Occasional article

Statistics on microcomputers

A non-algebraic guide to the appropriate use of statistics packages in biomedical research and pathology laboratory practice

5 Analysis of categorical data

R A BROWN, J SWANSON BECK*

*From the Departments of Mathematical Sciences and *Pathology, University of Dundee, Scotland*

Analysis of counts and proportions

The previous articles in this series have been concerned with the analysis of data measured on an objective scale ("continuous" data). There are, however, many occasions when some or all of the data are collected as *nominal variables*—that is, something which may fall into one of a set of named categories and is not ordinarily measured—or as *ordinal data*—that is, a feature which can be ranked subjectively by the observer but not measured objectively, as, for example, the severity of an illness. Thus nominal variables arise when each patient may be assigned to one of several mutually exclusive categories distinguished by names—for example, the sex of a patient is either male or female, and there is no meaningful way of ordering such categories. In contrast, an ordinal variable entails a set of mutually exclusive categories which can be sensibly ordered such as the staging or grading of a neoplasm into categories of increasing extent of disease or greater apparent aggressiveness.

Nominal data may be coded—for example, male = 1, female = 2—but the coding is an arbitrary label as the numerical values chosen as codes have no intrinsic meaning. Ordinal data are commonly coded with the simplest series of ascending integers—for example, none = 0, mild = 1, moderate = 2, severe = 3—as the integers 0, 1, 2, 3 preserve the relative ordering of the categories, but any other ascending series could be used such as none = 0, mild = 6, moderate = 20 and severe = 500. It is essential to remember that the actual increments in the

ordinal series (none, mild, moderate, severe) are probably unequal and so when using a coding such as 0, 1, 2, 3 it would be quite improper to assume that "2" was twice as far up the grading scale as "1". It is for this reason that descriptive statistics such as the sample mean and standard deviation should not be used to summarise either nominal or ordinal data (the average sex of a sample of subjects is meaningless).

There are two special situations in which pathologists commonly encounter data as counts or proportions. The first concerns measurements using radioactive tracers, such as uptake of ³H-TdR as an indicator of DNA synthesis. Here the counts are expressed as cpm and are often large. The second arises when particular features are counted directly as in enumeration of blood cells or point counting in histometric methods. In both cases relatively large numbers are counted and there will be as many counts or proportions as there are specimens; consequently it is often possible to treat the data as if they were true measurements, but care must be taken to ensure that the data conform to the basic assumptions of the chosen statistical methods of analysis which were devised for measured data.

ESTIMATING A SINGLE PROPORTION

It is very unlikely that the sole purpose of an investigation would be the estimation of a single proportion—for example, the proportion of leprosy cases showing the lepromatous form of the disease in a particular region—but it is instructive to look at this example to appreciate some of the limitations of estimation when dealing with proportions. The investigation would proceed by taking a random sample of cases and classifying each one as either lepromatous or not. Thus

of 250 cases 175 might be lepromatous and 75 might not so that the estimate (p) of the population proportion would be 175/250, or 0.7. To calculate the 95% confidence interval (as a measure of the precision of this estimate) it is necessary to obtain the standard error of the estimate by substitution of the value of the estimate in the formula:

$$\text{SE of a proportion} = \sqrt{(p(1 - p)/n)}$$

and so in this example the standard error is $\sqrt{(0.7 \times 0.3/250)}$, or 0.029. The critical value used in calculating the 95% confidence interval is 1.96. By using the formula:

$$\text{estimate} \pm \text{critical value} \times \text{SE of estimate},$$

it is clear that the 95% confidence interval for the proportion of patients with leprosy showing the lepromatous form of the disease stretches from 0.643 up to 0.757.

This method of calculation of the 95% confidence limits is valid only when the sample size is reasonably large. For very small sample sizes, say 30 or less, the confidence interval can be read directly from a chart given in Neave's tables.¹ For example, if the estimated value of the proportion is 0.7 based on a sample of 10 then Neave's tables give 95% confidence limits stretching from 0.35 up to 0.88. Under these conditions misuse of the formula would give the interval as 0.42 up to 0.98. Notably, the correct interval is not symmetrical about the estimated value of 0.7, but both intervals are so wide as to be virtually useless. This example emphasises how little reliance can be placed on estimates of proportions based on small samples of cases.

One very simple way of assessing the sample size needed to achieve a given precision is to specify the width of confidence interval which is acceptable before the start of the investigation. For example, if the investigator has prior knowledge from reports of similar investigations that the proportion is likely to be around 0.4 and a 95% confidence interval of width 0.1 is required, as the width is twice the critical value multiplied by the standard error of the estimate then:

$$\text{width} = 2 \times 1.96 \times \sqrt{(p(1 - p)/n)}.$$

Introducing the prior estimate $p = 0.4$ into this formula results in the equation:

$$0.1 = 3.92 \times \sqrt{(0.24/n)}$$

for which the solution is $n = 369$. This is quite a lot larger than might be anticipated by an inexperienced investigator and shows how much effort is needed to get even moderately precise estimates of proportions.

This calculation requires prior knowledge of the probable size of the proportion which the investigation is supposed to estimate; if it is not possible to make a reasonable guess at this value then $p = 0.5$ should be substituted in the formula relating width to

sample size since the standard error takes its maximum value when $p = 0.5$. This gives the simple result:

$$n = \delta (1.96/\text{width})^2$$

for calculating the sample size required for a 95% confidence interval of a given width. This approach will err on the side of caution but it will avoid unwarranted optimism in planning investigations.

COMPARISON OF TWO PROPORTIONS

It is valid to compare the proportions showing a given character only when the samples have been selected randomly from two populations. In general, the sample sizes need not be equal but it is best if they are not too dissimilar; bearing in mind the imprecision inherent in estimates of proportions from small samples it would not be sensible to sample 200 subjects from one population and only 20 from the other if a precise estimate of the difference between the proportions was required.

Table 1 gives the results of a comparison of the prevalence of the Kml immunoglobulin allotype in patients with pulmonary tuberculosis and healthy controls in Indonesia. The Kml allotype is present in 32% of the patients and in 57% of the controls. It is more important to determine a confidence interval for the difference between the population proportions than to perform a significance test to determine whether the data provide evidence of a significant difference between the population proportions. The first of these procedures is obviously more informative as it is concerned with practical significance, the second may lead us into the trap of mistakenly emphasising statistical significance.

The χ^2 is a significance testing procedure which is appropriate in various situations where categorical data must be analysed.² In this particular example it provides a test of the hypothesis that the proportion of subjects with the Kml allotype is the same in patients and controls. The test is based on a comparison of what would be expected if this were the case with what is actually observed. As a working hypothesis, suppose that the proportion of subjects with the Kml allotype is the same in both populations, then the information in the two samples may be combined to give 111/253, or 0.4387 as the best estimate of this proportion. On

Table 1 Comparison between prevalence of Kml immunoglobulin allotype in Indonesian patients with pulmonary tuberculosis and healthy controls

	Kml positive	Kml negative	Total
Patients	39	82	121
Controls	72	60	132
Total	111	142	253

```

MTB > NAME C1 'KM1 +'
MTB > NAME C2 'KM1 -'
MTB > CHISQ 'KM1 +' 'KM1 -'

Expected counts are printed below observed counts

```

	KM1 +	KM1 -	Total
1	39	82	121
	53.09	67.91	
2	72	60	132
	57.91	74.09	
Total	111	142	253
ChiSq =	3.738 +	2.922 +	
	3.427 +	2.679 =	12.765
df = 1			
MTB > CDF 12.765;			
SUBC > CHISQ 1.			
	12.7650	0.9996	
MTB > NOTE: So the p-value is 1 - 0.9996 = 0.0004			

Fig 1 Minitab output for the χ^2 analysis of incidence of the Kml immunoglobulin allotype (table 1). CDF command is used to determine the p value of the χ^2 statistic.

average 43.87% of subjects in each sample would be expected to bear the Kml allotype—that is, 53.08 patients and 57.91 controls—and by similar reasoning 67.92 patients and 74.09 controls would not be expected to carry the Kml allotype. These are the “expected counts” and they should be interpreted as long run averages; it is clearly impossible to observe fractions of one subject. The overall discrepancy between the observed and expected counts is known as the χ^2 statistic and is the sum over the four “cells” of the table of the values of:

$$(\text{observed count} - \text{expected count})^2 / \text{expected count}.$$

Regrettably, both Minitab and Statgraphics concentrate on significance testing and ignore estimation. The Minitab output for the χ^2 test is displayed in fig 1. The statgraphics output is less satisfactory as it includes a considerable number of extra statistics which are irrelevant to this particular situation while omitting the expected counts. The χ^2 value of 12.765, which has a p value of 0.0004, indicates that there is extremely strong evidence that the proportion of subjects bearing the Kml allotype is not the same in the population of patients with pulmonary tuberculosis as it is in the population of healthy controls.

The next and more important step is to calculate a

Table 2 Hypothetical results of much smaller investigation of prevalence of Kml allotype

	Kml positive	Kml negative	Total
Patients	2	5	7
Controls	7	6	13
Total	9	11	20

```

MTB > CHISQUARE C1 C2

Expected counts are printed below observed counts

```

	C1	C2	Total
1	2	5	7
	3.15	3.85	
2	7	6	13
	5.85	7.15	
Total	9	11	20
ChiSq =	0.420 +	0.344 +	
	0.226 +	0.185 =	1.174
df = 1			
2 cells with expected counts less than 5.0			

Fig 2 Minitab output for the χ^2 analysis of the data shown in table 2. The warning about expected count less than 5 indicates that the χ^2 test may be misleading.

confidence interval so as to interpret this result. Unfortunately, neither package does this, but it is quite easy to calculate the 95% confidence interval using the formulae given in the appendix. The result of the investigation may be summarised as follows: the proportion of patients with pulmonary tuberculosis bearing the Kml allotype is estimated to be 32.3% with a standard error of 4.2%, and for normal controls the proportion is 54.5% with a standard error of 4.3%; this difference is highly significant ($\chi^2 = 12.765$, $p = 0.0004$), and the 95% confidence interval for the difference indicates that the proportion of patients bearing the Kml allotype is likely to be between 10.4 and 34.2 percentage points lower than the corresponding proportion for healthy controls.

If the sample sizes are very small then a problem may arise. Suppose that we try to analyse the hypothetical data in table 2 which correspond to a drastically scaled down version of the previous investigation. Minitab produces the display shown in fig 2: note the warning message “2 cells with expected counts less than 5.0”, which indicates that the sample sizes are too small for the χ^2 test to be reliable. The general rule is that the test is reliable only when all expected counts are greater than 5. (Note that the rule refers to expected counts, so it is quite in order to apply the test when some observed counts are less than 5, provided all expected counts exceed 5). There is a test known as Fisher’s exact test² which can be used in this situation and details of which are given in the appendix. Statgraphics automatically performs the exact test whenever the χ^2 test is likely to be invalid but manual calculation will be necessary for users of the Minitab package.

The best advice that can be given here is that the investigator should avoid getting into this situation by keeping sample sizes reasonably large. When an investigation aims to compare proportions which are expected to be very small, expert advice should be sought as it is likely that extremely large sample sizes

Table 3 *Leprosin A skin tests in various subject groups*

	Test result		Total
	Positive	Negative	
Healthy controls (A)	30	20	50
Hospital contacts (B)	18	2	20
Household contacts (C)	45	33	78
Treated patients (D)	35	18	53
New (untreated) patients (E)	21	31	52
Total	149	104	253

will be needed to obtain worthwhile data and this may involve multicentre collaboration or data collection over a long period of time. In such circumstances diagnostic criteria, laboratory procedures, and record keeping systems must be carefully standardised and active quality control of the results (perhaps by means of periodic quality audits) will be essential.

COMPARISON OF SEVERAL PROPORTIONS

Comparison of several samples from different populations to detect and estimate differences between proportions is dealt with by a straightforward extension of the χ^2 method. The situation is analogous to that in which the analysis of variance is used for measured data, so a single significance test is first applied to all of the data to assess the evidence for any differences between the samples, and a follow

MTB > NAME C1 'POS.'				
MTB > NAME C2 'NEG.'				
MTB > CHISQ C1 C2				
Expected counts are printed below observed counts				
	POS.	NEG.	Total	
1	30	20	50	
	29.45	20.55		
2	18	2	20	
	11.78	8.22		
3	45	33	78	
	45.94	32.06		
4	35	18	53	
	31.21	21.79		
5	21	31	52	
	30.62	21.38		
Total	149	104	253	
ChiSq =	0.010	+	0.015	+
	3.286	+	4.708	+
	0.019	+	0.027	+
	0.459	+	0.658	+
	3.025	+	4.334	=
				16.541
df = 4				
MTB > CDF 16.541;				
SUBC > CHISQ 4.				
16.5410			0.9976	
NOTE: So the p-value of the test is 1 - 0.9976 = 0.0024				

Fig 3 Minitab output for the χ^2 analysis of the Leprosin A skin test data (table 3). The highly significant χ^2 value indicates that some groups differ in responsiveness.

Fig 4 Multiple comparison confidence intervals for the differences between proportions of positive skin tests for all pairs of groups of subjects. Most intervals include zero, indicating that the corresponding pair of groups cannot be regarded as differing in responsiveness. To indicate the groups being compared the lines representing the confidence intervals are labelled with letters assigned to the groups in table 3.

up investigation determines which proportions are different: the problem of multiple comparison analogous to that seen with measured data (article 3) arises in the follow up stage.

Table 3 shows the results of Leprosin A skin testing in healthy controls, two groups of patients with leprosy, and two groups of contacts of leprosy victims. The relevant questions here are concerned with whether there are differences between the responsiveness of subjects in the various groups to Leprosin A. The Minitab output is shown in fig 3, with expected counts calculated on the same principle as before; there are a total of 149 positive responses in 253 subjects so if there were no differences between the rates of positive response in the various subpopulations the best estimate of the common rate of positive responses is 149/253, or 0.589, so 58.9% of each sample might be expected to show a positive skin test and 41.1% to show a negative response. The χ^2 statistic is calculated by summing:

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

for all 10 cells of the table, giving the value 16.54 which has a p value of 0.0024. This indicates that the data provide extremely strong evidence of differences between some of the population proportions. To follow up this finding we can calculate a series of multiple confidence intervals for the difference between positive response rates. The details of this procedure are given in the appendix. Confidence intervals for the differences between positive response rates for each pair of groups are displayed in fig 4, from which it is apparent that the main difference is between hospital contacts of patients with leprosy who have the highest rate of

Statistics on microcomputers: analysis of categorical data

Table 4 Relation between BCG vaccination status and cavitation of lung in Indonesian patients with pulmonary tuberculosis

Extent of cavitation	Vaccinated	Not vaccinated	Total
0	4	31	35
1	4	9	13
2	13	18	31
3	3	27	30
Total	24	85	109

positive response and the group of new patients who have the lowest positive response rate; the remaining groups fall in an intermediate position with no detectable differences between their response rates.

Multiple category responses

The χ^2 test may also be applied to situations where there are more than two categories of response. Table 4 shows the results of an investigation of the relation between BCG vaccination status and extent of cavitation of the lungs (graded on a four point ordinary scale) in Indonesian patients with pulmonary tuberculosis. The hypothesis that the underlying distribution of the extent of cavitation does not depend on the vaccination status of the patients is tested by the χ^2 method. A significant result will mean that extent of cavitation does depend on vaccination status. Expected values and the χ^2 statistic are calculated in the same way as in the previous examples, and the Minitab output (fig 5a) indicates a problem with small sample sizes as one expected count is considerably less than 5. This may be dealt with by combining the first two cavitation categories (cavitation index none or one) to obtain higher expected counts. This leads to the analysis shown in fig 5b, from which we conclude that extent of cavitation is not independent of BCG vaccination status. The main difference between the distributions of extent of cavitation would seem to be a relative excess of cavitation grade 3 (cavities in two lung zones) among the vaccinated patients. Interpretation of this type of investigation is necessarily more complex as differences in the overall pattern of the multiple category grading variable are being looked for and such differences cannot be summarised in terms of one or two simply interpretable numbers.

Point or structure counts, radioactive counts

Where an investigation entails direct counts of structures in tissue sections or estimation of proportionate areas or volumes occupied by various features, then although the end result will be expressed as a proportion, the method of analysis to be used is quite different from that described in the earlier sections of this article. The same will be true of investigations which use radioactive cpm as an indirect method of measur-

MTB > CHISQ C1 C2			
Expected counts are printed below observed counts			
	C1	C2	Total
1	4 7.71	31 27.29	35
2	4 2.86	9 10.14	13
3	13 6.83	18 24.17	31
4	3 6.61	27 23.39	30
Total	24	85	109
ChiSq =	1.783 + 0.452 + 5.585 + 1.968 +	0.503 + 0.128 + 1.577 + 0.556 =	12.551
df = 3			
1 cells with expected counts less than 5.0			

(a)

MTB > CHISQ C1 C2			
Expected counts are printed below observed counts			
	C1	C2	Total
1	8 10.57	40 37.43	48
2	13 6.83	18 24.17	31
3	3 6.61	27 23.39	30
Total	24	85	109
ChiSq =	0.624 + 5.585 + 1.968 +	0.176 + 1.577 + 0.556 =	10.486
df = 2			
MTB > CDF 10.486; SUBC > CHISQ 2. 10.4860 0.9947			
NOTE: So the p-value of the test is 1 - 0.9947 = 0.0053.			

(b)

Fig 5 (a) Minitab output for the data relating extent of lung cavitation to BCG vaccination status (table 4). One expected count is less than 5 so the χ^2 test could be misleading. (b) By combining results for adjacent cavitation grades (0, 1) the expected counts are raised to permit valid use of the χ^2 test.

ing some aspect of cellular activity, such as DNA synthesis. The essential difference is that each case will result in a separate proportion or count rather than contributing a small amount to a single proportion. In this case proportions and cpm can be treated as if they were measurements, and all of the methods appropriate to measured data may be used after suitable transformation of the data.

Proportions and cpm, however, are not true measurements and they do not conform to all the assumptions that the standard methods devised for continuous data require for their validity. In particular, the assumption of constant residual variability is not met. The variability in a set of proportions is related to the size of the proportions, thus small and large proportions have less scope for being variable because they cannot extend beyond zero or unity whereas the intermediate proportions have more room for variability. To nullify this effect proportions

should always be transformed by means of the angular transformation² (also known as the arc-sine transformation) before methods such as analysis of variance or regression analysis are applied.

From the theory of radiation physics it can be shown that radioactive counts conform to the Poisson distribution, not the normal distribution. One of the important consequences of this is that the variability of radioactive cpm will be directly proportional to their average.² In this situation the raw data must be transformed by taking square roots of counts before applying analysis of variance or regression methods.

Appendix

EXPECTED COUNTS AND DEGREES OF FREEDOM

There is a simple rule for calculating the expected counts, for each cell of the table:

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{grand total}}.$$

The degrees of freedom associated with the χ^2 statistic is equal to one less than the number of rows in the table multiplied by one less than the number of columns.

CONFIDENCE INTERVALS

The standard error of the estimate (p) of a proportion is $\sqrt{p(1-p)/n}$ and the confidence interval is:

$$\text{estimate} \pm \text{critical value} \times \text{SE of estimate}.$$

The critical value is 1.96 for a 95% interval, 2.58 for a 99% interval, and 2.81 for a 99.5% interval. These critical values are the upper 2.5%, 0.5%, and 0.25% points of the normal distribution.

If the estimates of two proportions are p_1 and p_2 based on samples of sizes n_1 and n_2 then the standard error of their difference is:

$$\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$$

and a confidence interval for the difference between the proportions is:

$$(p_1 - p_2) \pm \text{critical value} \times \text{SE of difference},$$

the same critical values being used.

MULTIPLE COMPARISON OF A SET OF PROPORTIONS

The confidence intervals shown in fig 4 were calculated by applying the formula for the confidence interval for the difference between two proportions to all possible pairs of proportions. To maintain an overall error rate of 5% in the complete set of comparisons the confidence coefficient of each interval must be $100\% - 5\% / (\text{number of comparisons})$. In the example there are 10 comparisons so the confidence coefficient must be 99.5% and each interval will be of the form:

$$\text{difference} \pm 2.81 \times \text{SE of difference}.$$

For example, the proportion of positive skin tests

0	7	7	1	6	7	2	5	7
9	4	13	8	5	13	7	6	13
9	11	20	9	11	20	9	11	20

Fig 6 The p value of Fisher's exact test is calculated by summing the probabilities of occurrence associated with these three tables (derived from table 2).

among the hospital controls is 0.9 and among the healthy controls it is 0.6. The difference is 0.3 with a standard error equal to $\sqrt{((0.6 \times 0.4/50) + (0.9 \times 0.1/20))}$, or 0.096, so the confidence interval is $0.3 \pm 2.81 \times 0.096$ or 0.3 ± 0.27 , (30% $\pm$ 27%). Note that a separate calculation of the standard error is needed for each pair of groups being compared.

FISHER'S EXACT TEST

The exact test calculates the p value associated with the observed table on the hypothesis that there is no difference between the corresponding population proportions. This is done by summing the probability of occurrence of the observed table and all tables which have the same marginal totals but are more extreme. In the example these are the tables shown in fig 6, which are obtained by putting 0, 1, and 2 in the upper top left corner (this then determines the remaining three entries since the row and column totals are fixed).

If O_1, O_2, O_3 and O_4 denote the values in the body of the table, R_1, R_2, C_1, C_2 and N the row totals, column totals, and grand total, respectively, then the probability associated with this particular table is:

$$\frac{R_1! \times R_2! \times C_1! \times C_2!}{N! \times O_1! \times O_2! \times O_3! \times O_4!}$$

where, for example $R_1!$ (read as " R_1 factorial") is the product of the integers from 1 up to R_1 (in this example $R_1 = 7$ and $7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$ or 5040). Note also that the numerical value of $0!$ (zero factorial) is 1.

The probabilities associated with the three tables in fig 6 are $(7! \times 13! \times 9! \times 11!)/(20! \times 7! \times 9! \times 4!)$, $(7! \times 13! \times 9! \times 11!)/(20! \times 1! \times 6! \times 8! \times 5!)$, and $(7! \times 13! \times 9! \times 11!)/(20! \times 2! \times 5! \times 7! \times 6!)$ or 0.004, 0.053, and 0.215, so the p value of the observed table is 0.272, leading to the conclusion that observed results provide no evidence of a difference between the proportions.

References

- 1 Neave HR. *Elementary statistics tables*. London: Allen and Unwin, 1981.
- 2 Armitage P. *Statistical methods in medical research*. Oxford: Blackwell, 1971.

Requests for reprints to: Professor J Swanson Beck, Department of Pathology, Ninewells Hospital and Medical School, PO Box 120, Dundee DD1 9SY, Scotland.