Sample size in guidelines trials

JM Bland


In clinical trials, the statistical concepts of significance and power are used in the determination of sample size for trials. The trialist must provide an estimate of standard deviation and a hypothetical population difference to be detected. This must be modified to deal with the designs encountered in guideline research. These are cluster randomized trials, because the patients of a single doctor or practice form a cluster. The trialist must be able to provide information about the effects of clustering, in the form of an intraclass correlation coefficient.

Introduction

In the study of guidelines, the main function of randomized clinical trials is to evaluate the attempt to persuade service providers such as GPs to adopt the guideline. It is not to evaluate the guideline itself. If this is soundly based on good evidence, then we know without further trial that its implementation would be beneficial to patients. It follows that the appropriate unit of analysis of the trial is not the patient, but the doctor.

In this paper, I shall outline the statistical concepts of significance and power and the way in which they are used in the determination of sample size for trials. I shall then go on to show how this must be modified to deal with the designs encountered in guideline research. These are called cluster randomized trials, because the patients of a single doctor or practice form a single unit called a cluster.

As we shall see, approaches to design and analysis which ignore the clustering may be highly misleading.

Sample size for significance tests

In a typical clinical trial, we allocate our subjects into two groups at random. These groups then form two samples from the same population. We apply different treatments to the two samples. If the treatments have the same effect, we still have samples from the same population; if not, the samples are now from different populations. For example, we might randomize general practices into two groups and provide one group with guidelines. We measure the extent to which each practice conforms to the guidelines. We then use a significance test to test the null hypothesis that the groups are from the same population, i.e. that the treatment has had no effect.

A typical test of significance works like this. Suppose we have two populations with means $\mu_1$ and $\mu_2$, standard deviation $\sigma$, all unknown. We have two samples, each size $n$, with means $\bar{x}_1$ and $\bar{x}_2$, and standard deviation $s$. The details, including the effects of unequal sample sizes, unequal variances, small samples, etc., are given in many books.1–4

The difference between $\bar{x}_1$ and $\bar{x}_2$ would vary from sample to sample. If we were to do the experiment again, we would get different means. These might not differ in the same way; indeed, they might differ in the opposite direction. We want to know whether the difference in our sample is large enough for us to conclude that there is a difference in the whole population. This is the function of the test of significance.

For one such test, the large sample $z$ test, we calculate:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{2s^2/n}}$$

If there were no population difference, this would be an observation from a standard normal distribution. If its absolute value exceeds 1.96, the difference is significant at the 5% level.

The test is more likely to give a significant difference if there is a large difference between two populations rather than a small one. It is also more likely to detect a population difference of a given size (i.e. be significant) if the sample is large than if it is small. We call the probability that a test will produce a significant difference at a given significance level the power of the test. Power is related to the postulated difference in the population, the sample size, and the significance level ($\alpha = 0.05$).

Figure 1 shows the effect of hypothesized population difference and sample size on the power of a test. A simple formula connects the number in each group, the significance level $\alpha$, the power P, the hypothesized difference $\mu_1 - \mu_2$ and the variance $\sigma^2$:

$$n = \frac{f(\alpha, P) \times 2\sigma^2}{(\mu_1 - \mu_2)^2}$$

Department of Public Health Sciences, St George’s Hospital Medical School, London SW17 0RE, UK.
Here $f(\alpha, P)$ is a simple function of $\alpha$ and $P$ derived from the Normal distribution, tabulated in Table 1. The usual value used for $\alpha$ is 0.05, and $P$ is usually 0.80 or 0.90.

Thus, to determine sample size, we need to choose the power, $P$, and the significance level, $\alpha$, know the standard deviation, $s$, and decide the population difference to be detected, $m_1 - m_2$. The significance is usually set to 0.05 and the power to 0.90 or 0.80. Personally, I think that 0.80 is too low, but most funding organizations seem happy with it. Researchers are often unable to decide on the size of difference they wish to detect. I think that they often choose the number of subjects they think they can get then calculate the target difference from that. They are often unable even to provide the standard deviation, but a pilot study can usually discover this.

**Cluster randomized studies**

In a cluster randomized study, a group of subjects are randomized to the same treatment together. For example, we might randomize GPs to receive guidelines or not. The patients of the GP or of the whole practice form the cluster. An example is given in Table 2.5,6

The analysis must take the clustering into account. Ignoring it may make confidence intervals far too narrow and $P$-values too small, resulting in spurious significant differences. This is done far too often. This should not surprise us, as cluster randomization has been ignored almost completely in textbooks of medical statistics and in statistical articles in the medical literature. Only recently have medical statisticians begun to publish guidance on this.7-9 The first edition of *Statistical Tables for the Design of Clinical Trials*10 did not include them, but the second edition does.4 Several methods of analysis can be used. The simplest is to combine the data for the patients from one practice into a single summary statistic.6,8 The patients here tell us something about the practice, and the proportion of referrals from the practice which conform to the guidelines provides a good measure of the practice conformity. We can then carry out a two-sample $t$-test on summary statistics. In Table 2, the numbers of referrals in the clusters varies considerably. We can do a $t$-test weighted by cluster size to take this into account.11 We may need to use a transformation to make the data approximately normal. Another approach is multilevel modelling,12 which increases the complexity considerably. As the focus of the analysis is here on the practitioner rather than the patients, this is seldom necessary.

### Sample size in cluster randomized studies

The presence of clusters alters the calculation of the sample size.9 We now have two different sample sizes: the number of clusters (practices), $c$, and the number of subjects (patients) within a cluster, $m$. We also have two different variances: the variance between clusters $\sigma_c^2$, and the variance within a cluster, $\sigma_w^2$. The formula for sample size now gives us the number of clusters required and becomes:

$$c = \frac{f(\alpha, P) \times 2(\sigma_c^2 + \sigma_w^2)}{(m_1 - m_2)^2}$$

The total number of patients is $n = cm$. 

---

**Table 1** Values of $f(\alpha, P)$ for different $P$ and $\alpha$

<table>
<thead>
<tr>
<th>Power, $P$</th>
<th>Significance level, $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.05</td>
</tr>
<tr>
<td>0.50</td>
<td>3.8</td>
</tr>
<tr>
<td>0.70</td>
<td>6.2</td>
</tr>
<tr>
<td>0.80</td>
<td>7.9</td>
</tr>
<tr>
<td>0.90</td>
<td>10.5</td>
</tr>
<tr>
<td>0.95</td>
<td>15.2</td>
</tr>
<tr>
<td>0.99</td>
<td>18.4</td>
</tr>
</tbody>
</table>

**Figure 1** Effects of population difference and sample size on power for a $z$ test between two means.
The ratio of the total number of subjects required using cluster randomization to the number required using simple randomization is called the design effect (Deff). 

\[
\text{Deff} = \frac{m\sigma^2_s + \sigma^2_w}{\sigma^2_s + \sigma^2_w} 
\]

We can calculate the sample size as for a simply randomized (non-cluster) study and multiply it by Deff to get the number of subjects required for the cluster design.

It can be useful to present the design effect in terms of the intra-cluster correlation coefficient (ICC):

\[
\text{ICC} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_w} 
\]

This is the correlation which we expect between observations on pairs of subjects drawn one pair from each cluster. The design effect is then

\[
\text{Deff} = 1 + (m - 1) \text{ICC} 
\]

To estimate our sample size, we need an estimate not just of the variance of our measurement within clusters but also between cluster variation or the ICC. Although ICCs in cluster randomized trials are often small, typically <0.1, their effect cannot be ignored. For example, if ICC = 0.05 and the cluster size is \(m = 30\), then Deff = 2.45.

The number of patients required is more than twice that for a trial where patients were randomized individually.

Although the focus of a guidelines study must be on the service provider, we still need to consider the number of subjects in the cluster, the patients used to provide information about the provider. To do this, we need information on the two variances, in particular the variance between clusters, or on the ICC. This may be difficult to come by.

Acknowledgements

Thanks to my collaborator Sally Kerry for many helpful discussions and for supplying the data.

Discussion

As pointed out by Professor Bland, knowledge about the calculation of sample size for cluster randomized trials has only recently been featured in the non-specialist literature. Most of the participants in this workshop session were not statisticians, so a considerable part of the general discussion was taken up with clarification and further explanation of the paper’s content.

Lack of knowledge of intra-cluster correlation coefficients (ICCs) was identified as an issue of concern for both grant writers and funding bodies. Professor Bland pointed out that ICCs could be calculated from previous studies, but they are rarely published. Values typically ranged from 0.001 to 0.01.

One suggestion was that a register of ICCs should be created. Other suggestions were: (i) that the ICC could be measured in a pilot study; (ii) that its assumed value could be checked by an interim analysis and recruitment targets could be adjusted; and (iii) that a sensitivity analysis of how power depended on the ICC could be included in applications for funding.

The issue of how imbalance in cluster sizes could be addressed was raised. The commonest view was that it probably had to be ignored in calculations of sample size. Modelling of the impact of various degrees of imbalance on statistical power could be carried out if deemed necessary. In analyses of results, imbalance was easier to deal with. The contribution of each cluster to the overall result could be weighted by the cluster’s size.

The remainder of the general discussion covered aspects of study design and interpretation, which are dealt with in other papers in this supplement. Of note, however, was the point that multilevel analysis is only relevant if
the outcome of interest is a patient-level variable. If the outcome is a clinician variable, for example the propensity to refer to a specialist, the unit of analysis is the clinician not the patient, and issues such as the ICC are irrelevant.

References