapter 13, sections 1 and 2). We do so because this test is widely used and is frequently encountered in the literature. The test compares the proportions in two independent samples and so is the binary variable analogue of the unpaired *t*-test. The analogue of the paired *t*-test is known as McNemar's test; this is not covered in the present course but details can be found in Bland in chapter 13, section 9.

Note: c is the Greek letter 'chi', pronounced 'kye', and the test is occasionally written as the chi-squared test

An Example

A study of pain control amongst patients with non-malignant terminal disease admitted to a hospice was reported by Zeppetella *et al.*[†]. Part of this study reported that 11 of the 27 patients who had breakthrough pain were satisfied with their pain control. On the other hand, 12 of the 16 patients who did not have breakthrough pain were satisfied with their pain control.

[†] Zeppetella, G., O'Doherty, C.A., and Collins, S. (2001) Prevalence and characteristics of breakthrough pain in patients with non-malignant terminal disease admitted to a hospice, *Palliative Medicine*, 15, 243-6.

The estimate of the population proportion of patients with breakthrough pain who are satisfied with their control is, therefore, 11/27 = 0.41 = 41%. For the patients without breakthrough pain the proportion is 12/16 = 0.75 = 75%. While these sample values are clearly different, they are based on small samples. Do they provide evidence that the two population proportions differ?

A way to answer this question is to apply the c^2 test.

The first step is to form a 2×2 table, as follows:

	Breakthrough pain	No breakthrough pain
Satisfied with pain control	11	12
Not satisfied with pain control	16	4

Two things should be noted about the way the table has been formed.

i) Each patient in the study is counted in precisely one cell of the table. The rows are classified by 'satisfied' and 'not satisfied', rather than, for example, 'satisfied' and 'total'. The latter would be wrong because each satisfied patient would be counted in two cells.
ii) The entries are counts, *not* percentages.

The above table is what is known as the table of observed values. The next step in the c^2 test is to form what is known as the table of expected values. The idea is to work out how we would 'expect' the above table to look *if the two population proportions are equal*, i.e. if the null hypothesis is true. This is done as follows.

If the null hypothesis is true, then the proportion of patients satisfied with their pain control will be the same in the patients with breakthrough pain and in those without breakthrough pain. Therefore the data from the two groups could be amalgamated and the best estimate of the proportion with satisfactory pain control is (12+11)/(27+16) = 23/43.

If this is the case, then in the group of size 27, the patients with breakthrough pain, the number we would expect to be satisfied with their pain control would be $27 \times (23/43) = 14.4$. The expected number with satisfactory pain relief in the group of 16 patients without breakthrough pain would be $16 \times (23/43) = 8.6$. In the two groups the expected numbers who do not have satisfactory pain control are, consequently, $27 - 14.4 = 12.6$ and $16 - 8.6 = 7.4$. Thus the table of expected values is

	Breakthrough pain	No breakthrough pain
Satisfied with pain control	14.4	8.6
Not satisfied with pain control	12.6	7.4

The idea of the c^2 test is that if the null hypothesis is true, then the observed table will be 'close to' the expected table. The test statistic is a measure of how different are the observed and expected tables. It cannot be negative and is zero only if the two tables are identical. Therefore there will be a range of values for this statistic even when the null hypothesis is true. However, larger values will be less likely and above a certain limit they will occur but rarely if the null hypothesis is true. This is the all we need to construct a hypothesis test.

To perform such a test in Minitab, enter the observed table into two columns, so the first column contains 11 and 16 and the second contains 12 and 4. Then click on **Stat** -> **Tables** -> **Chi-Square Test**… and select the columns you have just entered into the **Columns containing the table:** box and then click on **OK**. The Session window then contains the following:

Chi-Square Test: C1, C2

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

          C1     C2  Total
    1     11     12     23
       14.44   8.56
       0.820  1.384

    2     16      4     20
       12.56   7.44
       0.943  1.592

Total     27     16     43

Chi-Sq = 4.740, DF = 1, P-Value = 0.029

MTB >
```

Each entry in the observed table is reproduced and the corresponding entries from the expected table are also given[‡]. The χ^2 value (here 4.74) is also printed. So too is the P-value, 0.029.

The P-value is interpreted in the same way as for the *t*-test. If the null hypothesis (the population proportions satisfied with their pain relief are the same in the two groups of patients) is true, then there is a chance of only 0.029 of obtaining a χ^2 value as large as 4.74. As this is small, it is reasonable to conclude that these data provide evidence that the proportion of satisfied patients is smaller among those with breakthrough pain.

[‡] The third item in each cell is the contribution to χ^2: the sum of these items is the χ^2 statistic. We will not attempt to interpret the contributions of each cell to this sum.