7 July 2006

Clinical Biostatistics: 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

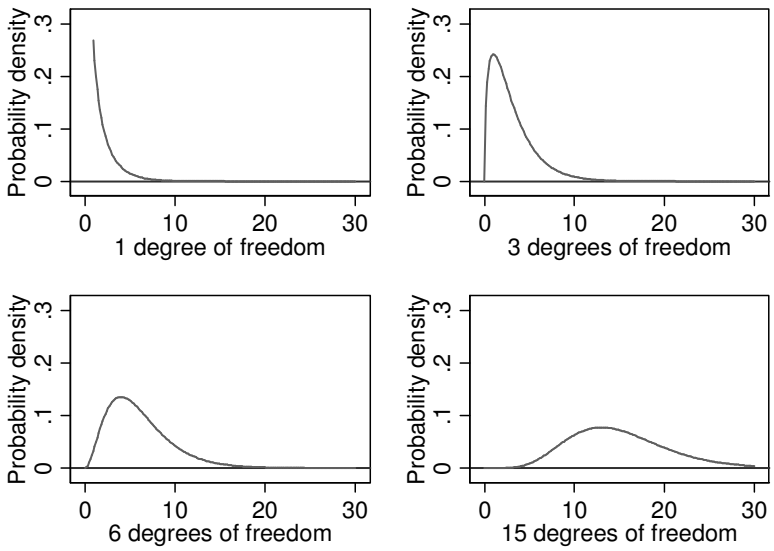

Figure 1. Some members of the Chi-squared distribution family.

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.

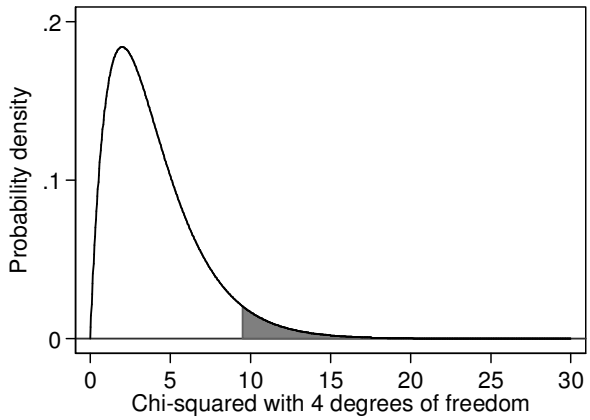

Figure 2. Upper 5% point of the Chi-squared distribution with 4 degrees of freedom, as shown in Table 2.

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, or, for 2 by 2 table, Yates' correction, both described below.

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

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

There are two different ways of doing Fisher's exact test. The test above was done using the statistical program Stata. I also tried it using my own program Clinstat. This gave 0.0049. This is not because I am a poor programmer, but because there are two different ways to calculate the Fisher probability. When the test was first devised, it was strictly for 2 by 2 tables. You could do a one-tailed test, by summing the probability for your observed table and for all the more extreme ones in the same direction. For Table 4, this meant making the number in the first cell (35) progressively bigger (36, 37, etc., up to 54) and calculating the probability for each of them. These probabilities would be added. This gave a one-sided test. To make a two-sided test it was not clear what tables you would use, nothing is symmetrical. What was done originally was to double the one-tailed P value. This is what Clinstat does. It could give probabilities greater than 1.0, but these would be set equal to 1.0. When you move to a table with more than two rows or columns, the direction of the effect doesn't mean anything and all tests will be two-sided. The way Stata does it is the only way for larger tables and so I now recommend it as the better way. Programs have been written which do both and let you make up your own mind. I would just do whatever your software does, but you should always state which program you use so that anybody who really wants to know can find out which method was used.

Even with the computers available at the time of writing, very large tables can defeat them. There are so many probabilities to compute that they cannot store them all and cannot calculate them before the summer holidays. Some programs, such as SPSS, offer a Monte Carlo option. This means that tables are generated at random and a sample of all the possible tables is produced and their probabilities calculated. This is used to estimate the P value. If you do not have enough simulations, you might get a slightly different answer if you do it again, so make sure you ask it for a few thousand.

Yates' correction

For tables with small expected frequencies, the chi-squared test gives smaller probabilities than Fisher's exact test. If you believe the exact probability, then the chi-squared test is wrong. Yates introduced a modified chi-squared test for a 2 by 2 table which approximates the Fisher's exact probability very closely. It works by making the difference between observed and expected frequency closer to zero by 0.5 before squaring. It works extremely well. For Table 4, Fisher's exact test gives P = 0.0049 and the ordinary chi-squared test gives the chi-squared statistic = 8.87, d.f. = 1, P = 0.0029. The chi-squared test with Yates' correction gives a smaller chi-squared statistic = 7.84, d.f. = 1, P = 0.0051. This is very close to the 0.0049 given by the exact test. It is a remarkably good modification to the chi-squared test and much easier to calculate without a computer than is the exact test.

Yates' correction is also called the continuity correction for the chi-squared test. This is because it makes allowance for the fact that for a table with few observations there are only a small number of possible values for the chi-squared statistic. We approximate the discrete distribution of the test statistic with the continuous Chi-squared distribution and Yates' correction improves this approximation.

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

Table 5. Number of antenatal visits by type of unit in a survey of women in Scotland (Hundley et al., 2002)

Type of maternity unit		Number of antenatal visits that women received				Total
		0-4	5-9	10-14	15 or more	
Traditional model	n	10	82	167	72	331
	%	37.0	30.8	40.7	46.2	38.5
New model	n	17	184	243	84	528
	%	63.0	63.0	69.2	59.3	61.5
Total	n	27	266	410	156	859
	%	100.0	100.0	100.0	100.0	100.0

Table 6. Change in self-reported smoking habit by pregnant women in a trial of motivational interviewing by midwives (Tappin *et al.*, 2005)

Change in smoking habit	Motivational interviewing		Control	
Quit	24	6.8%	31	7.5%
Cut down	31	8.8%	52	12.7%
Same	278	79.2%	286	69.6%
More	18	5.1%	42	10.2%
Total	341	100.0%	417	100.0%

Even Yates' correction makes the calculation without a computer more time-consuming than the simple chi-squared test and it used to be recommended only when there was at least one expected frequency less than five. This convention continued when people began routinely to do these calculations using computers. However, if Yates' correction is better for small tables, it must be better for all tables.

Yates' correction is now obsolete as we can always do the exact test and we will start to see it less often.

The chi-squared test for trend

Table 5 shows data from a study of maternity care in Scotland. We might carry out a chi-squared test for association on this table, which would give chi-squared = 11.36, d.f. = 3, P = 0.01. Many researchers would do exactly this. Hundley *et al.* (2002) did not, however, they reported chi-squared = 9.33, d.f. = 1, P = 0.002. They used a different test called the chi-squared test for trend.

In Table 5, the categories '0-4', '5-9', '10-14', and '15 or more' have a clear order, unlike the marital status categories in Table 1. The chi-squared test for association does not take the ordering of the categories into account. We would get the same chi-squared statistic however we shuffled the rows of Table 1. This is as it should be, we would not want to treat 'married', 'living with partner', 'single', and 'divorced/widowed/separated' as a meaningful ranking, because it isn't. If we write the 'single' row first, we want to get the same P value. When we have ordered categories, we should use the information which the ordering gives us.

In the chi-squared test for trend, we not only use the order of the categories, but attach a numerical value to them. Hundley *et al.* (2002) have attached the values 1, 2, 3, and 4 to the categories '0-4', '5-9', '10-14', '15 or more'. This is reasonable, as these categories are derived from numbers of visits and each category except the last covers the same range of visits, 5. Often we do not have an underlying qualitative variable, such as when a subject's condition is classified as 'poor', 'good', fair', or 'excellent'. These are clearly in ascending order, but our decision as to what numerical value to attach is arbitrary, a matter of judgement. When we have no reason to decide otherwise we usually make the intervals equal, i.e. we score them 1, 2, 3, etc.

The chi-squared for trend statistic is always less than the chi-squared for association statistic. In this example, it is 9.33 compared to 11.36. However, it has fewer degrees of freedom and, if there really is a trend, will have a smaller P value. Hence it is a more powerful test when categories are ordered. The difference between the two chi-squared statistics also follows a Chi-squared distribution if the null hypothesis is true, with degrees of freedom equal to the difference between the two degrees of freedom. For the example we have chi-squared = 11.36 – 9.33 = 2.03, with degrees of freedom 3 – 1 = 2, P = 0.4. This is called the chi-squared about trend, and tests whether there is evidence for any association in the table which would not be explained by a linear trend. In this example, the data are clearly consistent with the null hypothesis that there is no non-linear association.

Many qualitative variables with more than two categories have ordered categories and should be analysed in some way which takes the ordering into account. There are some slight variations on the formulae used in this test and different programs might give slightly different answers. A different approach, which some programs use, is rank correlation, in particular Kendall's tau b. Stata does this and for Table 5 gives Kendall's tau-b = –0.1031, SE = 0.032. If we use this for a test of significance we get

P = 0.001, slightly more significant than the chi-squared test for trend in this example. This may be because the category representing the lower extreme, '0-4' has a greater percentage in the 'traditional model' than the next category, '5-9', though after the percentages rise. Because there are only a few subjects in this category, this is insufficient to produce a significant association not explained by the trend.

There are two other points worth noting about Table 5. First, I have given the percentages for the columns, using the number of women in the visit category as the denominator. It would have been more logical to take them for the rows, with the number of women in the type of unit as the denominator, since it is the types of unit which we want to compare. This is what Hundley *et al.* (2002) did. My percentages were calculated to bring out the trend across the table. Second, there were 6 traditional units and 14 new model units. So the women were not really independent, as we might expect the number of visits to be related to the practice in the particular unit. An assumption of the test has been violated. It might have been better to estimate an average number for each unit and compare these with a two-sample method for continuous data. There are often many possible ways to analyse data, each of which may have advantages and disadvantages compared to the others.

For another example, Tappin *et al* (2005) carried out a trial of motivational interviewing by midwives to help pregnant smokers to quit or cut down Table 6 shows the change in self-reported smoking habits by treatment group. The authors reported χ^2 for trend = 0.01, P=0.98, χ^2 non-linear = 11.26, P=0.004 (2 d.f.). There is a significant relationship overall, χ^2 = 11.26, 3 d.f., P = 0.01. All the chi-squared is in the part which is about trend. There is a relationship, but it is not a linear trend. The proportion of women in the motivation interviewing group goes down from 'quit' to 'cut down', then goes up to 'same', then drops to 'more'. Figure 3 illustrates the two trend tests. Neither shows a completely smooth trend in one direction, but for the antenatal visits the three large groups clearly do. There is not discernable direction for the smoking study.

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.

We can also carry out a test of the null hypothesis that the proportions healed are the same for the populations of ulcer patients treated by elastic and inelastic bandages. This can be done using the standard error for testing the null hypothesis. This test gives a test statistic which, follows a Standard Normal distribution if the null hypothesis is true. For Table 4 this is z = 2.98, which has two-tailed probability P = 0.0029. We would usually present this as P = 0.003. This is exactly equivalent to the chi-squared test for association, and always gives an identical P value. The test statistic here is the square root of the chi-squared.

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. We deal with this using logarithms. Taking the log of a ratio gives us the difference between the logs of the two numbers. If we take the log of the rate ratio, we have something which is found by adding and subtracting log frequencies. The distribution becomes approximately Normal and, provided the frequencies are not small, it has a simple standard error.

For Table 4, the log risk ratio is $\log_e(RR) = 0.6412$. Its standard error is 0.2256 and so we can find a 95% confidence interval for $\log_e(RR)$ by

$$0.6412 - 1.96 \times 0.2256 \text{ to } 0.6412 + 1.96 \times 0.2256$$
$$= 0.1990 \text{ to } 1.0834.$$

Most of us do not think easily in terms of logarithms, so we transform this back to the natural scale by the antilog or exponential: 95% CI for RR = exp(0.1990) to exp(1.0834) = 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. Although we have a different confidence interval for the risk ratio, we use exactly the same P value as before.

The risk ratio is not in the middle of its confidence interval, unlike the risk difference. The confidence interval is symmetrical on the log scale, not the natural scale.

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

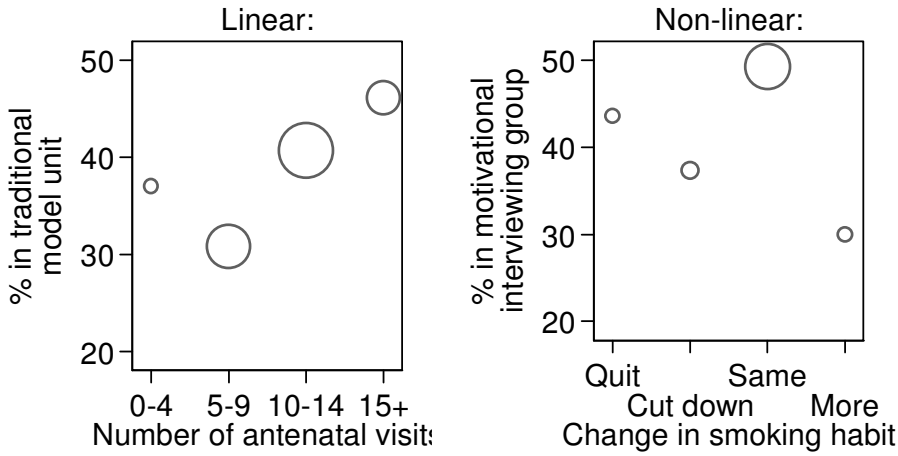

Figure 3. Two trend analyses

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

Table 8. Smoking history of stroke patients (cases) and controls, with row percentages (data of Markus *et al.*, 1995)

Patient group	Smoked		Never smoked		Total	
	n	%	n	%	n	%
Stroke patients	71	70.3	30	29.7	101	100.0
Healthy controls	36	26.3	101	73.7	137	100.0
Total	107	45.0	131	55.0	238	100.0

Table 9. Artificial data obtained by multiplying the frequencies for controls by 100.

Patient group	Smoked		Never smoked		Total	
	n	%	n	%	n	%
Stroke patients	71	70.3	30	29.7	101	100.0
Healthy controls	3600	26.3	10100	73.7	13700	100.0
Total	3671	45.0	10130	55.0	13801	100.0

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 we estimate the confidence interval in the same way. We use the log odds ratio. When we do this the distribution becomes approximately Normal and, provided the frequencies are not small, it has a simple standard error.

$\log_e(\text{OR}) = \log_e(2.95) = 1.0809.$

SE $\log_e(\text{OR}) = 0.3679.$

As usual, we will not go into the details of how this was calculated, as we would expect anyone who wanted to do statistical analysis would use a computer, but Bland (2000) gives details. We find the 95% CI for $\log_e(\text{OR})$ using ±1.96 standard errors by the usual large sample method:

$$1.0809 - 1.96 \times 0.3679 \text{ to } 1.0809 + 1.96 \times 0.3679$$
$$= 0.3598 \text{ to } 1.8020.$$

To find the 95% CI for the odds ratio itself we antilog these limits:

95% CI for OR = exp(0.3598) to exp(1.8020) = 1.43 to 6.06.

Hence our estimated odds ratio is OR = 2.95, 95% CI = 1.43 to 6.06.

As for RR, OR is not in the middle of its confidence interval. The interval is symmetrical on the log scale, not the natural scale.

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.

Odd ratios in case control studies

A case-control study is one where we take a group of subjects with a characteristic, the cases, and compare them to another group without the characteristic, the controls. Table 8 shows data from a case-control study, where a group of stroke patients, the cases, were compared with a group of healthy controls. The purpose of the study was to examine a gene (Markus *et al.*, 1995), but for this example we shall look at cigarette smoking as a risk factor for stroke. Because we started with stroke patients and controls, rather than smokers and non-smokers, we cannot estimate the proportion of smokers who have strokes. Hence we cannot calculate the risk of a stroke for a smoker or for a non-smoker, hence we cannot divide one by the other to get the relative risk. We can evaluate the odds ratio:

OR = (71×101)/(30×36) = 6.64.

Now, stroke is a common disease, but even so not many people in the population have had one. We don't know what the prevalence of past stroke is among the population being studied here, who were aged between 35 and 91years, but it is quite small. Purely for illustration, we are going to suppose it is 0.7% The reason for this that if we multiply the frequencies for the healthy controls by 100, the proportion of stroke patients will be 0.7%. Table 9 shows these artificial data. The row percentages are unchanged, and so is the odds ratio. It is still 6.64.

OR = (71×10100)/(30×3600) = 6.64.

Because we should now have the correct proportion of stroke cases, the proportion of stroke cases among the smokers should also be correct, 1.9, as should the proportion among the non-smokers. The relative risk should also be correct:

RR = (71/3671)/(30/10130) = 6.53

This is very similar to the OR. This is because when the frequencies in one category are much smaller than those in the other, OR and RR are much the same. If we did the same for Table 8, we would get

RR = (71/107)/(30/131) = 2.90

This means that in a case-control study, provided what defines a case is rare in the population, the odds ratio can be used as an estimate of the relative risk. The 95% confidence interval is 4.27 to 10.56, so we can say that we estimate that the risk of stroke among smokers is between 4 and 10 times that in non-smokers.

Risk ratio or odds ratio?

Risk ratio and odds ratio are two rather similar methods of summarising the data from a 2 by 2 table. Why do we need them both? The short answer is that RR has a more intuitive interpretation than OR, but OR has much more convenient mathematical properties and so is better for statistical analysis.

The first thing to note is that RR is more dependent on the order of rows and columns than is OR. For example, switching the columns does ***not*** invert the risk ratio. If we calculate the risk ratio for not healing given elastic bandage, we get:

RR = (30/65)/(48/67) = 0.644.

Compare this to the risk ratio for healing given elastic bandage, which we calculated earlier:

RR = (35/65)/(19/67) = 1.89

This is not inverse of 0.644, as is the case for OR, because 1/1.89 = 0.529, not 0.644.

Finding risks down the columns instead of across the rows produces more values for the risk ratio. The risk ratio for elastic bandage given not healing is

RR = (30/78)/(35/54) = 0.593.

and the risk ratio for inelastic bandage given not healing is

RR = (48/78)/(19/54) = 1.749.

Altogether there are eight possible rate ratios.

Consider two hypothetical tables:

	Success	Fail
Treat	20	80
Control	10	90

RR = (20/100)/(10/100) = 2.00

OR = (20×90)/(80×10) = 2.25

	Success	Fail
Treat	90	10
Control	80	20

(90/100)/(80/100) = 1.125

(90×20)/(10×80) = 2.25

These tables have the same data, different RRs, same OR. OR is a much better measure of the strength of the relationship than RR. However, RR has a more intuitive interpretation and for most people it is what they want to see. RR has some awkward properties, however, and OR is much more convenient for statistical analysis. As we shall see when we discuss logistic regression, OR also extends smoothly into more complex situations.

All this applies only to 2 by 2 tables. For tables with more than two rows or columns, estimation is complicated and such estimates are seldom used in health research. We can do tests of significance, using the chi-squared and Fisher's exact tests, but they do not provide estimates of the size of effects.

Number needed to treat

The **number needed to treat**, usually abbreviated to **NNT**, was devised as a useful way to present the results of a clinical trial so that clinicians could easily appreciate the effectiveness of a treatment. It is the number of patients we would need to treat with one treatment rather than another to benefit one patient.

An example will show what we mean. For the wound healing data of Table 4, the difference between the proportions healed is 0.538 – 0.284 = 0.254, or 53.8% – 28.4% = 25.4 percentage points. How many people must we treat with elastic rather than inelastic bandages to heal or benefit one extra person? If the difference is 25.4 percentage points, this means that for every 100 people we treat with elastic rather than inelastic bandages, 25.4 would heal who would not if given inelastic bandages. Hence to heal one extra person we must treat 100/25.4 = 3.9 patients. For every 3.9 people treated with elastic bandages rather than inelastic we estimate that one extra person is healed. Another way to write this is that the extra people healed per person treated = 0.254. The number needed to treat = 1/0.254 = 3.9. Clearly a small NNT is good, as we need treat only a few patients to achieve another healing.

We can find a 95% confidence interval easily. We find the 95% CI for the difference between the two proportions and invert it. For the difference, the 95% CI = 0.093 to 0.417. The 95% CI for the NNT = 1/ 0.093 to 1/0.417 = 10.8 to 2.4. We turn this round to give 95% CI = 2.4 to 10.8.

Table 10. Leg ulcer wound healing by type of bandage in a randomised trial (Northeast *et al*., 1990)

Bandage	Healed		Did not heal		Total	
	n	%	n	%	n	%
Elastic	31	63.3	18	36.7	49	100.0
Inelastic	26	50.0	26	50.0	52	100.0
Total	49	53.4	52	43.6	101	100.0

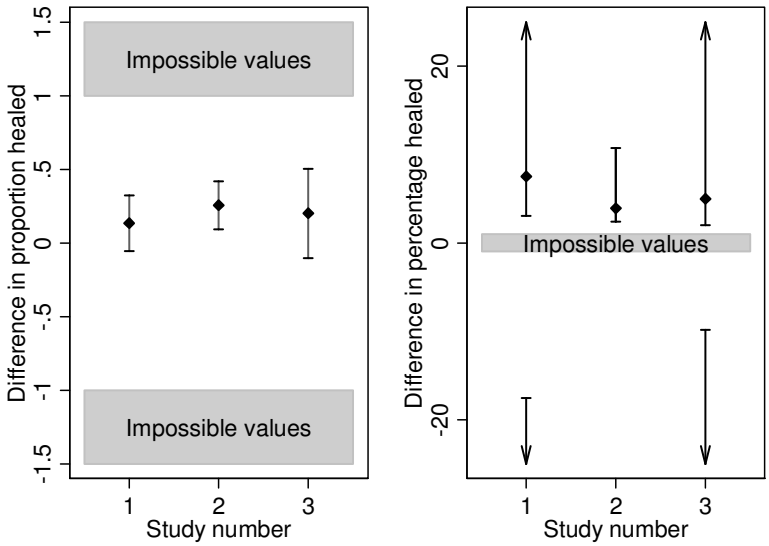

Figure 4. 95% confidence intervals for three trials, showing differences between proportions healed and the numbers needed to treat

NNT is straightforward when the difference is significant and the confidence interval for the difference does not include zero. Consider Table 10, which shows a different wound healing study. The proportions healed are 0.633 and 0.500, so the difference, elastic minus inelastic, is 0.633 – 0.500 = 0.133. The number needed to treat is therefore NNT = 1/0.133 = 7.5. We estimate that we need to treat 7.5 patients to heal one extra person. What about the confidence interval for this estimate? The 95% CI for the difference is –0.059 to 0.324. The 95% CI includes 0.0, and the difference is not significant (chi-squared = 1.81, d.f. = 1, P=0.2). If we take one over these limits we get 95% CI = 1/(–0.059) to 1/ 0.324 = –16.9 to 3.1. What does this mean? Can NNT be negative? A negative NNT arises because the proportion healed on the test treatment (elastic bandage) is less than the proportion healed on the control treatment (inelastic). The treatment does more harm than good. This is sometimes called the **number needed to treat to harm**, **NNTH**, or **NNH**. Now, the NNT cannot be between –1 and +1. It is 1.0 divided by the difference between two proportions, each of which must be 1.0 or less, so their difference must be between –1.0 and +1.0. Hence the reciprocal must be greater than 1.0, if the difference is positive, and less than –1.0 (i.e. more negative) if the difference is negative. If the difference = 0.0, then the NNT will be infinite, i.e. no matter how many patients we treat no extra person will heal or be harmed. When the difference is not significant, the confidence interval goes off to infinity in either direction. Figure 4 shows forest plots for the differences between proportions healed and the number needed to treat for the three ulcer trials described by Fletcher *et al.* (1997). So how can we describe this bizarre interval? For Study 1, we should say: '95% CI = 3.1 to ∞, NNTH = 17.5 to ∞' or '95% CI = 3.1 to ∞, NNTH = –17.5 to –∞'. The symbol '∞' means 'infinity'. But really the number needed to treat is not helpful when the difference is not significant and there is little to be gained by quoting it.

The NNT is also called the **number needed to treat to benefit** or **NNTB**. Others have extended the idea to include the number needed to screen (Rembold 1998) and other such things, but it all the same really.

Pitfalls in the analysis of categorical data

A common pitfall in the analysis of categorical data is ignoring ordering in the categories. A chi-squared test of association is done when a chi-squared test for trend would be more powerful.

A second pitfall is to treat each category as though it were a separate variable. This takes the form of separate tests for each row or column of a table. For example, Meadows et al. (1994) did not use the chi-squared test for association which we used their data to illustrate. Rather than produce one P value for the table, they produced four. They compared each category with the other three combined. For example, married women were compared with all unmarried, whether cohabiting, single, divorced, widowed, or separated. This is, of course, multiple testing.

Another pitfall is to carry out the chi-squared test for association when the expected frequencies are too small. The effect of this is to make the P values too small. Most programs will either warn you when the number of expected frequencies below five is too big, or tell you how many there were, or display them for you if you ask.

A pitfall we might encounter is trying to calculate relative risks and odds ratios when some of the frequencies are very small. This can make the ratio very unstable. The

large sample confidence interval estimates will be inaccurate, but will also be very wide, reflecting this.

Another common pitfall is the interpretation of relative risks without regard to the absolute risk. We might say that some risk factor doubles the risk of a bad outcome, but if that bad outcome is a rare event, the absolute risk is increased by very little. For example, being red-green colour blind is very strongly associated with being male, because the main forms are carried on the X chromosome. Men have only one of these, so if it has the gene they will be colour-deficient, about 8% so being. Women have two X chromosomes and need to have both carrying one of the colour-blindness genes, 0.4%, having this. The relative risk of colour-blindness for men is therefore 8/0.4 = 20, a huge relative risk, considerably greater than the relative risk of lung cancer for smokers compared to non-smokers. However, the absolute risk is small and the absolute increase in risk for a man compared to a woman is only 7.6 percentage points. Despite the huge relative risk, most men are not colour-blind.

Although relative risks and odds ratios are similar when the event is rare, they are not otherwise and we should not interpret an odds ratio as if it were a relative risk.

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.

Fletcher A**,** Cullum N**,** Sheldon TA. (1997) A systematic review of compression treatment for venous leg ulcers. *British Medical Journal* **315**, 576-580.

Hundley V, Penney G, Fitzmaurice A, van Teijlinen E, Graham E. (2002) A comparison of data obtained from service providers and service users to assess the quality of maternity care. *Midwifery* **18**, p 126-135.

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

Northeast ADR, Layer GT, Wilson NM, Browse NL, Burnand KG. (1990) Increased compression expedites venous ulcer healing. *Royal Society of Medicine Venous Forum*. London: Royal Society of Medicine.

Rembold CM. (1998) Number needed to screen: development of a statistic for disease screening. *British Medical Journal* **317**, 307-312.

Tappin DM, Lumsden MA, Gilmour WH, Crawford F, McIntyre D, Stone DH, Webber R, MacIndoe S, Mohammed E. (2005) Randomised controlled trial of home based motivational interviewing by midwives to help pregnant smokers quit or cut down. *British Medical Journal* **331**, 373-5.