STATISTICS FROM THE INSIDE

10. Counted data (2)

M J R Healy

McNemar’s test for paired data

The example of the comparison between two independent counted proportions that I discussed in my previous article was analogous to the continuous-data structure that is dealt with by unpaired *t*. What is the counted-data analogue of the paired *t* situation? Suppose that our samples are not independent, but that we have investigated the effect of a treatment using matched pairs of patients, one member of each pair being allotted to the treatment and the other to the control in a random fashion. There are then four possible patterns of response for a particular pair and it is necessary to summarise the findings in a different manner by enumerating these four patterns, as in the example in table 1. Note carefully that the counts in this table are counts of pairs – the study utilised a total of 100 subjects. In this example there are 14+24=38 concordant pairs in which both subjects respond or both subjects fail to respond. These provide no evidence as to the superiority or otherwise of the treatment over the control and have to be ignored. There are two kinds of discordant pairs, response in treated but not in control and response in control but not in treated. The first of these suggests that the treatment has a positive effect, the second that it has a negative effect, and on the null hypothesis the two are equally likely. We are thus led to consider our data as a binomial sample of size no more than 8+4=12, with a probability of 0.5 and 8 successes observed. The mean is 6, the variance is 12×0.5×0.5=3.0 and the standard deviation is 1.73 so that the discrepancy corrected for continuity amounts to 1.5/1.73=0.87 standard deviations, given an approximate two tailed significance level of 0.384 (the exact value is 0.387). This procedure is called *McNemar’s test*. In a general way, pairing and matching, which can be very important with a continuous variate, are much less so when the variate is of the yes/no type.

Fisher’s exact test

With an ordinary 2×2 table, the analysis using χ^2 which I described in the previous article involves the Normal approximation to the binomial distribution and this is quite satisfactory provided none of the expected values in the table falls below 5 or so (note that small observed numbers are quite acceptable). It is however possible to make an exact analysis involving no approximation – the arithmetic can be heavy but a good computer package will take care of this. The procedure is

known as *Fisher’s exact test*. I shall not go into details here, but instead I emphasise a crucial and rather worrying point. The procedure as originally presented by R A Fisher provides a one tailed significance level. Many computer programs follow suit, and many users of these programs quote these one tailed probabilities without being aware that they are doing so. A two tailed test is what is usually required, but there are several ways of obtaining a two tailed significance level, the results can be markedly different, and there is sharp disagreement between statisticians as to which method is to be preferred. With considerable hesitation I report the received wisdom which states that the one tailed significance probability should simply be doubled, merely pointing out that the resulting quantity does not seem to me to be the probability of any definable event and referring the interested reader to a magisterial paper by F Yates, Tests of significance for 2×2 contingency tables (*Journal of the Royal Statistical Society A* 1984; 147: 426–63) and to the subsequent discussion. It is fairly remarkable that after 100 years or so of development of statistical methodology there is still controversy about the correct way to compare two proportions.

Tables with more than two rows or columns

In the previous article I promised two reasons for introducing the use of χ^2 when comparing two counted proportions. The first was that the methodology showed how to introduce a continuity correction. The second reason is that it readily generalises to the analysis of tables of counts with more than two rows and/or columns (these are sometimes referred to as *contingency tables*). As an example, consider the problem of comparing three or more proportions, such as the rates of response to increasing doses of a drug (see table 2). This is a contingency table with two rows and five columns. As before we can draw up a table of the values expected on the null hypothesis by making the proportion responding

Table 1 Patterns of response in paired data

	Treated		Total
	Responder	Non-responder	
Control			
Responder	14	4	18
Non-responder	8	24	32
Total	22	28	50

23 Coleridge Court,
Milton Road,
Harpenden, Herts
AL5 5LD
Correspondence to:
Professor Healy.
No reprints available.

Table 2 Observed numbers

	Dose					Totals
	0	1	2	3	4	
Responders	7	9	12	12	17	57
Non-responders	13	12	7	10	6	48
Totals	20	21	19	22	23	105
Proportion	0.350	0.429	0.632	0.546	0.739	0.543

Table 3 Expected numbers

	Dose					Totals
	0	1	2	3	4	
Responders	10.86	11.40	10.31	11.94	12.49	57
Non-responders	9.14	9.60	8.69	10.06	10.51	48
Totals	20	21	19	22	23	105

Table 4 Residuals (observed-expected)

	Dose					Totals
	0	1	2	3	4	
Responders	-3.86	-2.40	+1.69	+0.06	+4.51	0
Non-responders	+3.86	+2.40	-1.69	-0.06	-4.51	0
Totals	0	0	0	0	0	0

at each dose equal to 0.543, the same as that overall (see table 3).

From table 3 we can get the table of discrepancies (see table 4) and from this the value of χ^2 can be obtained as 8.28 (note that no continuity correction is required in this context). In order to find the corresponding significance probability we need to take account of the fact that we now have more than two columns. There are on the face of it $5 \times 2 = 10$ numbers in our data. But we are taking the totals in the margins of the original data table for granted (they tell us nothing about the question we are interested in) and as a reflection of this, the marginal totals of the tables of observed and expected values are the same. It follows from this that, once we have filled in the first four numbers in the top row in the table of expected values, all the other entries in the table can be completed simply by making them add up to the correct row and column totals. We have in fact only 4 degrees of freedom. This can be verified by looking at the discrepancies, which add up to 0 in both directions so that the first four in the top row determine all the rest. Looking in the table of χ^2 and using 4 degrees of freedom, the significance probability for our data is somewhere between 0.10 and 0.05 – suggestive of a real effect of dose but not very impressive.

Testing for trend

We have not taken account of one feature of our data, the fact that the five columns in the table are not arbitrary but correspond to five increasing doses. Suppose we do a simple regression of the proportions responding on dose, using the binomial variance of $0.543 \times (1 - 0.543) = 0.248$ as the residual variance. The regression will need to be weighted to take account of the different numbers in the five groups, but this is a trivial complication of the arithmetic. The resulting slope turns out to be 0.0895 (SE 0.0340) corres-

ponding to a Normal deviate of 2.63 and a significance probability of 0.0036. Once we ask a specific question ('do the proportions change steadily with dose?') rather than the vague omnibus question ('do the proportions differ in any way at all?'), we get a much more definite answer.

It may have occurred to the reader that the χ^2 that we calculated for the 2×2 table in the previous article had only 1 degree of freedom (as you can see from that article, the four discrepancies were the same number four times over, apart from the sign). It will be remembered that the square root of this was a Normal deviate. We have just obtained a Normal deviate of 2.63, and in exactly the same way, the square of this is a χ^2 with 1 degree of freedom, $2.63^2 = 6.92$. This quantity is usually called χ^2 for trend. It forms part of the overall χ^2 of 8.28, the remainder of which can be used to check that the simple straight line relationship that we have assumed adequately fits the data. In our case this remainder is only 1.36, and as a χ^2 with $(4 - 1) = 3$ degrees of freedom this is actually somewhat better than we have a right to expect by chance.

Other uses of χ^2

Two further uses of the χ^2 distribution may be mentioned here. Suppose that we culture some organism and count the numbers of cells that appear on the grid of a haemocytometer. If the cells are distributed randomly on the slide, the counts in the grid squares should be a sample from a Poisson distribution. Can we test this supposition as a null hypothesis? Table 5 shows some data of this kind. The Poisson expected frequencies have been calculated from the formula given in the previous article with the mean μ set equal to the mean of the sample counts, 2.66.

The value of χ^2 from table 5 is 17.9. What about degrees of freedom? The Poisson expected counts add up to the same total as the observed counts, and they also have exactly the same mean as the observed counts. This has used up 2 degrees of freedom leaving 9 degrees of freedom over. A value of χ^2 of 17.9 on 9 degrees of freedom corresponds to a probability of only 0.036 – it appears that the Poisson distribution is not a very good model for our data.

But it must be remembered that the p values in the table of χ^2 are only approximations, usable provided only that the expected values used in the calculation do not go below about 5. Here

Table 5 Haemocytometer counts

Cells per square	Observed (O) No of squares	Expected (E) Nos (Poisson)	O-E
0	18	14.3	+3.7
1	42	37.7	+4.3
2	48	49.7	-1.7
3	35	43.8	-8.8
4	27	28.9	-1.9
5	14	15.2	-1.2
6	6	6.7	-0.7
7	6	2.5	+3.5
8	3	0.8	+2.2
9	1	0.2	+0.6
10+	0	0.2	-0.2
	200	200.0	0.0

Table 6 Haemocytometer counts

Cells per square	Observed (O) No of squares	Expected (E) Nos (Poisson)	O-E
0	18	14.3	+3.7
1	42	37.7	+4.3
2	48	49.7	-1.7
3	35	43.8	-8.8
4	27	28.9	-1.9
5	14	15.2	-1.2
6+	16	10.4	+5.6
	200	200.0	0.0

just this has happened; the Poisson expected counts for 7 or more cells per square are distinctly on the small side. To get round this it is necessary to lump together the small frequencies as in table 6. The value of χ^2 is now 6.51 with 5 degrees of freedom, giving a p value of 0.26 – most of the evidence for a bad fit by the Poisson model seems to have gone away. Unfortunately, much of this evidence lay in the values of the counts with small expectations at 7+ cells per square, and it is difficult for the χ^2 approximation to assess these properly.

We can do a bit better by approaching from a different direction. Suppose for the moment that the Poisson distribution is not a good fit to the data, what sort of misfit might we suspect? One real possibility is that the cells are slightly sticky, so that once we find one cell in a square we are rather likely to find another one sticking to it. Relative to the Poisson distribution based on a wholly random pattern of cells, there will be too many squares with high counts of cells and to balance this, too many squares with low counts. A glance at table 6 will show that this is what has actually happened. We shall in fact find a distribution which, for a given mean, has a larger variability than the Poisson distribution as measured by the standard deviation or the variance.

The mean of our observed distribution is 2.66 and this is also the mean of the fitted Poisson distribution. As we saw above, the variance of a Poisson distribution is equal to its mean, so that the expected value of the variance is also 2.66. The actual variance of the observed values can be calculated to be 3.51. Is this too big? We can answer this question by dividing the observed variance by that expected and then multiplying by the degrees of freedom, obtaining $3.51/2.66 \times 199 = 262.6$. This quantity is (to an excellent degree of approximation) a χ^2 value with 199 degrees of freedom and gives a p value of 0.001. There is in fact very strong evidence that the Poisson model does not provide a good fit to our data. Once again we see that a significance test chosen to be especially sensitive to a certain kind

of departure from the null situation can be much more incisive than an 'omnibus' test that has to try to guard against every possibility.

χ^2 and the sample variance

The use of χ^2 as a test of goodness of fit of a theoretical distribution is historically one of the oldest of statistical techniques (it dates back to the 1890s) and it is commonly given most emphasis in the textbooks. As we have seen, it is not always the best solution to the problem, and in clinical and epidemiological practice it is in any case a problem which does not arise very frequently. It is less appreciated that χ^2 provides a test for the agreement of an observed variance with a theoretical value. Suppose that we take a sample of size n from a Normal distribution and calculate the usual estimate of variance s^2 with $(n-1)$ degrees of freedom. Then, if σ^2 is the true variance of the population, the quantity $(n-1)s^2/\sigma^2$ is distributed (exactly) in the χ^2 form. The test for goodness of fit of a Poisson distribution described in the previous section is an application of this fact.

In the context of a continuous variable, where a theoretical variance is not usually available, the need for such a test does not often arise. However, as usual the testing procedure can be rearranged to provide confidence limits. Suppose that the 2½% and 97½% significance levels of χ^2 with $(n-1)$ degrees of freedom are h and k respectively. Then 95% confidence limits for σ^2 are given by $(n-1)s^2/k$ and $(n-1)s^2/h$.

Suppose for example that you wish to establish reference values for a new test whose results are Normally distributed. You measure a sample of 20 subjects and obtain a mean of 75 and a sample variance of 62 giving a standard deviation of 7.9. With 19 degrees of freedom the values of h and k are 8.91 and 32.85, so that the 95% confidence limits for σ^2 are $19 \times 62/32.85$ and $19 \times 62/8.91$, 35.86 and 132.21. These correspond to standard deviations of 6.0 and 11.5.

You should observe the substantial width of the confidence interval. The mean of a small sample may be quite precise if the variability of the material is small, but this is not true of the variance or standard deviation. To estimate a population standard deviation with 95% confidence limits of $\pm 20\%$, you cannot get away with fewer than 50 degrees of freedom.

Unlike the methods which I have discussed in previous notes in this series, this use of the χ^2 distribution depends crucially upon the Normality of the data. Relatively small departures from Normality, unlikely to be spotted in practice, can render the calculated confidence intervals noticeably too narrow.