PRACTICE

4

IN BRIEF

- The rationale underlying the choice of the optimal sample size in a clinical trial
- An explanation of Type I and Type II errors in hypothesis testing and their relevance to the significance level and power of the test
- A discussion of the factors that need to be considered when estimating the optimal sample size
- The use of Altman's nomogram to estimate the required sample sizes of two groups of observations which are to be compared.

Further statistics in dentistry Part 4: Clinical trials 2

A. Petrie[1] J. S. Bulman[2] and J. F. Osborn[3]

The principles which underlie a well-designed clinical trial were introduced in a previous paper.[1] The trial should be *controlled* (to ensure that the appropriate comparisons are made), *randomised* (to avoid allocation bias) and, preferably, *blinded* (to obviate assessment bias). However, taken in isolation, these concepts will not necessarily ensure that meaningful conclusions can be drawn from the study. It is essential that the sample size is large enough to enable the effects of interest to be estimated precisely, and to detect any real treatment differences.

FURTHER STATISTICS IN DENTISTRY:

1. Research designs 1
2. Research designs 2
3. Clinical trials 1
4. Clinical trials 2
5. Diagnostic tests for oral conditions
6. Multiple linear regression
7. Repeated measures
8. Systematic reviews and meta-analyses
9. Bayesian statistics
10. Sherlock Holmes, evidence and evidence-based dentistry

[1]Senior Lecturer in Statistics, Eastman Dental Institute for Oral Health Care Sciences, University College London; [2]Honorary Reader in Dental Public Health, Eastman Dental Institute for Oral Health Care Sciences, University College London; [3]Professor of Epidemiological Methods, University of Rome, La Sapienza
Correspondence to: Aviva Petrie, Senior Lecturer in Statistics, Biostatistics Unit, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London WC1X 8LD
E-mail: a.petrie@eastman.ucl.ac.uk

Refereed Paper

SAMPLE SIZE ESTIMATION

'How large a sample do I need?' is one of the most commonly asked questions of a statistician. It is also one of the hardest questions to answer. The researcher posing the query usually believes, quite wrongly, that the statistician can produce a figure, as if by magic, without any information about the why's and wherefore's of the trial. Unfortunately, both life and sample size estimation are not so simple! It is necessary to have some idea of the results that are expected from the trial, *before it has been conducted,* in order to evaluate the actual sample size required. If the proposed sample size appears outrageous, it is important to realise that if the numbers are reduced substantially, it may not be possible to detect real treatment differences, even if they exist. At the other extreme, if more patients than are really needed to compare treatments are used, the study may fall short of the ethical prerequisites.

TYPE I AND TYPE II ERRORS

The fundamental ideas of sample size estimation are most easily understood in the context of a trial to compare two arithmetic means using independent samples. The null hypothesis is that the true means in the populations from which the samples are derived are equal. An example is a clinical trial to compare the cariostatic action of two toothpastes in children of a given age; the children are to be randomly allocated the toothpastes and the mean dmfs increment observed after, say, 2 years will be compared in the two groups. Alternatively, a study might be designed to investigate whether the presence of fillings affects the level of *Streptococcus mutans* in the saliva; children recruited to the study will be divided into two groups not randomly, but according to the presence or absence of fillings, and the mean level of *S. mutans* observed in samples of saliva taken from the children will be compared. At the design stage of studies such as these, it will be necessary to know how many children to include in each sample.

The decision whether or not to reject the null hypothesis depends on the magnitude of the *P*-value obtained from the test and the cut-off value for it which determines significance, ie the **significance level.** Very often, although not necessarily, this level is chosen to be 0.05 so that the null hypothesis is rejected if the *P*-value is less than 0.05. If this is so, the result is said to be statistically significant and it is concluded that there is enough evidence to reject the null hypothesis. In the examples quoted, this would imply that there is evidence to suggest that one of the toothpastes is, on average, more cariostatic than the other, or that, on average, children with fillings tend to have higher (or lower) levels of *S. mutans* than children without fillings. Alternatively, if the *P*-value is greater than the cut-off level, there is not enough evidence to reject the null hypothesis, and the observed difference between the sample means is said to be not statistically significant at the chosen level. Note, however, that

this does not necessarily imply that the means in the populations of children are equal, only that there is insufficient evidence to show that these means are different.

It must be recognised that coming to either of these conclusions may or may not be correct. Rejecting the null hypothesis when it is true (concluding that there is evidence to show that the population means differ when, in fact, they are equal) leads to what is termed a **Type I error.** A **Type II error** is made when the null hypothesis is not rejected when it is false, ie when it is concluded that there is insufficient evidence to show that the population means differ when, in fact, these means are not equal. Table 1 summarises the consequences of rejecting and not rejecting the null hypothesis in the circumstances in which it is either true or false.

Table 1 Errors in testing the null hypothesis, H_0

	H_0 rejected	H_0 not rejected
H_0 false	No error	Type II error
H_0 true	Type I error	No error

Clearly both Type I and Type II errors are undesirable but because they arise as a consequence of sampling, and thus not having all information from the population available, the chances of making these errors cannot be entirely eliminated. The chances (ie probabilities) of making the Type I and Type II errors are usually denoted by the Greek letters, alpha (α) and beta (β), respectively. The aim in designing a study is to control α and β so that they are acceptable in the context of the proposed study. Since they both increase as the sample size of the study decreases, all other relevant factors remaining constant, choosing the optimal sample size becomes an integral part of study design.

The first step is to decide, in advance of collecting the data, on the worst case scenario for a Type I error. This means choosing the significance level of the test, the maximum value of α, and this is commonly but not necessarily assigned the value 0.05. Then the probability of incorrectly rejecting the null hypothesis cannot exceed 0.05 since H_0 is not rejected if $P > 0.05$. Now, if β is the probability of *not* rejecting the null hypothesis when it is false, then $(1-\beta)$ is the probability of rejecting it when it is false. $(1-\beta)$ is called the **power** of the test; it is the probability (often expressed as a percentage) of *correctly* concluding that a treatment difference of a specified size exists. It is usual to ensure that the study has a high power, very often in excess of 80%, since there is no justification for embarking on a study if it is known in advance that the study has little chance of detecting a real treatment effect. Having decided on values for the significance level and the power of the test, both of which will be chosen according to the circumstances of the investigation and the null hypothesis under test, it is then possible to evaluate the optimal sample size. However, as indicated in the following section, other factors apart from the power and the significance level come into play in the sample size determination.

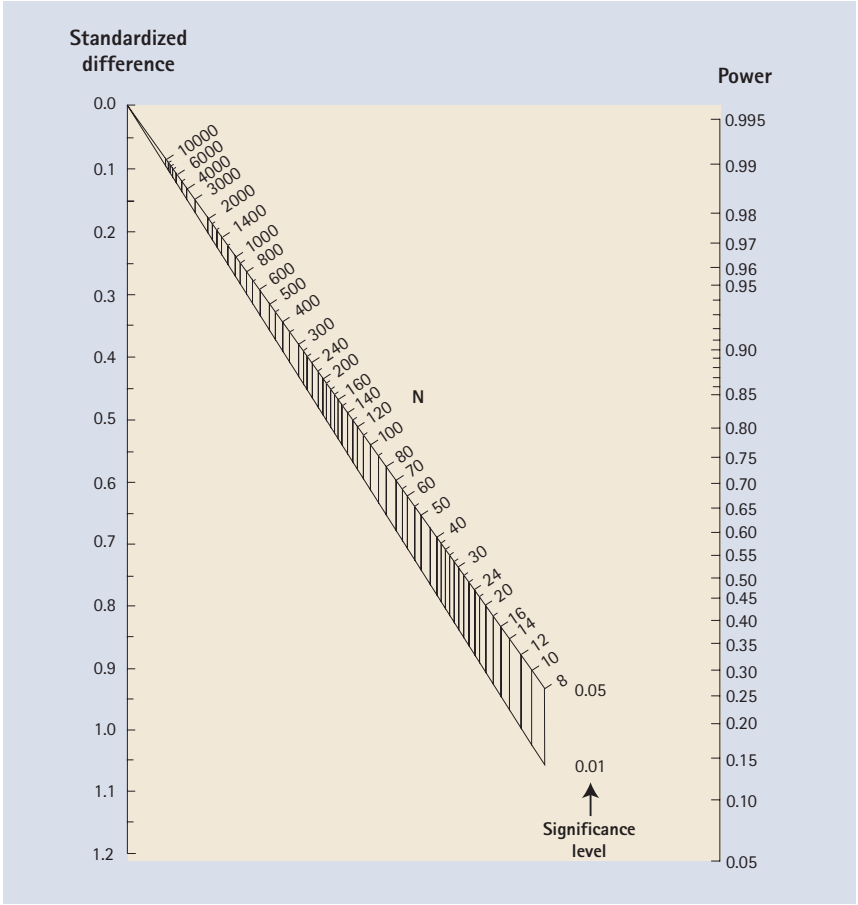


Fig. 1 Altman's nomogram for the calculation of sample size or power (extracted from Altman, 1982 How large a sample? in *Statistics in Practice*. Eds S. M. Gore and D. G. Altman. BMA London. Copyright BMJ Publishing Group, with permission.)

ALTMAN'S NOMOGRAM

There are various approaches one can use to determine the optimal sample size, each incorporating the same relevant factors into the calculations. It is usually the specification of these factors, provided in the bullet points in the following subsections, which creates the greatest difficulty in sample size calculations. Computer programs are available, for example *nQueryAdvisor,*[2] which produce useful tables and graphs. Specific formulae can be used to test different hypotheses, but these formulae tend to be cumbersome and their use time-consuming. Special tables exist for sample size calculations (Machin *et al.,* 1997)[3] but the disadvantage to this approach is the need for a separate table for each type of hypothesis. An alternative, and relatively simple approach (Altman, 1982),[4] is to use the nomogram shown in Figure 1.

The right-hand vertical axis of the nomogram represents different power values, ranging from 0.05 to 0.995. The left-hand vertical axis represents what is termed the standardized difference. This is a ratio which relates the difference of interest to the standard deviation of the observations. The exact form of the standardized difference varies according

to the nature of the variable under investigation and the specific hypothesis test. There are two axes within the nomogram, one for a significance level of 0.05, the other for 0.01, with total sample sizes indicated on each. The nomogram can be used to evaluate the optimal sample size once the power is specified, the significance level 5% or 1% is chosen, and the standardized difference is calculated. Alternatively, the procedure can be reversed, and the power of the study determined for a specified sample size. The nomogram is used under the assumption that equal sized samples are required, but the procedure can be modified to accommodate unequal sample sizes.

Sample size calculations for the comparison of two means from independent samples

Suppose that Altman's nomogram is to be used to estimate the optimal sample size for a trial in which the mean values of a single continuous variable using independent samples are to be compared. Typically, the data would be analysed by performing a *two sample t-test,* provided the data in each group are approximately Normally distributed and the observations in the two groups are *homoscedastic* (have the same variance). But first, it would be necessary to decide how many observations to include in each sample. Suppose it were decided to have equal sample sizes (n) in each group, with a total of $N = 2n$ observations.

In order to use the nomogram, the following factors must be specified:

- The **significance level** of the test: it is usually fixed at 0.05 or, occasionally, at 0.01, and a two-sided test adopted.
- The **power** of the test: this is usually required to be of the order of 80–90%.
- The assumed constant **variance** (σ^2), of the observations in each group. Inevitably, it is difficult to specify the variance of the observations before the data have been collected. However, since the variability of the observations has a direct bearing on sample size, some estimate of it must be obtained. The more variable the data, the larger the samples that are required to detect a real treatment difference of a specified size, if all other factors remain constant. It may be that a rough estimate of σ^2 can be obtained using the variance of the observations from a past experiment that has been performed which is similar in nature to that which is now planned. Perhaps the information can be found from published papers. If all else fails, it may be necessary to resort to a *pilot study* which is a small investigation, a small 'dress rehearsal' of the planned study, which may be used to provide an estimate of the variance.
- **The clinically important difference in the mean responses** (δ) which is considered to be so clinically or biologically important that if it were really to exist, it should be detectable by the proposed study. This is not the same as the difference in the mean responses which will be observed. It is a quantity that the investigator, not the statistician, must specify when he or she gives consideration to the consequences which may arise from the investigation. Note that it is easier to detect a large difference than a small one, so that the sample size is inversely proportional to δ.

In this particular problem of determining the optimal sample size to compare two means using the two sample t-test, the **standardized difference** is δ/σ, the clinically important difference divided by the assumed equal standard deviation of the observations in each group. This is the quantity on the left-hand vertical axis of the nomogram.

So, taking the first example in which two toothpastes are to be compared, suppose the investigator argues that if one toothpaste were to reduce the mean dmfs increment by 0.5 compared with the other, this would be regarded as a worthwhile treatment difference. There is an implication that if the true difference were less than 0.5, the investigator would not be too disappointed if the result were not statistically significant. The investigator must also obtain an estimate of the standard deviation, σ, of the dmfs increments. Researching the literature for other studies of the progression of dental caries in children, one obtains an estimate of the standard deviation, say σ =1.25 dmfs increment. Then the standardized difference is δ/σ = 0.5/1.25 = 0.4. If the investigator specifies that the level of significance to be adopted for the two-sided test is 0.05 and that the power should be 90% (often a rather arbitrary decision), the nomogram can be used to determine the total number of children required in the study. The line produced by connecting (using a ruler) the value of 0.4 for the standardized difference to the power value of 0.90 cuts the axis for a significance level of 0.05 at about N = 260 (Fig. 2). This indicates that there should be approximately n = 130 children in each toothpaste group. The investigator should then include in the protocol, paper or grant application a power statement such as *'it was decided to have 130 children in each of the two toothpaste groups in order to have a 90% chance of detecting a difference in mean dmfs increments of 0.5 at the 5% level of significance, assuming the standard deviation of dmfs increments to be about 1.25 in each of the groups'.*

It must be remembered that a sample size calculation can never be totally precise since the quantities used to calculate the sample size are often guessed or imprecisely estimated. The

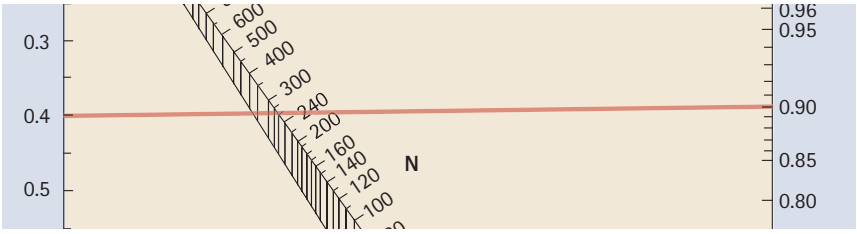


Fig. 2 The line produced by connecting the value of 0.4 for the standardized difference to the power value of 0.90 cuts the axis for a significance level of 0.05 at about N = 260 (a magnified section of Fig. 1 is shown)

Power, variability and sample size

Sample size is directly proportional to:

- **Power — the greater the power of the test, the larger the optimal sample size**
- **Variability — the more variable the observations, the larger the optimal sample size**

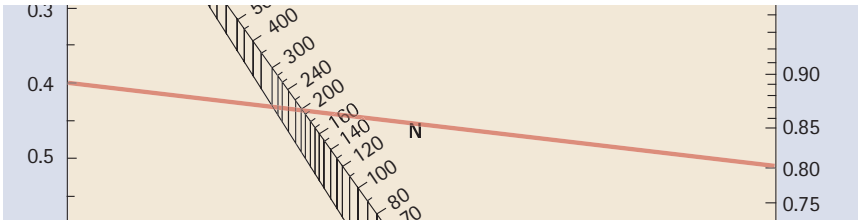


Fig. 3 The line produced by connecting the value of 0.4 for the standardized difference to the power value of 0.80 cuts the axis for a significance level of 0.05 at about N = 200 (a magnified section of Fig. 1 is shown)

aim of the calculation is to obtain a 'ball-park' figure for the sample size that is practically viable and results in a test which has sufficient power to detect a real and important treatment difference. The appeal of the nomogram is that it is easy to repeat the calculations after altering one or more of the quantities required to estimate the sample size. This is not to say it should be used deviously or to 'fiddle the figures'. As an illustration, suppose the sample size of 260 in the example quoted is unrealistic. What will be the effect on sample size of reducing the power specification from 90% to 80%? Again, using a ruler to connect the appropriate numbers in the nomogram, it can be seen that this would reduce the total sample size from 260 to about 200 (Fig. 3).

Clinically important treatment effect

It is harder to detect a small treatment effect than a large one so the optimal sample size is inversely proportional to the clinically important treatment effect

Sample size calculations for other comparisons

The use of the nomogram is fairly straightforward when it is necessary to compare two means from independent groups. The principles underlying the use of the nomogram remain the same for other types of experiment; essentially, it is the form of the standardized difference which changes. The two sided significance level (usually 0.05) and the power, usually between 80% and 90%, must still be specified for these calculations. The standardized differences that are required for different comparisons are indicated in the following bullet points.

- *Comparing two groups of paired numerical data* . The paired *t*-test is used to test the null hypothesis that the mean difference of a quantitative variable in two dependent or paired groups is zero. The standardized difference is

$$2\delta/\sigma_d$$

where:
δ is the clinically important difference.
σ_d is the standard deviation of the *differences*. This is much harder to estimate than the standard deviation of the individual observations. A pilot study is usually indicated.

The N in the nomogram represents the number of *pairs* of observations required in the experiment. If each pair represents two matched patients, the number of patients required is $2N$.

- *Comparing two proportions*. The *chi-square test* or, equivalently, a *standardized Normal deviate* is used to test the null hypothesis that the proportion of individuals possessing a certain attribute is the same in two groups. The response variable in this instance is a binary categorical variable. The standardized difference is

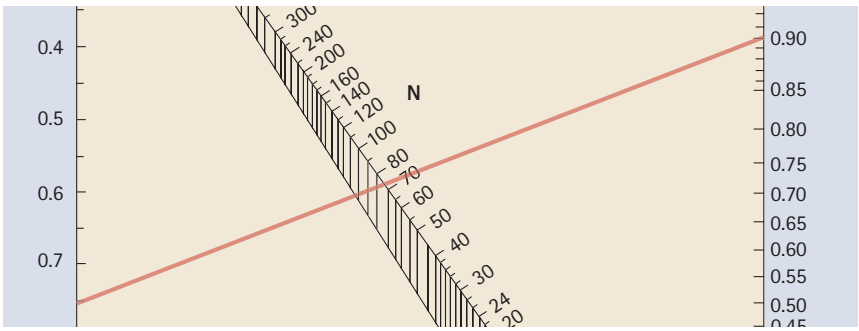
$$\frac{p_1 - p_2}{\sqrt{\{\bar{p}(1-\bar{p})\}}}$$

where:
$p_1 - p_2$ is the difference between the proportions of individuals with the attribute, that, if it really existed, would be considered clinically important.
$\bar{p}$ is the mean of p_1 and p_2.

The $N = 2n$ in the nomogram represents the total number of individuals required in the sample, with n in each group.

As an example, suppose it is of interest to establish the optimal sample size for a proposed study which will compare, in a given district, the dental fluorosis rates in children aged between 5 and 18 who have been either lifelong consumers of moderate- to high-fluoride surface water (≥ 0.50 mg F/L) or low-fluoride surface water (approximately 0.10 mg F/L). It is believed that the dental fluorosis rate in the low fluoride group is about 15%. A difference between the two groups in fluorosis rates of about 35% would be regarded as clinically important. In this case $\bar{p}$ =32.5% and the standardized difference is $(50-15)/\sqrt{(32.5 \times 67.5)} = 0.75$. Note that percentages have been used instead of proportions in this example, so the 'one' in the denominator of the standardized difference is replaced by '100'. Thus, by using the nomogram, it can be seen that in order to have a power of 90% of detecting a difference of 35% in fluorosis rates at the 5% level of significance about 70 children would be required (Fig. 4), with approximately 35 in each of the fluoride groups.

Fig. 4 The line produced by connecting the value of 0.75 for the standardized difference to the power value of 0.90 cuts the axis for a significance level of 0.05 at about N = 70 (a magnified section of Fig. 1 is shown)



SEQUENTIAL ANALYSIS

The methods discussed so far for determining the optimal sample size in an experiment relate to **fixed sample size** designs. It is assumed that the total sample size is a finite number which is fixed before the experiment is started. It is chosen in accordance with relevant power considerations, but also with reference to the expected patient accrual rate and the proposed time of investigation and costs.

As an alternative approach, the patients can

be entered *one at a time* into the clinical trial, and their responses, as they occur, can be used to test the hypothesis of interest. Either the trial is stopped in favour of one of the treatments when a significant treatment effect is observed; or it is stopped when it is considered that that no treatment difference is likely to arise. In both cases, the decision to stop is made with reference to a chart which is constructed by considering the significance level, the power and the size of the effect, all of which are specified at the outset. Clearly, in such a **sequential** trial (Armitage, 1975),[5] there is no need to estimate the patient numbers at the design stage, because the sample size depends on the results.

The advantage of a sequential trial is that it requires less patients than its fixed sample size counterpart if there is a large treatment effect. However, sequential trials are rarely performed, mostly because they are restricted to conditions in which there is only one response and when the time required to observe the response to treatment is not prolonged. Furthermore, it can be difficult to estimate the effect of interest and provide confidence intervals in a sequential study.

INTERIM ANALYSES

Sometimes clinical trials are designed so that the investigators can check the results at one or more *predefined* intermediate stages; these trials are often called **group sequential trials.** Apart from ensuring that the trial is running smoothly as regards compliance and that there is no concern about side-effects, the investigators may wish to perform significance tests to evaluate treatment effects at these times. Then, if one treatment is found to be superior, the trial can be stopped early and all the patients will go on to receive the most effective treatment.

Clearly, there are ethical advantages to this approach. However, be warned that such a proposal is not as straightforward as it might at first appear, and it is open to criticism if the statistical methods are not handled appropriately.

The problem with performing significance tests at intermediate stages is that the significance level at the end of the trial is larger than it would be if there were no repeated tests. In other words, there will be a greater chance of concluding that there is a significant difference between treatments when in reality there is no difference between them. Hence it is necessary to adjust the significance levels used for the intermediate or interim analyses to ensure that the final significance level is as expected, typically 0.05 or, perhaps, 0.01. Pocock (1983)[6] provides a table which shows, under certain conditions, which significance level to use at each intermediate stage (this is called the *nominal significance level*) if the significance level at the final stage (this is called the *overall significance level*) is 0.05 or 0.01. For example, if there were to be five repeated significance tests, the nominal level for each should be 0.016 (ie each repeated test is significant if $P < 0.016$) in order to have the overall significance level at 0.05.

1. Petrie A, Bulman J O, Osborn J F. Further Statistics in Dentistry. Part 3. Clinical Trials 1. *Br Dent J* 2002; **193:** 495-498.
2. *nQueryAdvisor Version 4.0* Statistical Solutions Ltd (2000) or *nQueryAdvisor Version 5.0* Statistical Solutions Ltd (Due in October 2002).
3. Machin D, Campbell M J, Fayers P M, Pinol A P Y. *Sample Size Tables for Clinical Studies.* 2nd edn. Oxford: Blackwell Science, 1997.
4. Altman D G. How large a sample? *Statistics in Practice* (eds Gore S.M. and Altman D.G.) British Medical Association, London, 1982
5. Armitage P. *Sequential Medical Trials.* 2nd edn. Oxford: Blackwell Scientific Publications, 1975.
6. Pocock S J. *Clinical Trials: a Practical Approach.* Wiley, Chichester, 1983.

Sequential analysis

The sample size in a sequential trial is not fixed in advance but depends on the results as they occur