IAPT: Analyses for qualitative data

Qualitative data

Qualitative data are also called nominal or categorical, and happen when we classify subjects into two or more categories. For example, we might classify a patient's condition as 'poor', 'fair', 'good' or 'excellent', or give as options for a question 'yes', 'no', or 'don't know'. This is different from quantitative data, where we have numbers which represent the magnitude of something, such as blood pressure. Even though these may actually be recorded as numerical codes 1, 2, 3, or 4, the number does have any numerical meaning. We could code 'yes' as 1 and 'no' as 2, or 'yes' as 2 and 'no' as 1 and it would not make any difference to the analysis. Categorical variables with only two categories, such as 'alive' or 'dead', or 'female' or 'male' are called **dichotomous**, **attribute**, **quantal**, or **binary**.

Many statistical methods have been developed to analyse such data. In this lecture I shall cover the chi-squared test for association, Fisher's exact test, the chi-squared test for trend, risk ratio, relative risk, or rate ratio, odds ratio, and the number needed to treat.

Contingency tables

A **contingency table** is a cross-tabulation of two categorical variables. For example, Table 1 shows data from a study of the acceptance of the HIV antibody test in antenatal clinics (Meadows *et al.*, 1994). The data are arranged in rows and columns. The part of a table defined by a particular row and column is called a **cell** of the table. The numbers in a contingency table are frequencies. The number in the first cell of Table 1, which is 71, tells us that in this study there were 71 women who both were married and accepted the HIV test. I have also added the totals for each row and column and for the whole table. The row and column totals are also called marginal totals, the total number of all the observations in the table is called the grand total. This kind of cross tabulation of frequencies is also called a **cross classification table**. We often refer to tables using the size of the table. Table 1 might be called a '4 by 2' or '4×2' table, because it has four rows and two columns. You sometimes hear the general term '$r \times c$ table', where r would denote the number of rows and c the number of columns.

The Chi-squared test

We often want to test the null hypothesis that there is no relationship between the two variables. We use the term '**association**' for a relationship between two categorical variables. If the sample is large, we can do this by a chi-squared test. If the sample is small, we must use a different test, Fisher's exact test, described below.

Our null hypothesis is that there is no association between the two variables. The alternative hypothesis is that there is an association of some type. The chi-squared test works by calculating the frequencies we would expect to see in the cells if there were absolutely no association. It works like this. For the HIV test data, the proportion who accepted the test is 134/788. Out of 486 married women, we would expect $486 \times 134/788 = 82.6$ to accept the test if the null hypothesis of not association were true. Similarly, the proportion who refused the test is = 654/788. Out of 899 486 married women, we would expect $486 \times 654/788 = 403.4$ to accept the test if the null hypothesis were true. Note that $82.6 + 403.4 = 788$. The expected frequencies sum to the same total as do the observed frequencies.

In the same way. out of 222 cohabiters we would expect 222 × 134/788 = 37.8 to accept the HIV test if the null hypothesis were true. We would also expect 222 × 654/788 = 184.2 to be refusers if the null hypothesis were true. Note that 37.8 + 184.2 = 222, the second row total. We continue in this way until we have expected frequencies for all the cells of the table (Table 2). Note that 82.6 + 37.8 + 8.5 + 5.1 = 134.0 and 403.4 + 184.2 + 41.5 + 24.9 = 654.0. The observed and expected frequencies have the same row and column totals.

We can see that for each cell of the table, we calculated the expected frequency by

$$\frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

These would be the expected frequencies if the null hypothesis were true. Here the word 'expected' is used to mean 'the average number of observations we would expect in the cell if we repeated this study over and over again'. It does not mean that we expect to see 82.6 married women accepting HIV testing. Statisticians often torture the language in this way, we regret to say. To be really pedantic, is the expected frequency if the null hypothesis were true and the row and column totals remained the same.

The chi-squared test for a contingency table uses the differences between the observed and expected frequencies. The bigger these differences are, the more evidence we will have that the two variables are associated. We cannot just add these differences, because they always sum to zero. Instead we do what we did when calculating standard deviation, we square them. Another problem is that the bigger the frequencies are, the greater is the possible size of the difference between observed and expected. We might expect that big samples would produce bigger differences than small samples, just by chance. It turns out that we can allow for this by dividing the squared difference by the expected frequency, to give:

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

The precise reasons for this choice are rather abstract and mathematical, so I won't go into them here, but Bland (2000) explains it. We work this (observed – expected)^2/expected out for each cell of the table and add them together. I don't usually give formulae in this course, but we shall come across this one quite often and it is pretty simple, so I have included it to make it easier to see what is going on. For Table 2, this sum is 9.15. This will be our test statistic (Week 3). Following the usual formulation of a significance test, this should follow some known distribution if the null hypothesis is true. It follows one called the Chi-squared distribution. (See Bland, 2000, for more about this and why it is true.) 'Chi-squared' is often written χ^2, where 'χ' is the Greek letter 'chi', pronounced 'ki' as in 'kite'. The sum of (observed – expected)^2/expected is called the chi-squared statistic, sometimes written as X^2. I shall give the Chi-squared distribution a capital 'C' to distinguish it from chi-squared statistics.

Table 1. Acceptance of HIV test grouped by marital status (Meadows et al., 1994)

Marital status	Acceptance of HIV test		Total
	Accepted	Rejected	
Married	71	415	486
Living with partner	41	181	222
Single	15	35	50
Divorced/widowed/separated	7	23	30
Total	134	654	788

Table 2. Acceptance of HIV test grouped by marital status, with expected frequencies

Marital status	Acceptance of HIV test				Total
	Accepted		Rejected		
	observed	expected	observed	expected	observed
Married	71	82.6	415	403.4	486
Living with partner	41	37.8	181	184.2	222
Single	15	8.5	35	41.5	50
Divorced/widowed/separated	7	5.1	23	24.9	30
Total	134		654		788

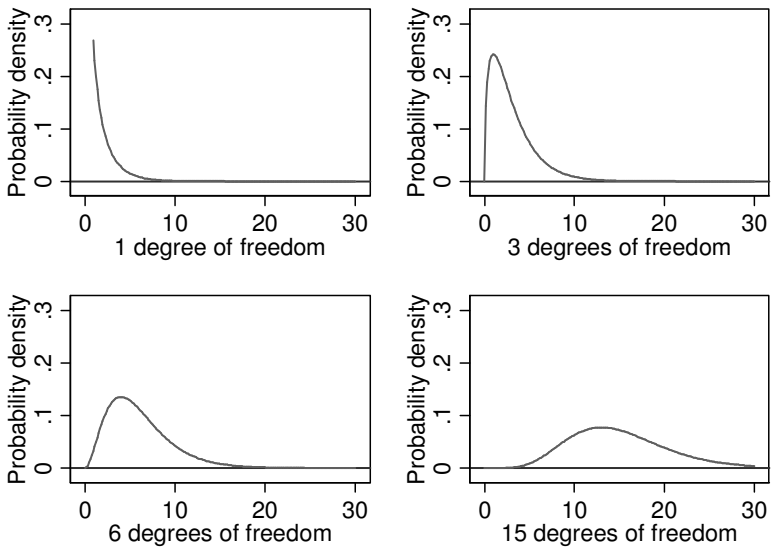

Figure 1. Some members of the Chi-squared distribution family.

Table 3. Percentage points of the Chi-squared distribution.

Degrees of freedom	Probability that the tabulated value is exceeded			
	10%	5%	1%	0.1%
1	2.71	3.84	6.63	10.83
2	4.61	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.09	20.52
6	10.64	12.59	16.81	22.46
7	12.02	14.07	18.48	24.32
8	13.36	15.51	20.09	26.13
9	14.68	16.92	21.67	27.88
10	15.99	18.31	23.21	29.59
11	17.28	19.68	24.73	31.26
12	18.55	21.03	26.22	32.91
13	19.81	22.36	27.69	34.53
14	21.06	23.68	29.14	36.12
15	22.31	25.00	30.58	37.70
16	23.54	26.30	32.00	39.25
17	24.77	27.59	33.41	40.79
18	25.99	28.87	34.81	42.31
19	27.20	30.14	36.19	43.82
20	28.41	31.41	37.57	45.32

The table shows the upper 5% or 0.05 point, as shown in Figure 2.

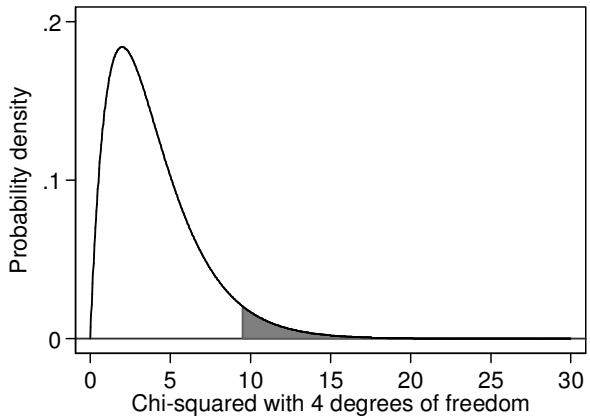

Figure 2. Upper 5% point of the Chi-squared distribution with 4 degrees of freedom, as shown in Table 2.

The Chi-squared distribution is very like the t distribution, to which it is closely related. It is a family of distributions, and the particular member of the family is defined by one parameter, called the degrees of freedom. Figure 1 shows a few members of the Chi-squared family. When the degrees of freedom is small it is highly skewed to the right, and as the degrees of freedom increases it becomes more symmetrical. Eventually it becomes like a Normal distribution. We would expect this to happen, because it is obtained by adding lots of things together and this tends to generate a Normal distribution as the number of things added increases. Like the t distribution and the Normal distribution, there is no simple formula for the area under the chi-squared curve and hence for the probability of exceeding any given value. We can use a table of probabilities laboriously calculated by a (very accurate) mathematical approximation. Table 3 shows some percentage points for the Chi-squared distribution with different degrees of freedom. Of course, computers are very good for laborious calculations and in practice we just let the computer program do the work and calculate the probability for us every time. One useful feature of the Chi-squared distribution is that its mean is equal to its degrees of freedom. So if the observed value of a chi-squared statistic is similar to or less than its degrees of freedom, the data will be consistent with the null hypothesis being tested.

We decide which member of the Chi-squared family applies to our table by calculating the degrees of freedom. For a contingency table, the degrees of freedom are given by:

(number of rows – 1) × (number of columns – 1).

Again, I won't go into the reasons for this, see Bland (2000) if you are curious. But we can see something of the logic by looking at how many cells of the table can be filled before the remaining cells can be calculated from the row and column totals. We can start filling the cells of the first row, until we reach the last cell. This must be fixed, because all the cell frequencies must sum to the row total. So the number of free choices in the first row is the number of columns minus one. We can fill in cells in the second row in the same way, getting another 'number of columns minus one' free choices. We go on doing this until we reach the last row. Now these frequencies will al be fixed, because the frequencies in each column must sum to the column totals. For all the rows but one, we have 'number of columns minus one' free choices. Hence the total number of free choices is 'number of columns minus one' multiplied by 'number of rows minus one'. This gives us the degrees of freedom for the table.

For Table 2, we have $(4 – 1) \times (2 – 1) = 3$ degrees of freedom. If we look in Table 3 at the 3 d.f. row, we see that the 5% point is 7.81 and the 1% point is 11.34. Our observed chi-squared statistic, 9.15, is between these two, so our probability is between 5% and 1%. We would write this as $P<5\%$ or $P<0.05$. Using a computer to do the calculation, we get the more accurate $P = 0.027$, which we could round to one significant figure and report as $P = 0.03$.

Like most significance tests, there are some assumptions we have to make about the data or conditions that the data must fulfil for the test to be valid. These are that the sample is large enough and that the observations are independent. The conventional criterion for the sample size is this: the chi-squared test is valid if at least 80% of the expected frequencies exceed 5 and all the expected frequencies exceed 1. This is a large sample test. The smaller the expected values become, the more dubious will be the test. For Table 2, they all exceed 5.0, so there is no problem. As there are 8

expected frequencies, we could accept $8 \times 0.2 = 1.6$ expected frequencies less than five. We should round this down to 1.0 and say that one expected frequency between one and five would not be a problem. If we have a 2 by 2 table, 20% of the cells is $4 \times 0.20 = 0.80$, which is less than one, so no cell should have an expected frequency less than 5. Note that the observed frequencies could be zero without affecting the validity of the test. As long as the expected frequencies are greater than 5.0, the test is valid. What should we do if this condition is not met? We can make the expected frequencies bigger by combining rows or columns. In Table 2, for example, we might combined the single and divorced/widowed/separated categories. Another approach is to drop a category altogether, if there is no obvious combination including it. It is easy to see how combining categories will affect expected frequencies, because when we do it the expected frequencies will simply be added together. The alternative approach is to use Fisher's exact test, described below, or, for 2 by 2 table, Yates' continuity correction, which we shall omit. The chi-squared test for association in a contingency table is also known as the Pearson chi-squared test.

The chi-squared statistic is not an index of the strength of the association. If we double the frequencies, this will double chi-squared, but the strength of the association is unchanged. There are indices of strength for use in special circumstances, but they are not seen much.

Fisher's exact test

When the chi-squared test is not valid because the expected frequencies are too small, we can use a different test, Fisher's exact test, also called the Fisher-Irwin exact test. This works for any sample size, though it used to be used only for small samples in 2 by 2 tables, because of computing problems. Now that we have powerful computers with efficient statistical programs it can be done for any table.

For Fisher's exact test, we calculate the probability of every possible table with the given row and column totals. We then sum the probabilities for all the tables as or less probable than the observed. You can do this by hand for a 2 by 2 table with small frequencies (Bland 2000 gives a formula) but you would not want to. For Table 1, Fisher's exact test gives P = 0.029. Compare this to the chi-squared test P = 0.027. They are very similar. This is not always the case. Table 4 shows the results of a clinical trial. Fisher's exact test for this table gives P = 0.004. The chi-squared test gives chi-squared = 8.87, 1 d.f., P = 0.0029. These are rather different, though both would lead to the same conclusion. Which should we choose? I think that Fisher's exact test is always to be preferred, because it is exact. So why have two tests? Fisher's can only be done for tables with more than two rows or columns or with a large number of subjects if you have a modern computer. Researchers' behaviour changes slowly, textbooks are always out of date, and it takes a long time for practice to change. You will see lots of chi-squared tests. There are many other situations where chi-squared statistics are used and the principles will be much the same. I consider one below under the chi-squared test for trend.

Not all statisticians would agree that Fisher's exact test is right and this is a matter of occasional lively debate among statisticians (yes, it really happens). I might be wrong. You will hear Fisher's exact test described as 'conservative' because it gives larger P values than does the chi-squared test. I think that the opposite is true, that the

Table 4. Leg ulcer wound healing by type of bandage in a randomised trial, with row percentages (Callam *et al*., 1992)

Bandage	Healed		Did not heal		Total	
	n	%	n	%	n	%
Elastic	35	53.8	30	46.2	65	100.0
Inelastic	19	28.4	48	71.6	67	100.0
Total	54	40.9	78	59.1	132	100.0

chi-squared tests gives P values which are too small and is 'anti-conservative'. This only matters when the frequencies are small.

Risk ratio

We now turn to methods for 2 by 2 tables only. Consider the venous leg ulcer bandaging trial shown in Table 4. We want an estimate of the size of the treatment effect. One we have already looked at is the difference between two proportions, for which we can find a standard error, a large sample confidence interval using the standard error, and a small sample confidence interval using exact probabilities. For Table 4 the two proportions are 0.538 and 0.284, or 53.8% and 28.4%, for elastic and inelastic bandaging respectively. The difference is 0.538 – 0.284 = 0.254 or 53.8% – 28.4% = 25.4 percentage points.

The proportion who heal is called the **risk** of healing for that population. This sounds rather odd, since risk usually refers to something bad and healing is a good thing, but because we use exactly the same methods to analyse the proportions experiencing desirable events as for the proportions who experience undesirable events we use the same term for both. The difference is called the **risk difference**, **absolute risk difference**, or **absolute risk reduction**.

We can look at the difference in risk between the two treatment groups in a different way. The ratio of the risk of healing in the elastic bandage group to the risk in the inelastic bandage group is called the **risk ratio**. For Table 4, the risk ratio = 0.538/0.284 = 1.89. The risk ratio is also called the **relative risk** and the **rate ratio**, all of which can be conveniently abbreviated to **RR**.

Having calculated our estimate of effect, we would like a confidence interval for it. Ratios are rather difficult things to deal with statistically. Because risk ratio is a ratio, it has a very awkward distribution. Variations which makes the denominator smaller have a much bigger effect on the ratio than do those which make the denominator larger, and we get a skew distribution. The result of this is that confidence intervals for risk ratios are not symmetrical. For the data of Table 4, the 95% CI for RR = 1.22 to 2.95. So we estimate that the proportion who will heal given elastic bandaging is between 1.22 and 2.95 times the proportion who heal given inelastic bandaging. For a test of significance of the null hypothesis that RR = 1, we use exactly the same chi-squared test and P value as before.

This confidence intervals are large sample approximations. There are several better ones which are used for small frequencies. A good guide is that if all four frequencies exceed five, the large sample method will be OK. We have problems when one of the frequencies is zero, as the relative risk may be zero or not able to be calculated at all.

Table 7. Leg ulcer wound healing by type of bandage in a randomised trial, with order of columns reversed (Callam *et al*., 1992)

Bandage	Did not heal		Healed		Total	
	n	%	n	%	n	%
Elastic	30	46.2	35	53.8	65	100.0
Inelastic	48	71.6	19	28.4	67	100.0
Total	78	59.1	54	40.9	132	100.0

Estimation is very difficult, but we can find an upper or a lower limit even if we cannot estimate the RR.

Odds ratios

The odds ratio is another method to estimate the relationship in a 2 by 2 table. First, we shall look at odds. This is a familiar concept from sports and gambling. In statistics the **odds** of an event is number of times it happens divided by the number of times it doesn't happen. For example, in Table 4 there were 65 patients who received elastic bandages, of whom 35 healed and 30 did not. The risk of healing = 35/65 = 0.538. The odds of healing = 35/30 = 1.17. So

Risk = number experiencing event divided by number who could.

Odds = number experiencing event divided by number who did not experience event.

Another way to look at this is that risk = 0.538 means that for every person treated, 0.538 people heal, or for every 100 people treated, 53.8 people heal. Odds = 1.17 means that for every person who do not heal, 1.17 people heal, or for every 100 people who do not heal, 117 people heal.

Another way to define the odds is that it the probability of the event divided by one minus the probability of the event. Hence the odds of healing is 0.538/(1 — 0.538) = 1.17.

The **odds ratio** is the odds of healing given elastic bandages divided by the odds of healing given inelastic bandages. As we have seen, the odds of healing given elastic bandages = 35/30 = 1.17. Similarly, the odds of healing given inelastic bandages = 19/48 = 0.40.

$$\text{Odds ratio} = \frac{35/30}{19/48} = \frac{1.17}{0.40} = 2.95.$$

For every person who does not heal, 2.95 times as many will heal with elastic bandages as will heal with inelastic bandages.

'Odds ratio' is often abbreviated to 'OR'. Like RR, OR has an awkward distribution and the confidence interval is not symmetrical. For the data of Table 4, the 95% CI for OR = = 1.43 to 6.06.

One of the great things about the odds ratio is that it doesn't matter which way round we do it. We can find the odds ratio for treatment by elastic bandage given healing. 35 patients healed and had elastic bandaging compared to 19 to healed with inelastic

bandaging, so the odds of elastic bandaging for those who healed was 35/19. Similarly, the odds of elastic bandaging for those who did not heal was 30/48. Hence

$$\text{Odds ratio for treatment: OR} = \frac{35/19}{30/48} = 2.95 \text{ as before.}$$

Both these versions of the odds ratio are the same:

$$\frac{35/30}{19/48} = \frac{35/19}{30/48} = \frac{35\times 48}{19\times 30} = 2.95$$

So both versions of the odds ratio = (35×48)/(30 ×19). We also call this the **ratio of cross products**, because the numerator is the top left cell frequency multiplied by the bottom right and the denominator is the top right cell frequency multiplied by the bottom left.

Switching the order of the rows or columns inverts the odds ratio. Table 7 shows the data of Table 4 with the order of the columns reversed. We can calculate the odds ratio for not healing given elastic bandage from the ratio of cross-products in this table: OR = (30/35)/(48/19) = 0.339 = 1/2.95. So this is one over the odds ratio for healing. In fact, there are only two possible odds ratios for a 2 by 2 table, reflecting the directions in which we might look at the relationship. On the log scale, these are equal and opposite: $\log_e(2.95) = 1.082$ and $\log_e(0.339) = -1.082$.

Odds ratios have the same problems when frequencies are small as do relative risks.

References

Bland M. (2000) *An Introduction to Medical Statistics, 3rd edition.* Oxford University Press.

Callam MJ, Harper DR, Dale JJ, Brown D, Gibson B, Prescott RJ, Ruckley CV. (1992) Lothian Forth Valley leg ulcer healing trial—part 1: elastic versus non-elastic bandaging in the treatment of chronic leg ulceration. *Phlebology* **7**, 136-41.

Meadows J, Jenkinson S, Catalan J. (1994) Who chooses to have the HIV antibody test in the antinatal clinic? *Midwifery* **10**, 44-48.