Statistics Series

Statistical Applications in Orthodontics. Part I. Confidence Intervals: an Introduction

R. G. NEWCOMBE, PH.D., C.STAT., HON. M.F.P.H.M.
Statistical adviser to Journal of Orthodontics.

Readers of the *Journal of Orthodontics* and other research publications will have met with a variety of methods for summarizing results. These include:

(1) *summary statistics*, such as means, standard deviations, and proportions;
(2) *confidence intervals* for these quantities;
(3) *hypothesis tests*, which attempt to identify when differences between groups are sufficiently well substantiated to be credible.

The use of summary statistics is fundamental and incontrovertible in both descriptive and comparative studies. For many decades, hypothesis testing was the mainstay of inference in comparative studies, yet it is far from ideal for this purpose. While from time to time very strong views are expressed (Savage, 1999), I hope that readers will find my response (Newcombe, 1999) presents a balanced view of the issues.

In particular, confidence intervals are often advocated as the most useful way to express the uncertainty attaching to research findings, which results from the fact that, of necessity, it is only possible to study a sample of limited size. *Study size* usually refers to the number of individuals in a study or a group. It may also refer to the number of manufactured specimens under test (e.g. archwires or brackets) or to a collection of teeth, though in the latter case great caution is needed as it may not be appropriate to regard different teeth from the same subject as *independent*. In this, the first article in a short series, I will show how to calculate confidence intervals for means and differences between means, and explain how these should be interpreted. A second article will deal with proportions and their differences. The third article will present various ways to assess what sample size should be studied.

A very useful resource, for both clinicians analysing data for the first time and experienced clinical researchers, is the book *Statistics with Confidence*, first published by the *British Medical Journal* in 1989. This comes with accompanying software called CIA (Confidence Interval Analysis). A second, considerably improved edition (Altman *et al*., 2000) is in print. All the calculations in this first article are easily performed using commercial statistical software, though it is important to remember that most packages require you to start from the raw data, not from the means and standard deviations.

Confidence Interval for a Single Mean

Heasman *et al*. (1998) studied manual toothbrushing forces in $n = 30$ children immediately before and at 2 and 14 weeks after attachment of fixed appliances to both arches. At baseline, the mean force was $\bar{x} = 194$ g, with a standard deviation (SD) of 124 g.

Usually, a 95 per cent *confidence interval* (CI) is calculated. This is defined as $\bar{x} \pm t \times \text{SE}$. SE denotes the *standard error of the mean*, which is $\text{SD}/\sqrt{n}$, here $124/\sqrt{20} = 22{\cdot}6$. It expresses the *precision* of the sample mean. The multiplier, t, is approximately 2 for a 95 per cent CI.

When the sample size is small, the required value of t is rather greater. Table 1 shows the value of t required to construct a 95 per cent CI, for various numbers of *degrees of freedom* (d.f.). For the case of a single mean, the number of degrees of freedom is $n - 1$. In this example, d.f. = 29, so $t = 2{\cdot}045$. More extensive tabulations, including values for 90 and 99 per cent confidence, are available in statistical textbooks, and also software such as Minitab can be used.

So we report the sample mean, 194 g, and the 95 per cent CI for the mean, which is calculated as $194 \pm 2{\cdot}045 \times 22{\cdot}6$, i.e. 194 ± 46 or 148–240 g. How is this to be interpreted?

First, the *point estimate*, the best single estimate of the mean brushing force at baseline, *in subjects such as those in the sample studied*, is 194 g. It is implicitly assumed that these 30 children are *representative* of some wider *population* of children who are candidates for orthodontic treatment. We are interested in this study group of 30 subjects in their own right, to some degree. However, of much greater interest is what the results from this limited *sample* tell us about the wider *population*, to which we want the results to be applicable. True *random* sampling of a defined population of interest would ensure representativity. In most studies, including our illustrative example, a *consecutive series* of subjects meeting specified eligibility criteria are investigated. Since this is not strictly a random sample of the relevant population, the reader should consider issues such as the following. The study was carried out in the

TABLE 1 *Values of* t *used to calculate 95 per cent CIs, by number of degrees of freedom*

d.f.	*t*	d.f.	*t*	d.f.	*t*
10	2·228	20	2·086	30	2·042
11	2·201	21	2·080		
12	2·179	22	2·074	>30	1·96 + 2·4/df
13	2·160	23	2·069		
14	2·145	24	2·064	40	2·02
15	2·132	25	2·060		
16	2·120	26	2·056	60	2·00
17	2·110	27	2·052		
18	2·101	28	2·048	120	1·98
19	2·093	29	2·045		
				∞	1·96

0301-228X/00/030000+00$02.00

North East of England: would similar results be expected in my catchment area? Are a 2:1 female:male ratio and a mean age at start of treatment of 13·7 years similar to my caseload? If the answers to such questions are reassuring, then the reader can reasonably expect the results to be relevant to his/her practice.

Next, again assuming the sample is relevant, we turn to the CI, 148–240 g. This is interpreted as giving a *margin of error* either side of the observed mean of 194 g. The *width* of the interval is approximately four times the SE: the interval width, like the SE, expresses the degree of *precision* within which we can claim the sample mean estimates the population mean. The interval is designed to have a 95 per cent chance of including the population mean, which is usually denoted by μ. There is a 2·5 per cent chance that sampling would produce a set of results that are so much too high that the lower confidence limit is greater than μ, and similarly 2·5 per cent chance that the upper limit is lower than μ. However, on 95 per cent of occasions when a 95 per cent CI is calculated, it fulfils its objective of including μ. Note that this does not imply that all values between 148 and 240 g are equally likely for μ. Values around 194 g are much more plausible than values towards one or other end of the interval.

Often, readers interpret such an interval to mean 'there is 95 per cent chance that μ is between 148g and 240g'. Strictly speaking, this is an incorrect inference. To be able to make such an assertion, we would need to take into account all information from other sources bearing on μ. This CI is designed to summarize the results of a *single* study *only*—we may later interpret the findings, informally, in the light of other knowledge (Newcombe, 1999).

Note that the lower and upper confidence bounds are in the original units, in this case grams. We return to this point later.

In the above example, the distribution of brushing force cannot be very close to Gaussian, because the standard deviation is quite large in relation to the mean, whereas values below zero would be meaningless. Notwithstanding this, the CI as calculated above is still reasonable. Only when the distribution is very far from Gaussian and the sample size is small, is the above simple calculation inappropriate.

Confidence Interval for a Difference Between Means of Two Independent Samples

In the Heasman example, the mean brushing force was 220 g (SD 136 g) in $n_1 = 10$ males and 181 g (SD 119 g) in $n_2 = 20$ females. Often a familiar *hypothesis test*, the unpaired *t*-test, would be used to compare these two means. An alternative, complementary approach is to calculate the observed difference, together with a confidence interval. In this instance, the difference is 220 – 181 = 39 g. The standard error of the difference is $\sqrt{(SD_1^2/n_1 + SD_2^2/n_2)} = \sqrt{(136^2/10 + 119^2/20)} = 50{\cdot}6$ g. Equivalently, the standard errors of the male and female means are $136/\sqrt{10} = 43{\cdot}0$ g and $119/\sqrt{20} = 26{\cdot}6$ g, and these combine by 'squaring and adding' to give $\sqrt{(43{\cdot}0^2 + 26{\cdot}6^2)} = 50{\cdot}6$ g. The number of degrees of freedom here is $n_1 + n_2 - 2 = 28$, so $t = 2{\cdot}048$. So the 95 per cent CI for the difference is $39 \pm 2{\cdot}048 \times 50{\cdot}6$ or 39 ± 104, i.e. from –65 to +141 g.

Note that this interval is very wide, reflecting the small sample size studied. Our best estimate is that boys use 39 g more force than girls, on average. The CI tells us that the difference between boys and girls in the population could be as large as 141 g. Or, conversely, it is possible that girls use 65 g greater force than boys. It is quite credible that the population difference could be zero. This corresponds to the fact that an unpaired *t*-test would not classify this as a statistically significant difference, $P > 0{\cdot}05$. However, the CI gives much more information: it gives a range of values, *in original (force) units*, for the effect size, whereas the hypothesis test gives only a very indirect measure, a kind of coincidence probability. We can interpret the *clinical importance* of the effect size. For example, suppose we felt that a 50 g difference would be important. Then the point estimate, 39 g, would suggest an unimportant difference. However, the wide confidence bounds indicate that an importantly greater brushing force in boys or even in girls cannot be ruled out by this small set of results. So we get a more realistic appraisal of the limitations resulting from the chosen sample size.

In an *observational* study such as this, of course, we cannot be sure that the observed difference is a *direct* consequence of gender. For example, the mean age at presentation could be different in males and females, and this would distort the difference. It is worth considering to what degree an observed difference could be affected by *confounding* by some other factor(s) in this way. Partial confounding can be adjusted for in the analysis, though methods for doing so are beyond the scope of this article.

Confidence Interval for a Paired Difference of Means

Still in the Heasman study, the mean brushing force increased from 194 g (SD 124 g) at baseline to 203 g (SD 77 g) 2 weeks after appliance attachment. Because *the same* 30 subjects were studied on both occasions, the analysis must take this into account. *Paired* methods of analysis are used whenever subjects are used as their own controls, as here, or in a cross-over or split-unit trial. They are also used when a specific control subject is identified for each case, for instance in a retrospective study comparing cases of oral cancer and individually matched controls for tobacco consumption.

To obtain a CI for a paired difference, we need to calculate the mean and SD of the individual paired differences—carefully heeding which are positive and which are negative, of course. The mean difference is then the same as the difference of the pre- and post-treatment means, here $\bar{d} = 203 - 194 = 9$ g. The SD of the paired differences cannot be calculated directly from the separate SDs for the two series, 124 and 77 g, without additional information. The SD of the paired differences is obtained directly from the raw data, in this case 147 g (P. A. Heasman, personal communication). A 95 per cent CI for the mean change is then simply obtained as $\bar{d} \pm t \times SE$, where $SE = 147/\sqrt{30} = 26{\cdot}8$ g, d.f. $= n - 1 = 29$, $t = 2{\cdot}045$, and the CI is $9 \pm 2{\cdot}045 \times 26{\cdot}8$ or 9 ± 55, i.e. –46 to +64 g, as in the original article. Once again, this includes 0, corresponding to a non-significant difference, but the interval also tells us what average degree of alteration in force, in either direction, is compatible with the observed results.

Note that this analysis is only stating something about the tendency for the average force to increase slightly. After appliance placement the SD shrinks considerably, even though the range remains virtually unaltered, suggesting that the applied force tends to become more controlled after attachment.

References

Altman, D. G., Bryant, T. N., Gardner, M. J. and Machin, D. (Eds) (2000)
Statistics with Confidence—Confidence Intervals and Statistical Guidelines, 2nd edn,
BMJ Books, 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. (1999)
The controversy over how to present research findings,
British Journal of Orthodontics, **26,** 233–234.

Savage, M. (1999)
Statistical significance testing,
British Journal of Orthodontics, **26,** 244.