Statistical Applications in Orthodontics

Part III. How Large a Study is Needed?

ROBERT G. NEWCOMBE
UWCM, Cardiff, UK

Every research study involves a finite number of subjects or of some other units such as tooth specimens. Two previous articles (Newcombe, 2000a, 2000b) outlined the rationale for confidence intervals (CIs) as the most helpful expression of sampling uncertainty, and demonstrated methods to calculate CIs for means, proportions, and their differences. This all relates to the analysis stage of the study, of course. Long before then, however, the investigators should plan carefully what sample size to study. Standard statistical methods assume that the sample size is laid down in advance. If the investigators drift through a study, with no clear protocol, and simply terminate when statistical significance is reached, conventional statistical analyses are meaningless. At an early stage of protocol development, careful thought should be given to the issue of study size. This should be large enough to yield reliable information. Conversely, an unnecessarily large study will be wasteful of resources, will not yield a timely result and may be ethically unacceptable. In practice, many studies could have benefited from being (much) larger, while relatively few are too big. While there is no such thing as getting the study size exactly right, what is needed is a study size that can be defended as reasonable, in the light of what is understood at the planning stage.

Two main approaches are available. In a descriptive study, we can specify that we want to estimate a particular mean or proportion within a certain margin of error with 95 per cent confidence. Similarly, in a comparative study, we can specify the desired interval width for an effect size such as a difference in means or proportions. Alternatively, we can calculate the sample size required giving a *power* of, say, 80 per cent to detect a pre-specified difference. Whichever we choose, the calculation should relate to the outcome measure we regard as of *primary* importance to the study, this choice should be specified in the protocol in any case. The sample size arrived at should be increased to allow for whatever degree of attrition or dropout it is reasonable to anticipate is liable to occur in the study.

The confidence interval approach

In a study of toothbrushing forces before and after orthodontic appliance attachment, Heasman *et al.* (1998) reported a mean force of 194 g (SD 124 g) for $n = 30$ children before appliance attachment. We showed in the first article that a 95 per cent CI for this mean is 194 ± 46 g, i.e. from 148 to 240 g. With the benefit of hindsight, we could well argue that the study was too small, we would wish to estimate the population mean to within a narrower margin of error. Suppose that in a new study we want to estimate the mean to within ±30 g, i.e. we plan to end up with a 95 per cent CI calculated as the observed mean $\bar{x}$ ± 30 g. How many subjects should be recruited?

A 95 per cent confidence interval for the mean is approximately $\bar{x} \pm 2\ \mathrm{SD}/\sqrt{n}$, when n is large. We want to choose n so that $2\ \mathrm{SD}/\sqrt{n}$ will be 30 units. Assuming a SD of 124 g, from the published study, this would require $2 \times 124/\sqrt{n} = 30$ and $n = 68$. Thus, in a study involving 68 subjects, we would expect to end up with a 95 per cent confidence interval of the form $\bar{x} \pm 30$. This would apply whether $\bar{x}$ turned out to be 160, 220 or even 250 g.

We might decide that even $\bar{x} \pm 30$ g is too wide to be informative. Suppose we wanted to estimate the mean to within 15 g. This would require $2 \times 124/\sqrt{n} = 15$ and $n = 273$. This sample size is (to within rounding error) *four* times what we would require in order to estimate the mean to within ±30 g. Doing so would greatly increase the cost of the study, up to four-fold.

We can plan a sample size for estimating a proportion similarly. Sargison *et al.* (1999) found that the site of bond failure for 19 out of 30 (63 per cent) of etched specimens was at the enamel-cement interface (ECI). A very simple confidence interval for this proportion is $P \pm 1{\cdot}96\sqrt{[P(1-P)/n]} = 0{\cdot}633 \pm 0{\cdot}172$, i.e. from 46·1 to 80·6 per cent. We might, however, want to estimate this proportion to within an absolute margin of error of ±10 per cent with 95 per cent confidence. To do so, we would need to make some assumption about P, as this determines the standard error. With $P = 0{\cdot}633$, we would then want $1{\cdot}96 \times \sqrt{(0{\cdot}633 \times 0{\cdot}367/n)} = 0{\cdot}10$ and $n = 89$. In planning a study *de novo*, of course, we would usually not have recourse to an exact numerical estimate of P, so we would need to make an informed guess. With $P = 0{\cdot}6$, we would need $1{\cdot}96 \times \sqrt{(0{\cdot}6 \times 0{\cdot}4/n)} = 0{\cdot}10$ and $n = 92$. What value of P we assume can make a big difference here: for example, with the very different assumption $P = 0{\cdot}9$, we would require only 35 subjects. If we really have no idea about P, a play-safe assumption is to substitute $P = 0{\cdot}5$, giving in this case $n = 96$. This guarantees estimation to within ±10 per cent or less, irrespective of the value of P. However, such an extreme shot in the dark is highly inadvisable; it is much more sensible to perform a small pilot study first, if only to decide whether P is around 0·5 or is at one or other end of the range. Moreover, in practice, whether the calculation produced the answer 89, 92, or 96, we would probably plan to use 100.

The approach extends to comparative studies, and applies equally to unpaired and paired analyses. For example, in the brushing study, the mean change in brushing force after appliance fitting was an increase of 9 g. The SD of the changes was 147 g, suggesting just over half the children increased their brushing force and just under half decreased it. A 95 per cent CI for the mean change is approximately $\bar{x}$

0301-228X/01/020000+00$02.00

$\pm 2 \times 147/\sqrt{n}$. To estimate the mean change to within ±20g would require a study with $2 \times 147/\sqrt{n} = 20$ and $n = 216$ subjects.

Type II error and power

In a simple hypothesis test such as a *t*-test or a chi-square test, the test statistic is calculated from the observed data. We assess the credibility of the *null hypothesis* (H_0) of no difference between the two groups, in the light of the data. If the observed difference is too large to be well explained as a chance difference, we reject H_0 in favour of the *alternative hypothesis* (H_1) that there is a difference. If the observed difference is compatible with the play of chance, we accept H_0 as a possible explanation.

When the null hypothesis is true, but is rejected on analysing the data, we make a *type I error*. Conventionally, the *type I error rate*, α, of a test is set at 5 per cent. When the calculated value of *t* exceeds the appropriate tabulated value (about 2), or chi-square for a 2 by 2 table exceeds 3·84, the null hypothesis is rejected at the 5 per cent level and the difference is declared statistically significant. The *P-value* is the probability of getting a difference as extreme as the one observed, or more so, purely by the play of chance assuming H_0 is true. H_0 is rejected if the calculated *P*-value is less than α. The implication of working at an α level of 0·05 is that if H_0 is in fact true, the data prompts rejection of H_0, inappropriately, in 5 per cent of studies.

A second type of error may occur. We may fail to detect a real difference, and get a non-significant result even though there really is a difference between the two underlying populations. This is a *type II error*. The *type II error rate*, β, is the probability of a non-significant difference when H_1 is true. The *power*, $1 - \beta$, is the probability of detecting the difference as statistically significant. Many studies are too small and operate at too low a power. Eighty to 95 per cent is regarded as reasonable, but lower than 80 per cent is very much 'hit and miss' and is unsatisfactory. Conversely, a sample size large enough to give a power of over 95 per cent may be a poor use of resources. We can plan a study to be large enough to yield a suitably high power to detect a specified size of difference. This should be large enough to be clinically important, but not so large as to be implausible in the light of existing knowledge. The time to consider power is when planning the study; '*post-hoc* power calculations' are sometimes seen, but are an exercise in self-deception, really no more than a rescaling of the *P*-value.

Several approaches to power assessment are available. A variety of formulae for direct calculation are found in statistics textbooks. Computer software, published tables, and nomograms are available (examples are included in the bibliography). A very useful, flexible, widely applicable method involving indirect calculation is demonstrated below. This involves choosing an arbitrary sample size, calculating the expected value of the test statistic, then comparing with the required value from a special table. The sample size is then scaled by an appropriate factor to produce the desired power.

Power assessment by indirect calculation

In the study of Heasman *et al*, the mean toothbrushing force at baseline was 220g (SD 136g) in boys and 181 g (SD 119 g) in girls. Suppose we decided that in a new study, we want an 80 per cent power to detect a difference of 30 g as statistically significant at the 5 per cent level. Based on the existing study, it seems reasonable to assume a SD of 130 g. Suppose that, as in the published study, we expect to recruit twice as many girls as boys—most other methods do not have the flexibility to build in this aspect.

Thus, take an entirely arbitrary starting guess, say 100 boys and 200 girls. Then the expected value of the unpaired *t*-test statistic is $30/[130\sqrt{(1/100 + 1/200)}] = 1{\cdot}88$. From Table 1, we want to plan for a *t*-value of 2·80; 1·88 is too low and corresponds to a power a little below 50 per cent. Because the expected value of *t* is proportional to the *square root* of the sample size, we need to multiply our initial guess by (2·80/1·88) *squared*, that is, by 2·21. So we would require to study approximately 221 males and 442 females. It is very advisable to check this result by calculating $30/[130\sqrt{(1/221 + 1/442)}]$, which is 2·80.

This would be a very time-consuming study. Alternatively, we could detect a difference of 60 g, *twice* as large as the value assumed above, with 80 per cent power using a study *one-quarter* as large as this, with 55 males and 110 females. However, we should not crank up our target difference in this way unless we (and our peers) are quite happy to regard this as a plausible size for the true difference.

A similar approach can be applied when the outcome variable of interest is *binary*. In the study of Sargison *et al*. (1999), bond failure at the ECI occurred in 100 per cent of sandblasted specimens, but only 63 per cent of etched ones. Suppose we repeated the study using a different type of bracket or bonding agent. Often investigators want to assume extreme values for the projected outcome, sometimes based on poorly controlled preliminary work, but experience suggests it is wiser to make more conservative assumptions. In this case, it might be reasonable to hypothesize ECI failure rates of 90 and 70 per cent in the two groups. Suppose we want a 90 per cent power to detect such a difference, if it exists. With 100 specimens per group, our expected 2 by 2 table would then be as shown in Table 2.

If the above results were to be observed in the new study, the calculated chi-square value would be $[(90 \times 30) - (70 \times 10)^2] \times 200/(100 \times 100 \times 160 \times 40)$ or 12·5. From Table 1, this is larger than the required value, 10·51. Because chi-square is proportional to the *actual* sample size, we multiply 100 by 10·51/12·5, so it is sufficient to use 84 specimens per group.

TABLE 1 *Target values for t- and for chi-square (1 degree of freedom) for power 80, 90, and 95 per cent, using a test with a conventional 5 per cent α level*

Power	80 per cent	90 per cent	95 per cent
t	2.80	3.24	3.60
Chi-square	7.85	10.51	12.99

TABLE 2 *Projected data in a study following Sargison et al. (1999)*

Site of failure	Sandblasting	Etching	Total
ECI	90	70	160
CBI	10	30	40
Total	100	100	200

Note that these two examples illustrate the principle that the appropriate sample size for a clinical study tends to be greater than for a laboratory one, because in the former, biological variation between individuals is usually the dominant source of variation. The two examples above relate to unpaired comparisons, but a similar approach can be used when there is paired data: as long as we can calculate the expected value of t, or of a 1 df chi-square statistic, the method works. For an unmatched study with three or more groups, the simplest approach is to use the above method as if there were just two groups, to give the number of subjects to be recruited to each group.

Power assessment using a nomogram

Altman (1980) has developed a nomogram linking the power of a study to the sample size. It is reproduced here as Figure 1, with permission from the BMJ Publishing Group and the author. It is both of practical use and also instructive, enabling the reader to appraise the interrelationship of sample size, target difference and power. It is designed for comparison of the means of two independent samples of equal size.

Thus, returning to Heasman *et al.* (1998), suppose that as before we want an 80 per cent power to detect a difference of 30 g as statistically significant at the 5 per cent level, and we assume an SD of 130 g. Then we calculate the

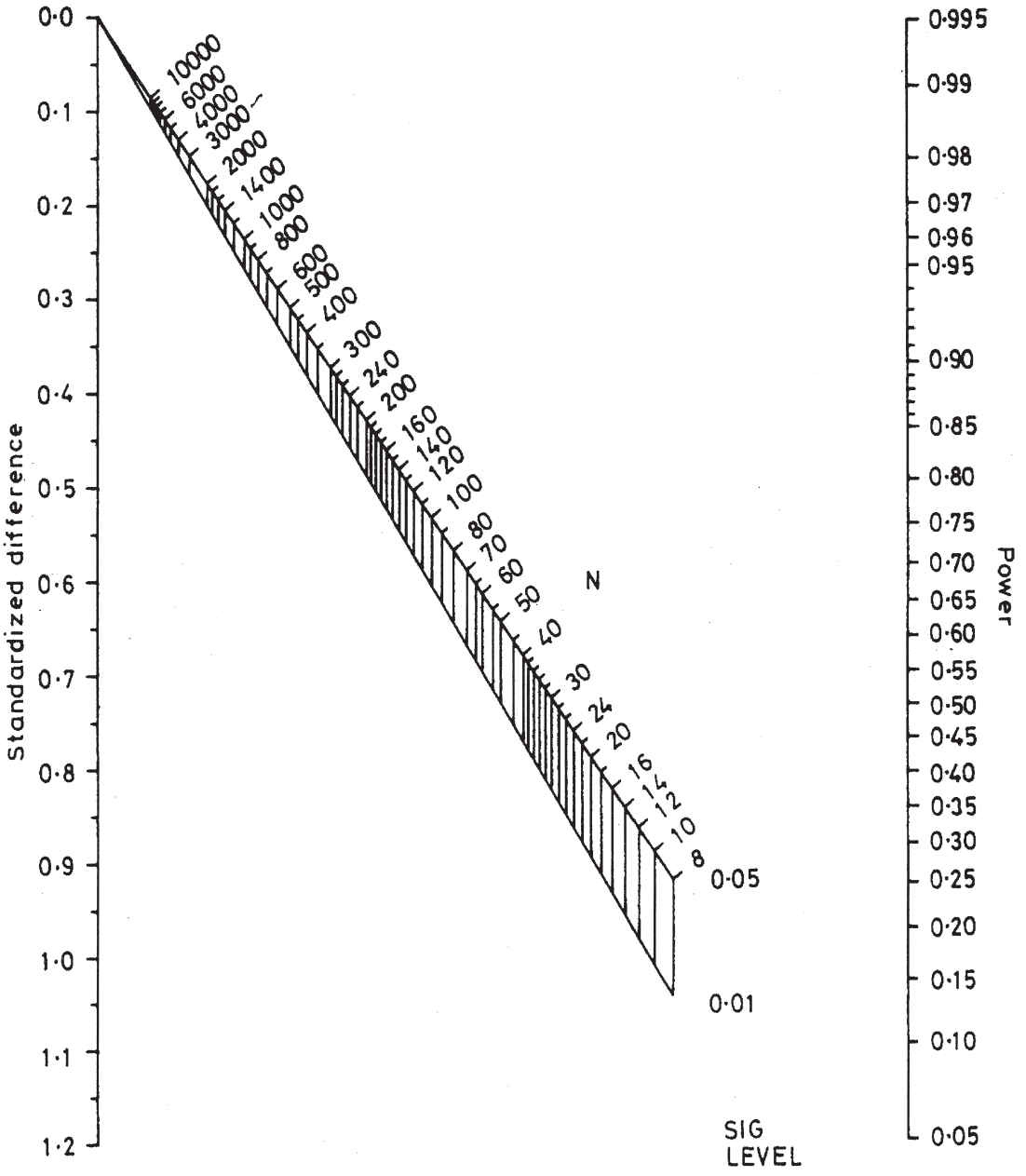

FIG. 1 Nomogram for sample size and power, for comparing two groups of equal size. Gaussian distributions assumed. (Reproduced from *British Medical Journal*, 1980, **281,** 1336–1338, with permission.)

standardized difference as 30/130 = 0·23. Then we locate 0·23 on the left-hand axis and the required power, 0·80, on the right-hand axis. We join these two points with a line or a straight edge. Finally, we read off from the diagonal line the sample size required. We use the upper diagonal line for a 5 per cent level test or the lower diagonal line if a 1 per cent level test is intended. Here, we read off from the upper diagonal line a value of approximately 550, indicating that a total of 550 subjects should be recruited, 275 in each group. (Note that the total number of subjects calculated here is rather lower than the figure of 221 + 442 = 663 obtained earlier. This is because, for a given total sample size, the greatest power is achieved by using two groups of equal size. However, it may be better to plan for unequal sizes if as here recruitment rates to the two groups are expected to be dissimilar.)

Suppose now that it is decided that only a total of 200 subjects can be recruited. We join the points 0·23 on the left-hand axis and 200 on the upper diagonal, and extend the line until it meets the right hand axis. We read off that this results in a power of 40 per cent, which is inadequate.

The investigator then asks, of course, what kind of difference could be reliably detected using 200 subjects. We join 0·80 on the right-hand axis and 200 on the upper diagonal, and read off 0·39 on the left-hand axis. The true difference would have to be as large as 0·39 × 130 g, i.e. just over 50 g to have an 80 per cent chance of being detected at the conventional 5 per cent significance level. Generally, neither the nomogram, nor a statistician using it, could tell the investigator whether it is plausible that a difference as large as this could exist. It is up to the investigator to discuss this with colleagues to arrive at a consensus, based on the best understanding at the time, and only to proceed with the study on 200 subjects if such a difference is judged plausible.

It is important to realize that, especially with an aid such as the nomogram that lends itself to interactive use, it is very easy to use power calculations inappropriately, to legitimize decisions that have already been made on grounds of convenience alone. It has been remarked cynically that the negotiations between investigator and statistician can be little more than a ritual dance. I hope that the issues I have developed here make it clear to readers that the whole exercise is quite pointless unless there is a commitment to take seriously and honestly the issue of choosing a realistic size for the treatment effect that we want to achieve a high power to detect.

References

Altman, D. G. (1980)
Statistics and ethics in medical research. III. How large a sample?
British Medical Journal, **281,** 1336–1338.

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.

Machin, D. and Campbell, M. G. (1987)
Statistical Tables for the Design of Clinical Trials,
Blackwell, Oxford.

Newcombe, R. G. (2000a)
Statistical methods in orthodontics. Part I. Confidence intervals: an introduction.
Journal of Orthodontics, **27,** 270–272.

Newcombe, R. G. (2000b)
Statistical methods in orthodontics. Part II. Confidence intervals for proportions and their differences,
Journal of Orthodontics, **27,** 339–340.

***nQuery Advisor*, Release 3 (1998)**
Statistical Solutions, Cork.

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.