Analysing ordered categorical data from two independent samples

Anthony P Morton, Annette J Dobson

Medical journals now require point estimates and confidence intervals in addition to p values.[1] *Statistics with Confidence* provides advice on calculating confidence intervals for many types of data[2] but not for ordered categorical data. Nevertheless, as much as 20% of reported medical research produces data in ordered categories.[3] The Apgar score, the Glasgow coma scale, quality of life and severity of illness indexes, and several methods for staging tumours and grading the severity of disorders such as congestive heart failure are some of the many methods used in clinical medicine which produce data in ordered categories.[4] Moreover, current interest in quality assurance in hospitals is likely to result in an increase in the use of methods which produce data in ordered categories.[5] Finally, it is often useful to place continuous data in ordered categories as the resulting tabulations aid in their interpretation.

In this paper we describe estimators which are appropriate for comparing data in which there are two independent exposure groups and an ordered categorical outcome. Because they offer familiar, practically useful interpretations they should prove suitable to communicate information to clinicians and other health workers. These estimators are related to the rank sum test, which can be used to obtain significance levels. In addition, bootstrapping,[6,7] a computer based method, can be used to calculate confidence intervals.

Methods and worked example

EXAMPLE

A trial carried out by the Medical Research Council in 1948 to test the efficacy of streptomycin for treating pulmonary tuberculosis was one of the earliest randomised experiments in medicine.[8] The results of the study were as shown in table I.

TABLE I—*Results of randomised trial of streptomycin for tuberculosis: categories of outcome in treated and control groups*

Category	Outcome	Treatment	Control	Total
1	Death	4	14	18
2	Considerable deterioration	6	6	12
3	Moderate deterioration	5	12	17
4	No change	2	3	5
5	Moderate improvement	10	13	23
6	Considerable improvement	28	4	32
Total		55	52	107

RANK SUM TEST

The rank sum test is appropriate for comparing two independent groups of ordered categorical data.[3] This test may be performed by three slightly different methods which give the same result: Kendall's S test, the Mann-Whitney U test, and Wilcoxon's two sample test. Moses *et al*[3] and Leach[9] describe how this test is used to analyse data in ordered categories. For the data from the streptomycin trial the rank sum test gives a test statistic of 20·2, which is compared with the χ^2 distribution with one degree of freedom. The resulting p value is less than 0·001, indicating significantly different responses between the two groups. The method of ranking data with tied values, which occurs when they are in ordered categories, is described by Swinscow[10] and Conover.[11]

An alternative approach is to use a χ^2 test for trend in proportions.[12] Thus one could use the scores -3, -2, -1, 0, 1, and 2 for the categories "death" to "considerable improvement" and test whether the proportion of subjects in the treatment group increases in successively better outcome categories. By this method $\chi^2=17{\cdot}93$ with one degree of freedom, close to that calculated from the rank sum test, and so the conclusion is the same. With an appropriate choice of scores the rank and score methods are in fact equivalent.[12]

These are appropriate methods for testing the hypothesis that there are no differences in the response patterns between the two groups, but they do not provide estimates of the magnitude of the differences. This problem has been the subject of several recent studies.[13-16] A rank sum estimator (δ) and a generalisation of the odds ratio (α) can be used to obtain point estimates and confidence intervals for the extent to which the responses in one group are better than the other.

POINT ESTIMATORS

The rank sum estimator (δ)

The rank sum estimator (δ)[17] is calculated as follows: δ=(average rank for treatment group−average rank for control group)/(average number of observations per group).

The calculation of ranks proceeds as follows. Suppose that all observations from both groups are numbered from 1 to 107 so that observations in category 1 have the numbers from 1 to 18, those in category 2 from 19 to 30 and so on. The median of the numbers for category 1 is 9·5 and this is the rank for category 1 (more simply, the rank is the average of the smallest and largest numbers for the category). Similarly, the rank for category 2 is 24·5. Table II shows the ranks for each category. All observations in each category have the same rank.

The rank total for the treatment group is $(4\times9{\cdot}5)+(6\times24{\cdot}5)+(5\times39)+(2\times50)+(10\times64)+(28\times91{\cdot}5)=3682$ and its average is 3682/55=66·945. The rank total for the control group is $(14\times9{\cdot}5)+(6\times24{\cdot}5)+$

TABLE II—*Ranking of observations from streptomycin trial to derive a median for each category—the category rank*

Category	Observation number	Category rank
1	1-18	9·5
2	19-30	24·5
3	31-47	39
4	48-52	50
5	53-75	64
6	56-107	91·5

Key Centre in Strategic Management, Queensland University of Technology, Brisbane 4000, Queensland, Australia
Anthony P Morton, MD, *research associate*

Department of Statistics, University of Newcastle, Newcastle 2308, New South Wales, Australia
Annette J Dobson, PHD, *professor of biostatistics*

Correspondence to:
Dr Morton.

Br Med J 1990;301:971-3

BMJ: first published as 10.1136/bmj.301.6758.971 on 27 October 1990. Downloaded from http://www.bmj.com/ on 23 April 2024 by guest. Protected by copyright.

$(12\times39)+(3\times50)+(13\times64)+(4\times91\cdot5)=2096$ and its average is $2096/52=40\cdot308$. The average number in the groups is $(55+52)/2=53\cdot5$.

Therefore $\delta=(66\cdot945-40\cdot308)/53\cdot5=0\cdot498$.

This estimator is a measure of the difference between the two groups. If all the observations in the treatment group are larger than any in the control group, δ has a value of 1; and if all the control group observations are larger than the treatment group observations its value is -1; if they have similar rankings its value is 0. If the data are ranked from the largest to smallest instead of having the smallest first δ has the same value but its sign is reversed. In these respects it behaves like a correlation coefficient. It can, however, also be interpreted in terms of probabilities.

If the treatment group contains N_a observations and the control group has N_b observations, there are $N_a\times N_b$ ways in which a treatment group observation can be compared with a control group observation. Thus there are $(28\times13)+(28\times3)+\ldots+(6\times14)=1942$ comparisons where a treatment group result is better than a control group result and there are $(4\times10)+(4\times2)+\ldots+(6\times4)=518$ comparisons where a control group result is better than a treatment group result. In $(28\times4)+(10\times13)+\ldots+(4\times14)=400$ comparisons the treatment and control group results are in the same categories. The total number of comparisons is 2860 or 55×52 as expected. The probability of a treatment group result being better than a control group result is $1942/2860=0\cdot679$ and of a control group result being better than a treatment group result $518/2860=0\cdot181$. The difference in these probabilities is 0·498. This is the same result as that obtained using the rank formula for δ. Thus it is also a generalisation of the familiar difference in proportions for two outcome data to data in ordered categories. This estimator was originally described by Somers, who applied it to ordered data in contingency tables.[18] When there are two study groups δ and Somers's estimator are equivalent.

The generalised odds ratio (α)

Calculation of the generalised odds ratio (α) proceeds as follows.

Let P be the probability that a patient in the treatment group has a better outcome than a patient in the control group and Q be the probability that a patient in the treatment group has a worse outcome than a control. Then $P=0\cdot679$ and $Q=0\cdot181$. The generalised odds ratio (α), which has been described by Agresti,[19] is the ratio of these probabilities (P/Q). Thus $\alpha=0\cdot679/0\cdot181=3\cdot75$.

CONFIDENCE INTERVALS

The usual method for calculating 95% confidence intervals, which involves adding to and subtracting from the value of the point estimate 1·96 times the standard error, is described in detail in the appendices to chapter 2 of *Statistics with Confidence*.[2] As pointed out there, this method requires that the estimate of interest, or often its logarithm, has a normal sampling distribution. Formulas for calculating standard errors of δ and of the logarithm of α are available,[18 19] but they are complicated and can sometimes give inaccurate results even when sample sizes are large. The sampling distribution of δ and of the logarithm of α are often far from normal. For example, if δ is between about $-0\cdot5$ and 0·5 the sampling distribution is fairly symmetrical but for values nearer to -1 and 1 the distribution is quite skewed (like that of the correlation coefficient) so that confidence intervals calculated from standard errors are incorrect.

Bootstrapping provides a simple alternative method for calculating confidence limits.[6 7] One way of thinking about confidence intervals is to imagine that repeated samples have been taken from the same population and the required point estimate calculated for each sample; then the interval which contains the middle 95% of these estimates is an approximate 95% confidence interval. Since it is not usually possible to take repeated samples from the parent population, it seems natural to use the data values in the study sample for this purpose and this is essentially how bootstrapping works. It provides a general way for calculating approximate confidence intervals in many situations where the sampling distribution of the estimator is not normal or formulas for calculating standard errors are either complicated or unavailable. Efron and Tibshirani[6] and Noreen[7] have described the process in detail, and Morton and Dobson[20] have applied the method to measures of agreement which have non-normal sampling distributions.

The method can be illustrated using the data given in table I. To obtain a bootstrap sample random numbers are used to allocate the sample of 55 observations from the treatment group randomly to the various outcome categories. Similarly, the 52 observations in the control group are randomly allocated to the outcome categories. This sampling is done "with replacement" so that only the totals for each group are fixed and the numbers in the categories vary. The sampling is done independently for each group. It produces samples such as those shown in tables III-V.

Samples produced by bootstrapping—using random numbers to allocate the 55 observations in the treatment group and 52 in the control group to each category

TABLE III

Outcome	Treatment	Control	Total
Death	3	12	15
Considerable deterioration	10	3	13
Moderate deterioration	3	15	18
No change	4	5	9
Moderate improvement	5	10	15
Considerable improvement	30	7	37
Total	55	52	107

TABLE IV

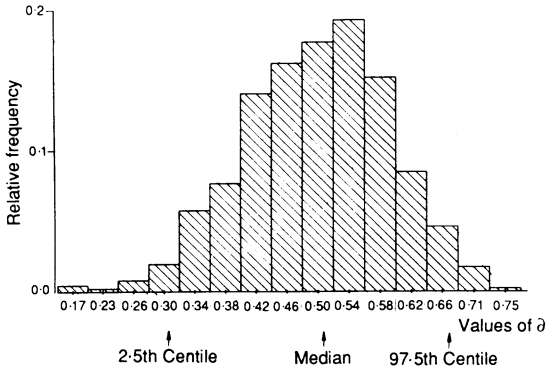
Outcome	Treatment	Control	Total
Death	2	10	12
Considerable deterioration	6	8	14
Moderate deterioration	7	11	18
No change	3	1	4
Moderate improvement	12	15	27
Considerable improvement	25	7	32
Total	55	52	107

TABLE V

Outcome	Treatment	Control	Total
Death	4	14	18
Considerable deterioration	2	6	8
Moderate deterioration	5	14	19
No change	2	3	5
Moderate improvement	10	10	20
Considerable improvement	32	5	37
Total	55	52	107

This process is repeated 1000 times.[6] In each case the estimators of interest—for example, δ and α—are calculated. For the data in table III $\delta=0\cdot40$ and $\alpha=2\cdot75$, for those in table IV $\delta=0\cdot41$ and $\alpha=2\cdot93$, and for those in table V $\delta=0\cdot59$ and $\alpha=5\cdot36$. The 1000 bootstrap estimates are then sorted into ascending order. The 95% bootstrap confidence intervals are then described by the 25th value from each end of the distribution; these are the 25th value and the $1000+1-25=976$th value. For the data in table I δ is 0·498 with α bootstrap 95% confidence interval of 0·30 to 0·66; α is 3·75 with a 95% confidence interval of 2·12

BMJ: first published as 10.1136/bmj.301.6758.971 on 27 October 1990. Downloaded from http://www.bmj.com/ on 23 April 2024 by guest. Protected by copyright.



Distribution of 1000 bootstrap samples for δ showing how confidence intervals are derived from 25th and 976th values

to 6·92. The distribution of 1000 bootstrap samples for δ is shown in the figure.

Discussion

Sometimes ordered categorical data are treated as if they were nominal, and a conventional χ^2 test is performed. This may lack power because it does not use the information available in the ordering of the data values. In addition, readily interpretable point estimates cannot be obtained. Occasionally ordered data are treated as if they were continuous. Although the resulting t test can be expected to provide a reasonable approximation for a p value,[3] the differences between the means or medians of the two samples are meaningless point estimates. The use of α and δ obviates these problems.

As well as providing a measure of the difference in the mean ranks of the two groups on a scale ranging from −1 to 1, δ describes the probability that an observation in the treatment group will be larger than an observation in the control group minus the probability that an observation in the control group will be larger than an observation in the treatment group. It thus generalises the difference in two independent proportions to data which are in ordered categories. In addition, α is a measure of the ratio of these probabilities and it thus generalises the odds ratio to data which are in ordered categories. Finally, these estimators correspond to the rank sum test and their confidence intervals agree well with that test—that is, when the significance level of the test is less than 0·05 the confidence interval for δ usually does not include 0 and the confidence interval for α usually does not include 1.

Halperin *et al* describe δ in a slightly different manner which is closely related to the Mann-Whitney test.[16] Their estimator (MW) is P+T/2 where P is the probability that a treatment group result will be greater than a control group result and T is the probability that the two group results will be tied. Since the total number of ways that the observations in the two groups can be compared is P+Q+T (where Q is the probability that a control group result will be greater than a treatment group result), MW and δ are related by MW=(δ+1)/2. Thus when there is no difference between the two groups δ will be 0 and MW will be 0·5. Although the two estimators are equivalent, MW may be useful—for example, in a clinical trial to give the probability that a patient would do better on the new treatment.

There may be concern that a computer intensive method such as bootstrapping could deter many potential users. Clinicians now use computers freely, however, and many are prepared to submit their data to analysis by one of the large statistical packages that are readily available. In some cases these data may be analysed by methods which may not be the most appropriate. The bootstrap method has some advantages in that it simulates the idea of confidence intervals (making inferences from repeated samples from a population) while using the observed data values to perform this process. Moreover, it does not depend on the data having a particular distribution or structure, thus making it relatively safe to use. A possible disadvantage is that repeating the analysis may produce a slightly different result each time; in most cases, however, the difference will be negligible. Finally, the use of two correction factors will usually improve the bootstrap approximation.[6] These are mathematically complex but can be incorporated into a computer program without difficulty.

1 Berry G. Statistical significance and confidence intervals. *Med J Aust* 1986;**144**:618-9.
2 Gardner M, Altman D. *Statistics with confidence*. London: British Medical Journal, 1989:1-79.
3 Moses L, Emerson J, Hosseini H. Analysing data from ordered categories. *N Engl J Med* 1984;**311**:442-8.
4 Feinstein A. *Clinimetrics*. New Haven: Yale University Press, 1987:225-44.
5 Donabedian A. The quality of care; how can it be assessed. *JAMA* 1988;**260**:1743-8.
6 Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science* 1980;**1**:54-77.
7 Noreen E. *Computer intensive methods for testing hypotheses*. New York: Wiley, 1989:63-81.
8 Bland M. *An introduction to medical statistics*. London: Oxford Medical Publications, 1987:241.
9 Leach C. *Introduction to statistics*. Chichester: Wiley, 1977:76-8.
10 Swinscow T. *Statistics at square one*. London: British Medical Journal, 1980;60, 74.
11 Conover W. *Practical nonparametric statistics*. New York: Wiley, 1980:219-33.
12 Armitage P, Berry G. *Statistical methods in medical research*. 2nd ed. Oxford: Blackwell Scientific, 1987:372-4.
13 Hochberg Y. On the variance estimate of a Wilcoxon-Mann-Whitney statistic for grouped ordered data. *Communications in Statistics—Theory and Methods* 1981;**A10**:1719-32.
14 Simonoff J, Hochberg Y, Reiser B. Alternative estimation procedures for Pr(X<Y) in categorized data. *Biometrics* 1986;**42**:895-908.
15 Halperin M, Gilbert P, Lachin J. Distribution-free confidence intervals for $Pr(X_1<X_2)$. *Biometrics* 1987;**43**:71-80.
16 Halperin M, Hamdy M, Thall P. Distribution-free confidence intervals for a parameter of Wilcoxon-Mann-Whitney type for ordered categories and progressive sensoring. *Biometrics* 1989;**45**:509-21.
17 Welkowitz J, Ewen R, Cohen J. *Introductory statistics for the behavioural sciences*. New York: Academic Press, 1976:266-7.
18 Reynolds H. *The analysis of cross classifications*. New York: The Free Press, 1977:85-9.
19 Agresti A. Generalised odds ratios for ordinal data. *Biometrics* 1980;**36**:59-67.
20 Morton A, Dobson J. Assessing agreement. *Med J Aust* 1989;**150**:384-7.

(Accepted 20 July 1990)

ANY QUESTIONS

What is the cause and possible treatment of brittle nails and ridging in the nails of a man of 70 who is otherwise in excellent health?

Brittle nails at any age in a person in excellent health are usually due to external causes. Excessive exposure to solvents, detergents, and soaps may weaken the nails, especially the ends, producing splitting and softening of the whole nail but particularly the ends, and a condition known as lamellar onycholysis may develop. Ridging of the fingernails may be transverse or longitudinal. Transverse ridging is known as Beau's lines and develops as a result of a systemic upset, either a severe infection or an operation, in which nail growth is temporarily impaired, so that a groove and ridge develop, often across all the fingernails. Multiple longitudinal ridging is due to an irregularity of growth of the nail under the nailfold and is constitutional and often more pronounced in the elderly, but a single groove running the length of one nail may be produced by a tumour under the nailfold.

Brittle nails will regain their normal strength if the causative external agents are avoided, so this means less wet work, less detergent in the washing up water, and avoiding contact with solvents. If the brittleness persists lotions containing formalin applied to the nail daily for a few weeks will harden the nails, but care should be taken to stop such lotions getting on to the skin to avoid an irritant and even an allergic contact eczema.

Beau's lines will grow out with the normal growth of the nail, though if the groove is deep enough the nail may break off. Longitudinal ridging is not amenable to treatment, though its severity will vary spontaneously from time to time.—ALAN B SHRANK, *consultant dermatologist, Shrewsbury*

BMJ: first published as 10.1136/bmj.301.6758.971 on 27 October 1990. Downloaded from http://www.bmj.com/ on 23 April 2024 by guest. Protected by copyright.