STATISTICS FROM THE INSIDE

9. Counted data (1)

M J R Healy

Continuous and discrete variates

The previous notes in this series have usually assumed that the variate under consideration was a *continuous* quantity such as stature, blood pressure, or serum sodium concentration. All these are in principle capable of taking any value whatsoever within an appropriate range. Many studies, on the other hand, deal with *discrete* quantities in the form of counts—number of siblings, number of children in a treated group achieving remission or cure, and so on. These variates can only take the values 0, 1, 2, . . .; fractional values (and usually negative values also) are not possible.

The distinction between continuous and discrete variates is not as absolute as it may seem. Stature is (or should be) measured in whole millimetres, and fractions of a millimetre do not occur in practice. White blood count is in principle discrete, a count of the number of white cells in a stated volume of blood, yet it is seldom recorded to more than three significant figures and is treated for statistical purposes as if it were continuous. The special statistical problems associated with discrete data are most acute when the 'graininess' of the data is appreciable, that is when the gaps between the possible data values are a noticeable fraction of the total range of values which may occur.

The methods of analysis that I have so far described for the handling of continuous data have been based upon the Normal distribution as an acceptable approximate model for the underlying population. Again on the face of it, the Normal distribution cannot be an exact model for counted data. Suppose for example that litter size in a strain of mice had a mean of 6.2 and a standard deviation of 2.5. A value of 3.1 would be 1.24 standard deviations below the mean and one of 3.9 would be 0.92 standard deviations below the mean. If we were to assume that the distribution of litter size was Normal, there would be a probability of 0.071 of finding a value between these two limits, and it would follow that 7.1% of all litters should consist of more than three but less than four animals. Note however that there is nothing strange about a distribution of whole number quantities having a non-integer mean—jokes about the average family man taking his 2.2 children for a ride in his 0.8 of a car must be regarded as a bit passé.

Distributions for counted data

More realistic models for counted data are provided by the two basic discrete distributions, the binomial and Poisson distributions. The

binomial distribution is appropriate when the data to be modelled consist of counts of what may be regarded as 'successes' and 'failures' in a sequence of independent trials. It can thus be used to model data from clinical studies in which the end point is a yes/no quantity such as survival or cure after a specified period, or in certain epidemiological contexts for the numbers in particular groups possessing some attribute such as a specified blood group or tissue type. The Poisson distribution is appropriate to counts of events taking place independently in time or space, when no rigid upper limit is present. The type case is that of counts per minute resulting from radioactive decay, as used by many analytic techniques. The assumption of independence is (as always) very important. It would be possible to record the numbers of cases of measles in several families of size four, counting each case as a 'success', but it is not to be expected that the binomial distribution would be a good model for the data; once one case is recorded in a family, the occurrence of another case becomes distinctly more probable.

For those interested in such things, the mathematical formulas for these two distributions are as follows:

BINOMIAL

Prob(r successes out of n trials, probability per trial π) = $\binom{n}{r} \pi^r (1-\pi)^{n-r}$

(Here π is the Greek equivalent of p , and does not stand for 3.14159 . . .)

POISSON

Prob(r events) = $e^{-\mu} \mu^r / r!$

Here $r!$ stands for $r(r-1)(r-2) \dots 2 \cdot 1$ (with 0! being taken to be equal to 1) and $\binom{n}{r}$ stands for $n! / (r!(n-r)!)$. These formulas are not of any great practical importance but two important facts follow from them. The first is that the means and variances of the two distributions are given in terms of the parameters as follows:

	Mean	Variance
Binomial	$n\pi$	$n\pi(1-\pi)$
Poisson	μ	μ

The second fact is that both distributions resemble a Normal distribution provided the mean is not too small—not less than 5, say (in the binomial case, both $n\pi$ and $n(1-\pi)$, the mean number of 'failures', should both be not less than 5). This statement appears to contra-

23 Coleridge Court,
Milton Road,
Harpenden,
Herts AL5 5LD

Correspondence to:
Professor Healy.
No reprints available.

dict what I said above about the Normal distribution being an inappropriate model for discrete data; what is meant is that the probability of a binomial or Poisson variate exceeding some whole number value can be read off from tables of the Normal distribution to a fair degree of approximation. Taking the binomial as an example, suppose we need the probability of obtaining 20 or more 'successes' in 50 trials when the probability per trial is 0.32. We have $n=50$, $\pi=0.32$, so that, using the results above,

$$\begin{aligned}\text{Mean count} &= 50 \times 0.32 = 16 \\ \text{Variance} &= 50 \times 0.32 \times 0.68 = 10.88 \\ \text{Standard deviation} &= \sqrt{10.88} = 3.298.\end{aligned}$$

A count of 20 thus exceeds the mean by 4, and this amounts to $4/3.298 = 1.213$ standard deviations; the probability of exceeding this in a Normal distribution is 0.113.

The true answer to our problem (using the formula above) is 0.145, so that the Normal approximation seems to be rather poor. It can be improved by reducing the deviation from the mean by $\frac{1}{2}$, from 4 to $3\frac{1}{2}$, and this is equivalent to $3.5/3.298 = 1.061$ standard deviations. The Normal probability corresponding to this is 0.144, in excellent agreement with the true value. This adjustment is called a *continuity correction* and its motivation can be appreciated from fig 1 which shows part of the binomial histogram and the approximating Normal curve with the same mean and standard deviation. The probability we require is the total area of the rectangles centred at 20, 21, 22, . . .; to obtain an approximation to this from the area under the curve, we need to start at the left hand side of the 20 rectangle, that is at $19\frac{1}{2}$.

Almost all the standard methods of analysing counted data utilise the Normal approximation to the binomial and Poisson distributions and it might seem that the methods used for continuous data which I have discussed in previous articles could be carried across without change. This is not quite true for three reasons. First, the binomial and Poisson distributions have only one parameter (the probability π for the binomial, the mean μ for the Poisson), whereas the

Figure 1 A binomial histogram and a Normal curve with the same mean and standard deviation.

Figure 2 Simple regression with a proportion as y-variate.

Normal distribution has two, the mean μ and the standard deviation σ . Once the mean of one of the discrete distributions has been estimated (using the sample mean), it is not necessary to make a separate estimate of the standard deviation or variance. Significance tests and confidence intervals can this be calculated using the Normal distribution rather than Student's t —in technical jargon, we can assume an infinite number of degrees of freedom.

The second point to notice is that for both the binomial and Poisson distributions the variance changes with the mean. It is thus often not reasonable to make the assumption that is commonly used with continuous data, that the variances of all the data items are the same. A more important point arises in a regression context. Suppose that the y-variate of a simple regression is a counted proportion (p , say) so that the sample data consist of a set of such proportions each with an associated x-value. A model of the form

$$E(p) = \alpha + \beta x$$

may then be unreasonable if the range of the x-values is at all wide, since it may predict proportions which are less than 0 or greater than 1 (fig 2). In the same way, a simple regression with a y-variate which has a Poisson distribution may predict negative means. One solution to this problem consists in transforming the data before analysis; I hope to treat this topic in a subsequent note.

Significance tests for proportions

The simplest problem that arises with counted data is the comparison of a counted proportion with a theoretical value—this does not occur very often in practice but has some place in the context of mendelian genetics. Suppose for example that in a large population theory leads us to expect 25% of affected individuals. In a random sample of size 20 we observe eight affected individuals; does this differ significantly from expectation? This problem can be solved using the method outlined above. On the null hypothesis we have

$$\begin{aligned}\text{Mean number affected} &= 20 \times 0.25 = 5 \\ \text{Variance} &= 20 \times 0.25 \times 0.75 = 3.75 \\ \text{Standard deviation} &= \sqrt{3.75} = 1.936.\end{aligned}$$

The observed deviation from the mean is $(8-5)=3$, so, bringing in a continuity correction, we need the probability in a Normal distribution of exceeding $2.5/1.936=1.291$ standard deviations, presumably in either direction; the answer is 0.197, and this is the two tailed significance level that we require.

Next comes the commoner problem of comparing two proportions. Suppose for example that we have two groups of patients, one treated and one as a control. Allocation of the patients to the groups has been random without pairing or matching, so that the two samples are independent. In the treated group, 18 patients out of 25 improve, in the controls nine out of 20 improve; is there evidence that the two groups differ?

The response rates in the two groups are $18/25=0.72$ in the treated group and $9/20=0.45$ in the controls. We are concerned with the difference between these rates which is equal to $0.72-0.45=0.27$. On the null hypothesis we are looking at two samples from a single binomial population with a response probability which can be estimated by the overall response rate of $(18+9)/(25+20)=27/45=0.60$. At this point we need a slight modification to the earlier results for the binomial distribution. I have quoted the mean and standard deviation of this distribution as $n\pi$ and $\sqrt{n\pi(1-\pi)}$, but these formulae apply to the actual counts. Here we are dealing with the rates obtained by dividing the counts by n ; the mean and standard deviation must both be divided by n ; giving π and $\sqrt{\pi(1-\pi)/n}$ respectively—the variance is just the square of the last expression. In our example, π is estimated at 0.60, so that $\pi(1-\pi)$ is $0.60 \times 0.40 = 0.24$. The sample sizes are 25 and 20, so that the variances of the sample rates are $0.24/25$ and $0.24/20$. At this point we appeal to our one piece of theory which tells us that the variance of the difference between the rates is the sum of the variances. In our example this gives the variance of the difference as

$$\frac{0.24}{25} + \frac{0.24}{20} = 0.0216$$

and the standard error of the difference is the square root of this, 0.147. Finally, then, our observed difference of 0.27 amounts to $0.27/0.147=1.84$ standard errors and the approximate significance level, using the Normal distribution, is 0.066.

The 2×2 table

It is a weakness in the above analysis that we have not supplied a continuity correction, so that the Normal approximation we have used may not be very good—indeed, it is not too clear just how the correction should be applied. We can however present the data for our problem in a rather different way, as shown in table 1. As before, we have overall 27 responders out of 45, and if the treatment was ineffective this would lead us to expect $(27/45) \times 25 = 15$ responders in the treated group. We have actually observed 18 responders in this group—is this surprising?

We can in fact produce a second table showing the numbers in each of the four categories which the null hypothesis would lead us to expect, by making the proportion of responders in each group equal to that in the total line (table 2). We need to judge whether the sizes of the four discrepancies between observed and expected numbers are too big to be due to chance. In doing this, it is clear in general terms that we should take the numbers expected into account; a discrepancy of 5 is trivial if 100 are expected, noticeable if 20 are expected, important if 8 are expected. It turns out that the quantity to calculate is

$$\text{Sum of } \frac{(\text{discrepancy})^2}{\text{expected}}$$

This quantity is always known as χ^2 (Greek chi-squared—the ‘ch’ is hard as in loch). In our example, the discrepancies are (table 3)

$$\chi^2 = \frac{9}{15} + \frac{9}{10} + \frac{9}{12} + \frac{9}{8} = 3.375.$$

Tables of χ^2 are widely available, but in the present context we do not need to consult them. Instead we simply observe that the square root of 3.375 is just 1.84, exactly the Normal deviate that we obtained in our previous analysis. All we seem to have done is to produce the identical result by a more tortuous path!

There are two reasons for pursuing this seemingly indirect type of analysis. First, it shows us how to apply a continuity correction. Just as before, we reduce the discrepancies by $1/2$, from 3 to $2\frac{1}{2}$, before calculating χ^2 . This gives

$$\chi^2 = \left(\frac{1}{2}\right)^2 \left(\frac{1}{15} + \frac{1}{10} + \frac{1}{12} + \frac{1}{8}\right) = 2.344$$

The square root of this is 1.53, corresponding to a two tailed Normal probability of 0.126. With the rather small numbers involved, the correction has made an appreciable difference.

Table 1 Observed values

	Responders	Non-responders	Total
Treated	18	7	25
Controls	9	11	20
Total	27	18	45

Table 2 Expected values

	Responders	Non-responders	Total
Treated	15	10	25
Controls	12	8	20
Total	27	18	45

Table 3 Discrepancies

	Responders	Non-responders	Total
Treated	+3	-3	0
Controls	-3	+3	0
Total	0	0	0

Notice, by the way, that the apparently substantial difference in response rates between the treated and control groups might well have been due to chance. In general, studies with a yes/no response are liable to require much larger numbers for a conclusive result than those whose response is a continuous measurement. It is usually not a very good idea to summarise this latter type of study by recording simply the numbers of cases whose measurements do and do not exceed some fixed level such as the 'upper limit of normal'.

Estimation from 2×2 tables

Our discussion so far has concentrated on significance testing, but we have seen in other contexts that it is usually far more interesting to use data to estimate some quantity of interest. We started by considering the difference between the response rates in the treated and control groups, 18/25 and 12/20 respectively, and we showed that an estimate of this difference was 0.27 with a standard error of 0.147. Using a Normal approximation, we get 95% confidence limits for the true difference of $0.27 \pm 1.96 \times 0.147$, or -0.02 to $+0.56$.

An alternative approach is often more appropriate. It may not be realistic to assume that the difference between treated and control rates will be 0.27 in all circumstances—what if the rate in the controls were only 0.25, for example? A different measure of the treatment effect is based on the *odds* on recovery in the two groups. In the treated group we observed 18 responders and seven non-responders; the odds on recovery in this group are therefore $18/7 = 2.57$ to 1. In the controls the odds on recovery are $9/11 = 0.818$ to 1 (a betting man might prefer to phrase this as 11 to 9 against). We can measure the size of the treatment effect by calculating the *odds ratio*, $2.57/0.818 = 3.14$ —it appears that the treatment multiplies a patient's odds on recovery by more than threefold.

The null value of the odds ratio (the value when the treatment is ineffective) is 1.0, and the ratio cannot go below zero. In addition, a ratio of 2 in favour of the control is the same as a ratio of $\frac{1}{2}$ in favour of the treatment. All this makes it more convenient from a mathematical viewpoint to deal with the logarithm of the odds ratio—this has a null value of 0 and is unlimited in both directions. The relevant formulas are simplified if 'natural' logarithms to the base e are used. Notice that the logarithm of the odds ratio can be regarded as a difference. It is in fact the difference between the log odds for the treated group and that for the controls.

The odds ratio in our 2×2 table is 3.14 and the natural logarithm of this is 1.144. This quantity is approximately Normally distributed with a standard deviation obtained by adding up the reciprocals of the numbers in the table and taking the square root. Here we get a standard deviation given by

$$\sqrt{\frac{1}{18} + \frac{1}{7} + \frac{1}{9} + \frac{1}{11}} = 0.633$$

A 95% confidence interval for the log odds ratio is thus given by $1.144 \pm 1.96 \times 0.633 = -0.097$ to 2.385 . Taking 'natural' antilogs (exponentials) the limits for the odds ratio itself are from -0.91 to 10.86 . It appears that the treatment is unlikely to be appreciably worse than the control and maybe considerably better. The confidence interval includes 1.0 so that the table does not attain the magic 5% level of significance. This agrees with the conclusion based on the difference between the rates, though the two tests do not give identical answers.

The use of log odds as a method of handling counted proportions is of considerable importance. We shall meet it again in a regression context in a subsequent note. Coincidentally, I also utilised a version of the same transformation when discussing confidence limits for a correlation coefficient.