Statistical Applications in Orthodontics

Part II. Confidence Intervals for Proportions and their Differences

R.G. NEWCOMBE, PH.D., C.STAT., HON.M.F.P.H.M.

Statistical Advisor to the Journal of Orthodontics

In a previous article (Newcombe, 2000) I outlined the reasons why confidence intervals (CIs) are regarded as a helpful way to express research findings, and how confidence intervals for means and their differences are calculated and interpreted. In this second article I will deal with proportions and differences between proportions.

Many variables in health related research are *binary*, having two possible values, e.g. gender (male or female), outcome (alive or dead), response to treatment (positive or negative), dental arch (maxillary or mandibular), etc. When a binary variable is recorded for each individual or unit in a sample, usually we report the *proportion* that falls in a particular group, often expressed as a *percentage*. For example, 20 out of a series of 30 orthodontic patients were female (Heasman *et al.*, 1998); the proportion here is 0·667 or 66·7 per cent. The remaining 10 subjects were male. These made up 100–66·7 = 33·3 per cent of the series. In a laboratory study (Sargison *et al.*, 1999), 19 (63·3 per cent) out of 30 bonds involving etched specimens failed at the enamel-cement interface (ECI). The remaining 11 (36·7 per cent) failed at the cement-bracket interface (CBI).

Though the calculations described below are quite practicable using an electronic calculator, it helps to have computer software available. Hitherto, widely available statistical software has provided little to assist the user in this area. The second edition of the Confidence Interval Analysis software accompanying the BMJ booklet *Statistics with Confidence* (Altman *et al.*, 2000) contains programmes to perform the calculations described below. Also, SPSS and Minitab macros to calculate these intervals are available at http://www.uwcm.ac.uk/uwcm/ms/Robert.html.

Confidence Interval for a Proportion

Suppose that out of n individuals or units, r are positive for the characteristic of interest. Then the proportion of positive responses is $p = r/n$. We want to calculate a confidence interval for the corresponding proportion in the population from which the sample has been drawn. A CI for p is commonly calculated as $p \pm z \times SE(p)$, where $SE(p) = \sqrt{(p(1-p)/n)}$ and z is 1·96 for a 95 per cent CI. For example, with $n = 30$ and $r = 19$, $p = 0·633$, as above, and the interval runs from 0·461 to 0·806, i.e. from 46·1 to 80·6 per cent.

While this is very easy to calculate, unfortunately it has several serious flaws. For example, in the same study, out of 30 sandblasted specimens, all 30 (100 per cent) failed at the ECI, 0 (0 per cent) failed at the CBI. If we substitute $p = 0$ in the above formula we get a zero SE, and the resulting interval is degenerate, the upper limit as well as the lower limit is zero. Similarly, when $p = 1$ both the lower and the upper limits are 1. Moreover, when r is small (1, 2, or sometimes 3), something equally absurd can happen: we can get a lower limit below 0. Similarly, when $n - r$ is small, the upper limit can exceed 1. Also, the interval is meant to have a 95 per cent chance of including the true population proportion, yet a simulation study shows that its true coverage probability is under 90 per cent for moderate values of n. Furthermore, the interval tends to be located too far out from 0·5, the midpoint of the scale: borrowing familiar terminology, the location of the interval is too *distal*. The consequence is that a calculated upper limit for, say, the incidence of some adverse effect will tend to be falsely reassuring (Newcombe, 1998a).

A variety of alternative methods have been formulated to get around these problems. I recommend a method due to Wilson (1927), known as the score method, which has very good properties for any data and is reasonably calculator-friendly. First, calculate the three quantities $A = 2r + z^2$; $B = z\sqrt{[z^2 + 4r(1 - r/n)]}$; and $C = 2(n + z^2)$. Then, the confidence interval for the proportion is given by $(A - B)/C$ to $(A + B)/C$.

Thus, with $n = 30$, $r = 19$ and $z = 1·96$, we calculate $A = 2 \times 19 + 1·96^2 = 41·84$, $B = 1·96 \times \sqrt{[1·96^2 + 4 \times 19 \times (1 - 19/30)]} = 11·03$, and $C = 2 \times (30 + 1·96^2) = 67·68$. Then, the 95 per cent confidence interval for the proportion of etched specimen bonds that fail at the CBI runs from $(41·84 - 11·03)/67·68 = 0·455$ to $(41·84 + 11·03)/67·68 = 0·781$, that is, from 45·5 to 78·1 per cent. Note that the observed proportion, 63·3 per cent, is not at the midpoint of the interval.

The interpretation is very similar to that of a CI for a mean. Assuming, of course, that the results of the laboratory study give us a reliable guide to what would happen in clinical practice, our best estimate is that for 63·3 per cent of bonds like this, failure would be at the CBI, rather than the ECI. We admit that this population proportion could be as low as 45·5 per cent or as high as 78·1 per cent, and still plausibly give rise to an observed proportion of 19 out of 30. The width of this interval is an expression of the degree of precision to which we have narrowed down where the true proportion is likely to lie.

When r and, hence, p is zero, the interval simplifies to 0 to $z^2/(n + z^2)$. When $r = n$ so that $p = 1$, the interval becomes $n/(n + z^2)$ to 1. Thus, a 95 per cent CI for the CBI failure rate for sandblasted specimens is 0 to $1·96^2/(30 + 1·96^2) = 0·114$, i.e. 0–11·4 per cent. Here, the lower limit is the same as the point estimate, at zero. The upper limit is greater than zero; with a true proportion of 11·4 per cent, occasionally we would get 0 positives in a sample of 30, we cannot rule out

the possibility that the true proportion could be around 11·4 per cent or approximately one in nine. Similarly, a 95 per cent CI for the ECI failure rate is $30/(30 + 1.96^2) = 0.886$ to 1. This conveys the same information, because if 11·4 per cent of bonds fail at the CBI, $100 - 11.4 = 88.6$ per cent of them must fail at the ECI.

Two Samples: unpaired case

In the enamel preparation study, 30 out of 30 (100 per cent) of bonds involving sandblasted specimens failed at the enamel-cement interface, but only 19 out of 30 (63·3 per cent) of etched specimens did so. The difference here is $D = p_1 - p_2 = 1.0 - 0.633 = 0.367$ or 36·7 per cent. We often want to calculate a confidence interval for a difference between two proportions. Here, we think that there is a 37 per cent greater chance that the failure will be at the ECI for sandblasted bonds, but we would like to express the uncertainty on this figure resulting from the limited sample size used. The simplest method of calculation is closely related to the simple method for the single proportion, and shares its drawbacks. A better method is to calculate l_1 and u_1, the lower and upper limits that define the 95 per cent CI for the first sample, and l_2 and u_2, the lower and upper limits for the second sample, using the score method as above. Then, the 95 per cent confidence interval for the difference is calculated as:

$$D - \sqrt{[(p_1 - l_1)^2 + (u_2 - p_2)^2]} \text{ to}$$
$$D + \sqrt{[(p_2 - l_2)^2 + (u_1 - p_1)^2]}$$

(Newcombe, 1998b). In our example, $p_1 = 1.0$, $p_2 = 0.633$, $D = 0.367$, $l_1 = 0.886$, $u_1 = 1.0$, $l_2 = 0.455$ and $u_2 = 0.781$. The 95 per cent CI for the difference is then

$$0.367 - \sqrt{[(1 - 0.886)^2 + (0.781 - 0.633)^2]} \text{ to}$$
$$0.367 + \sqrt{[(0.633 - 0.455)^2 + (1 - 1)^2]},$$

that is, from 0·180 to 0·545. Thus, although the best estimate for the difference between the proportions of bracket failures at the ECI is 37 per cent, the 95 per cent CI ranges from 18 to 54 per cent, showing the imprecision due to the limited sample size. This CI does not include the value 0, corresponding to a difference that was judged statistically significant in the original article using a chi-square test. Indeed, supposing we felt that, say, a 15 per cent difference in ECI failure rates would be clinically important, we would then note that the whole of the interval lies above this value, and we could assert that *even at a conservative estimate* the difference is large enough to be clinically important.

When the two proportions have equal denominators, we have to consider the study design carefully to ascertain whether we should regard the two sets of results as individually paired or not. In the above example, there is no suggestion that a paired design was used. It could be advant-

ageous to design a study of this kind so that each of 30 subjects provided a pair of contralateral premolars, of which one would be allocated (randomly) to each preparation method. This would be a *paired* design and the analysis should correspond. A confidence interval for a difference between proportions, based on individually paired data, can be calculated in a closely related way: for further details, see Altman & Machin (2000), or the software referred to above.

References

Altman, D. G., Bryant, T. N., Gardner, M. J. and Machin, D. (Eds) (2000)
Statistics with Confidence—confidence intervals and statistical guidelines, 2nd edition,
British Medical Journal, London.

Heasman, P. A., MacGregor I. D. M., Wilson, Z. and Kelly, P. J. (1998)
Toothbrushing forces in children with fixed orthodontic appliances,
British Journal of Orthodontics, 25, 187–190.

Newcombe, R. G. (1998a)
Two-sided confidence intervals for the single proportion: comparison of seven methods,
Statistics in Medicine, **17,** 857–872.

Newcombe, R. G. (1998b)
Interval estimation for the difference between independent proportions: comparison of eleven methods,
Statistics in Medicine, **17,** 873–890.

Newcombe, R. G. (2000)
Confidence intervals: an introduction,
Journal of Orthodontics, **27,** 270–272.

Sargison, A. E., McCabe, J. F. and Millett, D. T. (1999)
A laboratory investigation to compare enamel preparation by sandblasting or acid etching prior to bracket bonding,
British Journal of Orthodontics, **26,** 141–146.

Wilson, E. B. (1927)
Probable inference, the law of succession, and statistical inference,
Journal of the American Statistical Association, **22,** 209–212.

Erratum

In Part I of the Statistics Series which appeared in Journal of Orthodontics Volume 27 Number 3 September 2000, an error was printed on page 270, second column. Lines 4 to 8 are reprinted below as they should have appeared.

Oxford University Press and the Editors apologise to the author and subscribers for this error.

Usually, a 95 per cent *confidence interval* (CI) is calculated. This is defined as $\bar{x} \pm t \times SE$. SE denotes the *standard error of the mean*, which is $SD\sqrt{n}$, here $124\sqrt{20} = 22.6$. It expresses the *precision* of the sample mean. The multiplier, t, is approximately 2 for a 95 per cent CI.