Confidence intervals and sample sizes: don't throw out all your old sample size tables

Leslie E Daly

Over the past few years it has been recommended that emphasis be placed on the confidence interval rather than on hypothesis testing in the statistical analysis of medical data.[1-3] Although the two methods approach the analysis and presentation of data differently and confidence intervals make assessment of results easier, differences in basic interpretation arise only in exceptional circumstances.[4] On the other hand, estimates of sample size, essential at the planning stage of a study, if based on a confidence interval approach are, in general, quite different from the traditional estimates based on hypothesis testing. In this paper I argue that the traditional methods are more appropriate for comparative studies and that, with a change in interpretation to fit in with a confidence interval analysis, standard sample size tables and formulas should still be used.

Confidence intervals versus p values

A large proportion of medical research concerns the comparison of two groups, each of which may be considered a sample from a larger population. A hypothesis testing approach to statistical analysis typically determines whether some appropriate comparative measure (such as the difference between means or a relative risk) is significantly different from its null value (for example, a mean difference of zero or a relative risk of one). A confidence interval approach, however, concentrates on an estimation of the comparative measure together with its confidence intervals. The confidence interval gives an indication of the degree of imprecision of the sample value as an estimate of the population value. It is important to note, however, that hypothesis testing and confidence intervals are intimately connected.[4] If a 95% confidence interval does not include the null value of the hypothesis test then we can infer a statistically significant result at the two sided 5% level.

In this paper I consider the comparison of means by using a difference measure in the two group parallel design, but similar considerations apply to other comparative measures, studies of more than two groups, and within subjects designs such as crossover trials. The arguments in this paper are not germane, however, when the sole purpose of an investigation is to obtain an estimate for a non-comparative measure, as, for example, in a descriptive prevalence study in which no comparisons are planned. Such situations are rare and hypothesis tests are inappropriate.

Department of Community Medicine and Epidemiology, University College Dublin, Dublin 2, Ireland
Leslie E Daly, PHD, *lecturer in medical statistics*

BMJ 1991;**302**:333-6

Sample size and hypothesis tests

One of the critical aspects of study design is to estimate the sample size required and, traditionally, such estimates are based on a hypothesis testing approach to data. Suppose a clinical trial is planned to compare an antihypertensive drug with a placebo and that suitable hypertensive patients are to be randomised into two equal sized groups. For simplicity, we assume the groups will have similar baseline blood pressures and that the treatment effect is to be evaluated by examining the difference in mean systolic blood pressure after a period of, say, six weeks.

To determine the sample size required for this trial several quantities must be considered[5,6]:

- Firstly, the *significance level* (usually 5% or 1%) at which we wish to perform our hypothesis test and if it is to be one tailed or two tailed. (Apart from in exceptional circumstances two tailed tests are usually more appropriate)
- Secondly, the *smallest clinically worthwhile difference* in blood pressure we wish to detect. We must distinguish here between the blood pressure difference that we might observe in our study (the sample result) and the real treatment effect. The real treatment effect can be thought of as the difference in blood pressure that would be observed in a study so large that sample variation was precluded, or, alternatively, as the blood pressure difference between the "populations" of treated and untreated patients. If there was a real treatment effect of important size we would want our study to reflect this with a statistically significant result. We would be unlikely, however, to be interested in detecting a very small (real population) difference of, say, only 1 mm Hg as from a clinical point of view such a treatment effect could be considered negligible. We therefore decide on the smallest difference worth detecting such that if the real difference was this large or larger we would be likely to achieve a significant result; on the other hand, for real differences smaller than this a non-significant result is judged acceptable.[7] In our trial the smallest clinically worthwhile difference might be set at 5 mm Hg
- Thirdly, the *power* of the study. This is the chance of obtaining a significant result if the real effect is as great or greater than the smallest worthwhile difference specified. Powers of 80% or 90% are typical choices
- Fourthly, for quantitative data, the *variability of the measure* in the study population. This is usually determined from a pilot investigation or from published results. (Note that these calculations assume that the distribution of the measure is at least approximately normal.) For illustrative purposes we shall take the standard deviation of systolic blood pressure in hypertensives to be 20 mm Hg. (For the comparison of percentages the corresponding parameter required for sample size estimation is based on prior estimates of the percentages in each of the comparison groups.)

Given the comparative measure being used and levels for the four quantities listed above standard formulas, tables, and graphs are available to enable calculation of the required sample size. These are reviewed by Lachin.[6] To illustrate the method table I gives required sample sizes in each group of our clinical

TABLE I—*Sample size in each group for an independent two group comparison of mean blood pressures, prespecifying power to detect a smallest worthwhile difference. (Two sided significance level of 5%, population standard deviation of 20 mm Hg)*

Smallest difference to be detected	Power		
	90%	80%	50%
5 mm Hg	336	251	123
10 mm Hg	84	63	31
15 mm Hg	38	28	14

TABLE II—*Sample size in each group for an independent two group comparison of mean blood pressures, prespecifying confidence interval width (95% confidence interval, population standard deviation of 20 mm Hg)*

Confidence interval width	Sample size required
10 mm Hg	123
20 mm Hg	31
30 mm Hg	14

trial for three different levels of the smallest worthwhile difference to be detected (5, 10, and 15 mm Hg) and powers of 50%, 80%, and 90%. A two tailed 5% significance level and a population blood pressure standard deviation of 20 mm Hg are assumed. The required sample size increases with the power but decreases for higher levels of the difference to be detected. The appendix gives the equation from which these figures are calculated.

Sample size and confidence intervals

Alternative sample size calculations have been proposed based on a confidence interval approach which give rise to different and generally smaller sample size requirements from those given by the standard methods. The vast majority of these are based on the expected width of the confidence interval for the comparative measure being analysed. (The width of a confidence interval is a measure of the imprecision of the sample estimate and is the difference between the upper and lower confidence limits. For example, if a confidence interval was determined to be 90 to 170, its width would be 80.) All else being equal, the larger the sample size the narrower the width of the confidence interval. Once the width has been prespecified the only additional requirements for determination of a sample size by this approach are the confidence level (95% or 99%) and an estimate of the variability of the comparative measure. These specifications are clinically understandable and the difficult concepts of power, null value, and smallest difference to be detected seem to be avoided altogether. In addition, concentration on the precision of the estimate seems to fit in fully with an analysis that is to be performed with confidence intervals. Tables and formulas using this approach are available for various comparative measures.[8-11] (A further refinement is to estimate sample sizes on the basis that the width of the confidence interval, rather than being fixed, is a percentage of the actual population value.[8,9,12]) Table II gives the sample size requirements in each group of our clinical trial to achieve confidence interval widths of 10, 20, or 30 mm Hg for the difference between the mean blood pressures (see appendix for computational details).

Confidence intervals and null values

Although the precision of any measure is very important, estimates of sample size based on the width of the confidence interval can be misleading. The consequences of employing such estimates do not seem clearly to be understood, and, in general, the published work does not consider the problems explicitly. One distinction between hypothesis tests and confidence intervals is important in this regard.[13,14] Hypothesis tests are essentially asymmetrical with the emphasis on rejection or non-rejection of the null hypothesis. Conversely, confidence intervals are symmetrical and estimate the magnitude of the difference between two groups without giving any special importance to the null value. It seems a mistake, however, to conclude that this null value is irrelevant to the interpretation of confidence intervals, even though it plays no part in their calculation. Irrespective of precision, there is a qualitative difference between a confidence interval that includes the null value and one that does not include it. If the null value is included the possibility of no difference must be accepted, while if it is not some difference has been shown at a given level of probability. Herein lies the crux of the problem. In a comparative study can we ever say that the primary goal is just estimation and ignore completely the qualitative distinction between a difference and no difference? The answer is clearly no, and I argue below that the null value must have a central role in the estimation of sample sizes with a confidence interval approach. This is not usually the case.

Confidence intervals, power, and worthwhile differences

The role of the smallest clinically worthwhile difference to be detected (as specified by the alternative hypothesis) has also been questioned in the context of sample sizes based on confidence intervals. Beal states: "With estimation as the primary goal, where construction of a confidence interval is the appropriate inferential procedure, the concept of an alternative hypothesis is inconsistent with the associated philosophy, even when used as an indirect approach to hypothesis testing. Thus one should not, in this situation, determine sample size by controlling the power at an appropriate alternative."[15] This viewpoint is untenable. For determination of a sample size it seems inappropriate to specify the precision of an estimate without any consideration of what the real differences between the groups might be. Unfortunately, though the problem has been recognised by some,[13] workers usually make a correspondence between the precision of the confidence interval and the smallest difference to be detected. In the clinical trial example we might decide that a confidence interval width for blood pressure difference of just under 10 mm Hg would be sufficient to distinguish a mean difference of 5 mm Hg from that of a zero difference. If the confidence interval were centred around this difference of 5 mm Hg the expected interval of 0 to 10 mm Hg would just exclude the null value. However, comparison of the sample sizes based on the hypothesis test approach in table I with those based on confidence intervals in table II shows that those based on confidence intervals would have only 50% power of detecting the corresponding smallest worthwhile differences. Even if the real difference were as large as postulated there would be a 50% chance of the confidence interval overlapping zero if these sample sizes were used.

Explanation of the anomaly

There are two reasons for this apparent anomaly. Assume that the real population blood pressure difference was 5 mm Hg and that, based on a prespecified confidence interval width of 10 mm Hg, a sample size of 123 was used in each group. Firstly, this width is only an expected or average width. The width we might obtain on any actual data from the study would be above its expected value about 50% of the time. Thus the confidence interval, if centred around 5 mm Hg, would have a 50% chance of including zero. Secondly, the sample value of the blood pressure difference calculated on the study results would be as likely to be above the population value of 5 mm Hg as below it. If our sample estimate were, for instance, 4 mm Hg then a confidence interval with the expected width of 10 mm Hg would run from −1 mm Hg to +9 mm Hg and would include zero difference as a likely true value. Thus, specification of the width of a confidence interval as described above without consideration of possible true values of the difference and the power of detecting them (with a confidence interval excluding the null value of zero) can lead to unacceptably small sample sizes with too low a power to detect the required effect.

Proposed solutions

Beal and Grieve propose sample size estimations based on a specification of confidence interval width together with a probability (somewhat akin to power) that the width be less than a given value.[15,16] This

overcomes the problem related to expected width discussed above but does not account for the true location of the parameter of interest. Sample sizes based on this approach are still much lower than traditional estimates.

In planning any investigation the question of power to detect the smallest clinically worthwhile difference must predominate over that of precision. In practice, of course, estimates based on samples large enough to detect small differences will have a high degree of precision. It is only when we are trying to detect large differences (not often found in medical research) that an imprecise estimate will result. In this situation it would in any case be possible to calculate a sample size based on precision also and use the larger of the two sizes so calculated. In line with this view, Bristol[14] gives tables and formulas relating to the width of the interval to the power for detecting various alternatives when comparing differences of means and proportions. However, if these factors have to be considered at all, why should estimates of sample size not explicitly specify power to detect the smallest worthwhile difference in the first place, rather than concealing the specification in a vaguer requirement for confidence interval precision?

Confidence intervals and standard sample size tables

I propose that sample size requirements, which explicitly consider power, null values, and smallest worthwhile differences, can easily be put into a confidence interval framework without the consideration of hypothesis tests in either design or analysis. Although discussion has been in the context of employing the difference between means as a comparative measure, this proposal has general applicability. For a calculation of sample size based on confidence intervals we should specify (a) the confidence level (95% or 99%), (b) the minimum size of the comparative measure we wish to estimate unambiguously (that is, with the confidence interval excluding the null value), (c) the chance of achieving this if the measure actually had this minimum value (in the population). These correspond, of course, to the traditional requirements of (a) the significance level, (b) the smallest worthwhile difference to be detected, (c) the power of the study. Thus with only a slight change of wording the standard procedures based on hypothesis testing can be used to estimate sample sizes in the context of a confidence interval analysis.

It is essential to note that this approach allows for the sampling variability of both the location and width of the confidence interval. The width of the interval, however, is not explicitly prespecified; it is instead determined by the more important criterion that we are unlikely to miss a difference we wish to detect.

Greenland comes nearest to this view in terms of confidence intervals and sample size.[13] The proposal outlined in this paper is based on distinguishing between a particular difference, if its exists, and the null value. Greenland, however, in a subtle modification of this approach, also suggests that the sample size should be large enough to distinguish between the null value and this difference, if the groups are the same. In most situations this extra requirement does not result in an increase of sample size and it seems an unnecessary refinement. A further proposal by Greenland, which greatly increases sample size requirements calculated with confidence intervals, is based on unnecessarily stringent criteria.

Conclusion

There is no doubt that the whole topic of traditional sample size calculation tends to be complex and misunderstood and, even today, many studies are carried out without computing the necessary numbers.[17,18] Estimating an appropriate sample size is a vital part of any research design, and it is important that the current emphasis on using confidence intervals in analysis and presentation does not mislead researchers to employ samples sizes based on the width of confidence intervals. Though apparently much simpler, such calculations can result in studies too small to achieve meaningful results.

Examination of precision may well be a useful adjunct to traditional estimation of sample size, but unless we place our primary emphasis on the question of power to detect an appropriate effect we could be making a serious mistake. The use of confidence intervals in analysis, however, must be encouraged, and this paper indicates how a realistic rewording of the usual specifications allows standard approaches to be used for calculations of sample size in a confidence interval framework.

There is no need to throw out our old sample size tables in this era of confidence intervals. In fact, we should guard them with care. Inadequate sample size has been a major problem in medical research, and we do not want to repeat those mistakes in the future. According to Altman: "However praiseworthy a study may be from other points of view, if the statistical aspects are substandard then the research will be unethical."[19] If we depart from the tried and tested approach for calculations of sample size we are in danger of disregarding this principle.

I thank the referees of this paper for valuable advice.

Appendix

This appendix gives the formulas on which the sample size calculations in this paper are based. The following notation is used, with significance, confidence, and power levels expressed as proportions

n = Sample size in each of the two groups
σ = Population standard deviation (assumed equal in the two groups)
Δ = Smallest worthwhile difference to be detected
z_k = 100kth percentile of normal distribution
α = Two tailed significance level
$1-\alpha$ = Confidence level
$1-\beta$ = Power of test
$\bar{x}_1, \bar{x}_2$ = Observed sample means

k	0·975	0·90	0·80	0·50
z_k	1·96	1·28	0·84	0·00

Sample size in each group for a two group comparison of a quantitative variable, specifying power $(1-\beta)$ to detect a minimum difference Δ (α two tailed significance level and population standard deviation of σ):

$$n \geq \frac{2\,(z_{1-\alpha/2}+z_{1-\beta})^2\,\sigma^2}{\Delta^2}$$

Expected confidence interval for the difference between means (using normal approximation):

$$\bar{x}_1-\bar{x}_2 \pm z_{1-\alpha/2}\,\sigma\sqrt{(2/n)}$$

(Note that the actual confidence interval for a given set of study data would use the sample standard deviation rather than σ and the appropriate critical value of the Student's t distribution rather than $z_{1-\alpha/2}$).

Expected width of confidence interval:

$$w = 2\,z_{1-\alpha/2}\,\sigma\sqrt{(2/n)}$$

Sample size in each group for a two group comparison of a quantitative variable, based on a required (expected) confidence interval width, w ($1-\alpha$ confidence level and population standard deviation of σ):

$$n \geq \frac{8\,(z_{1-\alpha/2}\,\sigma)^2}{w^2}$$

1 Langman MJS. Towards estimation and confidence intervals [Editorial]. *BMJ* 1986;**292**:716.
2 Anonymous. Report with confidence [Editorial]. *Lancet* 1987;ii:488.
3 Daly L. The statistical ring of confidence. *Ir Med J* 1989;**82**:49-50.

4 Gardner MJ, Altman DG. *Statistics with confidence*. London: *BMJ*, 1989.
5 Donner A. Approaches to sample size estimation in the design of clinical trials—a review. *Stat Med* 1984;3:199-214.
6 Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clin Trials* 1981;2:93-113.
7 Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med* 1986;5:1-13.
8 McHugh RB, Le CT. Confidence estimation and the size of a clinical trial. *Controlled Clin Trials* 1984;5:157-63.
9 O'Neill RT. Samples sizes for estimation of the odds ratio in unmatched case-control studies. *Am J Epidemiol* 1984;**120**:145-53.
10 Day SJ. Clinical trial numbers and confidence intervals of prespecified size. *Lancet* 1988;ii:1427.
11 Gordon I. Sample size estimation in occupational mortality studies with use of confidence interval theory. *Am J Epidemiol* 1987;**125**:158-62.
12 Lemeshow S, Hosmer DW, Klar J. Sample size requirements for studies estimating odds ratios or relative risks. *Stat Med* 1988;7:759-64.
13 Greenland S. On sample-size and power calculations for studies using confidence intervals. *Am J Epidemiol* 1988;**128**:231-7.
14 Bristol DR. Sample sizes for constructing confidence intervals and testing hypotheses. *Stat Med* 1989;**8**:803-11.
15 Beal SL. Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* 1989;**45**:969-77.
16 Grieve AP. Confidence intervals and trial sizes. *Lancet* 1989;i:337.
17 Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet* 1990;**335**:149-53.
18 Gardner MJ, Bond J. An exploratory study of statistical assessment of papers published in the Brtish Medical Journal. *JAMA* 1990;**263**:1355-7.
19 Altman DG. Statistics and ethics in medical research. In: Altman DG, Gore SM, eds. *Statistics in practice*. London: British Medical Association, 1982.

(Accepted 19 November 1990)

Medicine and the Media

BBC 1 *40 Minutes*: "Where There's Hope" 31 January

Real change?

Autism is one of the most bewildering and distressing of childhood disturbances; it is intellectually the most fascinating of medical disorders, challenging as it does what constitutes the essential nature of being human and bearing on the idea of its in part being our individual ability to entertain a theory of mind. Children with autism seem to lack such ability. This *40 Minutes* programme was about the despair experienced by many parents of children with autism and the draw of anything holding out hope.

The source of hope is the Higashi School in Massachusetts. It is a Japanese inspired and run school, though not all staff are Japanese. A basic tenet, it seems, is that it is the high state of anxiety in autism that blocks learning. Put in different words this is the tenet of many approaches to autism—that attempts to insist on the child interrelating (whether this be for learning or for socialisation) result in anxiety and avoidance. "Holding therapy" is one approach to the problem; low intensity persistent firmness is another. At this school, which clearly employs the second to the extremes of patience in the staff, a new approach is added: physical activities to reduce anxiety levels.

So it seems that the school's programme is a mixture of these two, about half the child's day being spent in sports and other physical activities and half in more formal learning (with a short time in the dormitories at night which is less structured). The atmosphere conveyed in the programme was not alarming —it seemed to be warm and caring, with staff showing great persistence and patience. Their input was enormous and one can understand the size of fees (£30 000 a year) as very high staffing levels must be needed. The school houses 88 children, all boarders.

The programme traced two British children, Ruth, aged 3, and Joseph, aged 8, in particular, with some reference to a third British child, John. We saw the enormous strain on the parents of Ruth and the distress that Joseph's behaviour caused to his parents. Insufficient detail and history were given to be able to say for sure whether they have classic primary autism, but their behaviours suggested this. We saw them at home, on arrival at the school and at the separation between parents and children, the children's involvement in some aspects of the programme at the school, and then the parents and children on reunion four months later.

The changes, as filmed by the crew, were impressive as was the obvious satisfaction of the parents. The children handled the reunion in a calm and socialised manner and the outing for a hamburger was remarkably contained. Clearly, considerable changes had been achieved in behaviour.

The question becomes one of whether this is purely an effect of behavioural training or whether there are some real changes in the core deficits. These core deficits can be seen as problems of relating to other people, of communication, and of imagination. Various studies have shown that it is possible to train children with autism into more socially acceptable behaviours—for example, toilet training, stopping tantrums, and getting children to sit and remain at a task. Children also come to behave in a socially more normal way with time. But the core problems have proved much more resistant to any real change. It was not possible from the way this programme was presented to assess whether there were any fundamental changes. The children behaved on greeting their parents as though they might actually be more interested and understanding of them, but this could have been an intensively rehearsed behaviour.


RACHAEL HOGG/BBC

Una Murphy and her son Joseph

This issue is important, and it was highly frustrating to hear that the British Autistic Society is about to publish its assessment of the school. I think it was wrong of the producers to put out this programme without a more "scientific" assessment of some sort as any parent of an autistic child watching the programme is likely to perceive the school as having achieved a miracle cure and will begin trying to raise the necessary money.

I suspect there is no magic other than the enormous and expert staff input into behavioural training. With the right staff training (to which no doubt the expertise learnt by staff at Higashi could contribute) and the necessary large resources this could be provided anywhere. It constitutes a far higher input to such children than anything currently provided in Britain. The issues raised are similar to those of using the Peto Institute in Hungary for brain damaged children: should society be providing on this level for certain children? Do the gains in the child's quality of life justify it?

Northwick Park Hospital, Harrow HA1 3UJ
Clare Sturge, MRCPSYCH, *consultant child psychiatrist*