The Stata Journal (2004)
4, Number 2, pp. 142–153

Sample size and power calculations using the noncentral t-distribution

David A. Harrison
ICNARC, London, UK
david@icnarc.org

Anthony R. Brady
ICNARC, London, UK
tony@icnarc.org

Abstract. The standard formulas for sample size and power calculation, as implemented in the command `sampsi`, make use of a normal approximation to the t-distribution. When the sample sizes are small, this approximation is poor, resulting in overestimating power (or underestimating sample size). One particular situation in which this is likely to be important is the field of cluster randomized trials. Although the total number of individuals in a cluster randomized trial may be large, the number of clusters will often be small. We present a simulation study from the design of a cluster randomized crossover trial that motivated this work and a command to perform more accurate sample size and power calculations based on the noncentral t-distribution.

Keywords: st0062, sampncti, sample size, power, noncentral t-distribution

1 Introduction

One-sample and two-sample t tests are frequently used in analyzing the results of clinical trials, comparing the mean responses in two groups or comparing a single group to a hypothesized value. The standard sample-size formulas found in statistical textbooks and implemented in the command `sampsi` assume that the t-distribution can be adequately approximated by a normal distribution. For small sample sizes, this assumption is poor, resulting in overestimation of the power or underestimation of the sample size. It has been suggested that the normal approximation is acceptable, provided the sample size in each arm is at least 30 (Lachin 1981). For smaller sample sizes, simple adjustments include adding one to the sample size if you are using a 5% significance level or adding two if you are using a 1% significance level (Snedecor and Cochran 1989, 104). To perform more accurate calculations of sample size and power requires using the noncentral t-distribution, which describes the distribution of the test statistic under the alternative hypothesis of unequal means.

Section 2 derives the formulas for sample size and power from the noncentral t-distribution, following the method of Chow, Shao, and Wang (2002), and section 3 illustrates the usefulness of this approach over a normal approximation in the context of designing a cluster randomized trial. Section 4 presents the syntax and usage of a command called `sampncti` to perform the calculations, section 5 gives two examples, and section 6 lists the results stored by `sampncti`.

2 Sample size and power calculations

2.1 One-sample test of mean

Suppose that we have a single sample, x_i, $i = 1, \ldots, n$, which we assume comes from a normal distribution with mean μ and standard deviation σ. We wish to test the hypothesis

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad \mathrm{H_a} : \mu \neq \mu_0$$

for some hypothesized value μ_0.

The standard parametric test for this situation is the one-sample t test. This test is based on the test statistic

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where $\overline{x}$ represents the sample mean and s the sample standard deviation.

Under the null hypothesis, T has a Student's t-distribution on $n - 1$ degrees of freedom. Under the alternative hypothesis, T has a noncentral t-distribution on $n - 1$ degrees of freedom with noncentrality parameter

$$\theta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

The power to detect a difference of $\delta = \mu - \mu_0$ with two-sided significance level α is given by

$$1 - \beta = T_{n-1}\left(t_{\alpha/2,n-1} \,\middle|\, \frac{\delta\sqrt{n}}{\sigma}\right) - T_{n-1}\left(-t_{\alpha/2,n-1} \,\middle|\, \frac{\delta\sqrt{n}}{\sigma}\right)$$

where $T_{df}(\,.\,|\theta)$ is the cumulative distribution function of the noncentral t-distribution with df degrees of freedom and noncentrality parameter θ and $t_{p,df}$ is the point of the central t-distribution with df degrees of freedom corresponding to an upper-tail probability of p.

Ignoring the smaller of the two terms (with value $< \alpha/2$), the power is approximately

$$1 - \beta = T_{n-1}\left(t_{\alpha/2,n-1} \,\middle|\, \frac{|\delta|\sqrt{n}}{\sigma}\right) \tag{1}$$

Hence, the sample size n can be obtained by solving (1) for given β.

For a one-sided test, the approximation in (1) is exact, with $\alpha/2$ replaced by the one-sided significance level α.

2.2 Two-sample test of equality of means

Now suppose that we have two independent samples, x_i, $i = 1, \ldots, n_1$, and y_j, $j = 1, \ldots, n_2$, and we assume that these come from normal distributions with means μ_1 and μ_2 and standard deviations σ_1 and σ_2, respectively.

We wish to test the following hypothesis:

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad \mathrm{H_a} : \mu_1 \neq \mu_2$$

The standard parametric test for this situation is the two-sample t test. If the variances are assumed to be equal ($\sigma_1 = \sigma_2$), as is usually the case when designing a clinical trial, the test is based on the statistic

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)s_x^2 + (n_2-1)s_y^2}{n_1+n_2-2}}}$$

where $\overline{x}$ and $\overline{y}$ represent the sample means and s_x and s_y the sample standard deviations for x and y, respectively.

If we do not assume equal variances in the two samples ($\sigma_1 \neq \sigma_2$), then the test statistic is given by

$$T = \frac{\overline{x} - \overline{y}}{\sqrt{s_x^2/n_1 + s_y^2/n_2}}$$

Under the null hypothesis, T has a Student's t-distribution on ν degrees of freedom. If $\sigma_1 = \sigma_2$, then $\nu = n_1 + n_2 - 2$. If $\sigma_1 \neq \sigma_2$, then ν can be approximated by Satterthwaite's formula (Satterthwaite 1946) as

$$\nu = \frac{\left(s_x^2/n_1 + s_y^2/n_2\right)^2}{\frac{(s_x^2/n_1)^2}{n_1-1} + \frac{\left(s_y^2/n_2\right)^2}{n_2-1}}$$

or by Welch's formula (Welch 1947) as

$$\nu = \frac{\left(s_x^2/n_1 + s_y^2/n_2\right)^2}{\frac{(s_x^2/n_1)^2}{n_1+1} + \frac{\left(s_y^2/n_2\right)^2}{n_2+1}} - 2$$

Under the alternative hypothesis, T has a noncentral t-distribution on ν degrees of freedom with noncentrality parameter

$$\theta = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

The power to detect a difference of $\delta = \mu_1 - \mu_2$ with two-sided significance level α is given by

$$1 - \beta = T_\nu\left(t_{\alpha/2,\nu} \left| \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right.\right) - T_\nu\left(-t_{\alpha/2,\nu} \left| \frac{\delta}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right.\right)$$

Ignoring the smaller of the two terms (with value $< \alpha/2$), the power is approximately

$$1 - \beta = T_\nu \left(t_{\alpha/2,\nu} \left| \frac{|\delta|}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right. \right) \tag{2}$$

Taking $r = n_2/n_1$ fixed, the sample size n_1 can be obtained by solving

$$T_\nu \left(t_{\alpha/2,\nu} \left| \frac{|\delta|\sqrt{n_1}}{\sqrt{\sigma_1^2 + r\sigma_2^2}} \right. \right) = 1 - \beta$$

For a one-sided test, the approximation in (2) is exact with $\alpha/2$ replaced by the one-sided significance level α.

3 Application to the design of a cluster randomized trial

One situation in which the inaccuracy of the normal approximation may be important is in the design of cluster randomized trials. While the total number of individuals in the trial may be large, it is the number of clusters that determines the degrees of freedom of the t-distribution, so if the number of clusters is small, then the normal approximation will be poor.

3.1 Setting

We were asked to design a study to investigate the impact of the introduction of a new health technology on the length of stay of patients in intensive care units (ICUs). Due to the nature of the intervention, it was considered necessary to randomize by cluster (ICU). It may be against best practice to encourage units to use two different types of equipment simultaneously (Department of Health 1998), and there may be potential for contamination if the presence of the new equipment changes staff practices when using the old equipment. A crossover design was considered to try to reduce the number of ICUs required for such a trial and to minimize any dropout that may have been caused by units being randomized to not receive the new technology.

3.2 Cluster randomized cluster crossover design

The study consists of two arms and is divided into two time periods. Clusters are randomly assigned to the two arms such that each arm contains c clusters. We will assume that, during each time period, every cluster recruits m individuals. During the first time period, individuals recruited to clusters in the second arm receive the experimental treatment, with individuals recruited to clusters in the first arm acting as controls (table 1). In the second time period, the clusters cross over, with individuals recruited to clusters in the first arm receiving the treatment and individuals recruited to clusters in the second arm acting as controls. Note that each individual receives only one

treatment. Thus when $m = 1$, the design *does not* reduce to the standard individually randomized crossover design. We use the term *cluster randomized cluster crossover* to distinguish from a cluster randomized individual crossover design in which randomization takes place at the cluster level but each individual receives both treatments in the randomly allocated order.

Table 1: Cluster randomized cluster crossover design with $c = 3$ clusters in each arm and $m = 5$ individuals recruited by each cluster in each time period (∘ control, • experimental)

		Period	
Arm	Cluster	$k = 1$	$k = 2$
$i = 1$	$j = 1$	∘ ∘ ∘ ∘ ∘	• • • • •
	$j = 2$	∘ ∘ ∘ ∘ ∘	• • • • •
	$j = 3$	∘ ∘ ∘ ∘ ∘	• • • • •
$i = 2$	$j = 1$	• • • • •	∘ ∘ ∘ ∘ ∘
	$j = 2$	• • • • •	∘ ∘ ∘ ∘ ∘
	$j = 3$	• • • • •	∘ ∘ ∘ ∘ ∘

We denote the continuous outcome measurement by Y_{ijkl} for individual l from time period k within cluster j of arm i ($i = 1, 2$; $j = 1, \ldots, c$; $k = 1, 2$; $l = 1, \ldots, m$). We will assume a mixed-effects model, with fixed effects for treatment and period and a random effect of cluster:

$$Y_{ijkl} = \mu + \pi_k + \tau X_{ik} + V_{ij} + \epsilon_{ijkl} \tag{3}$$

where

$$\begin{aligned}
\mu &= \text{mean for untreated subjects in the first time period} \\
\pi_k &= \text{fixed period effect; } \pi_1 = 0,\ \pi_2 = \pi \\
\tau &= \text{fixed treatment effect} \\
X_{ik} &= \text{design matrix; } X_{11} = X_{22} = 0,\ X_{12} = X_{21} = 1 \\
V_{ij} &\sim N\left(0, \sigma_A^2\right) = \text{random cluster effect} \\
\epsilon_{ijkl} &\sim N\left(0, \sigma_W^2\right) = \text{independent random error}
\end{aligned}$$

We have assumed that there is no carryover effect from one period to the next and no treatment by period interaction, and that cluster effects are the same in both periods. We have also assumed equal variances in the two arms of the trial.

For a cluster-level analysis of this design, we collapse the measurements to cluster means within each time period and apply the techniques for an individually randomized crossover design. We calculate the difference in outcomes between the first and second period $d_{ij} = \overline{Y}_{ij2\cdot} - \overline{Y}_{ij1\cdot}$ for each cluster and compare these between the two arms using a two-sample t test, observing that the expected value of the difference in means between the two arms will be 2τ as

$$d_{1j} = \overline{Y}_{1j2\cdot} - \overline{Y}_{1j1\cdot} \sim N\big(\pi + \tau, 2\sigma_W^2/m\big)$$
$$d_{2j} = \overline{Y}_{2j2\cdot} - \overline{Y}_{2j1\cdot} \sim N\big(\pi - \tau, 2\sigma_W^2/m\big)$$

Using the standard normal approximation, the power to detect a treatment effect of $\tau = \Delta$ is given by

$$\begin{aligned} 1 - \beta &= \Phi\left(\frac{|2\Delta|\sqrt{c/2}}{\sqrt{2\sigma_W^2/m}} - Z_{\alpha/2}\right) \\ &= \Phi\left(\frac{|\Delta|\sqrt{n/2}}{\sigma_W} - Z_{\alpha/2}\right) \end{aligned}$$

where $n = 2cm$ is the total number of individuals in each arm of the trial, $\Phi(\,.\,)$ is the cumulative distribution function of the standard normal distribution, and Z_p is the point of a standard normal distribution corresponding to an upper-tail probability of p. Using the normal approximation, the power is independent of the number of clusters, provided the total sample size is maintained.

Using the noncentral t-distribution, as in (2), the power is given by

$$\begin{aligned} 1 - \beta &= T_\nu\left(t_{\alpha/2,\nu} \,\middle|\, \frac{|2\Delta|\sqrt{c/2}}{\sqrt{2\sigma_W^2/m}}\right) \\ &= T_{2c-2}\left(t_{\alpha/2,2c-2} \,\middle|\, \frac{|\Delta|\sqrt{n/2}}{\sigma_W}\right) \end{aligned}$$

We now see that the power does indeed depend on the number of clusters through the degrees of freedom of the t-statistic. When the number of clusters is small, the normal approximation to the t-distribution will be poor, and this will be an important factor.

3.3 Simulation study

We illustrate the results above with a simulation study to investigate the empirical power of the cluster-level analysis for various numbers of clusters and cluster sizes compared with the estimated power from the normal approximation and using the noncentral t-distribution. Fixed parameter values used in the simulation are shown in table 2. The mean values and the within and among cluster standard deviations were chosen to be consistent with values from the ICNARC Case Mix Programme Database (Rowan and Black 2000) for an outcome variable of the length of stay in intensive care measured in days. The power values represent the power to detect (at the 5% significance level) a one-day reduction in intensive care stay for the given sample size and number of clusters using the cluster crossover design.

Table 2: Parameter values for simulation study.

Description	Parameter	Value
Mean control outcome	μ	4.4
Treatment effect	τ	-1
Period effect	π	0.3
Within cluster standard deviation	σ_W	7.9
Among cluster standard deviation	σ_A	1.3

One thousand datasets were simulated from the true model (3) for each combination of $c = 5$, 10, 25, and 50 clusters per arm and a total sample size of $n = 500$, 1000, 1500, and 2000. Expected power using the normal approximation was calculated by the `sampsi` command, and expected power using the noncentral t-distribution was calculated using the command `sampncti` described in the following sections. The observed power from each simulation is plotted against the expected power using the normal approximation in figure 1 and using the noncentral t-distribution in figure 2. The power was clearly overestimated by the normal approximation approach at low numbers of clusters. Observe how the expected power using the normal approximation is independent of the number of clusters (for a given total sample size), whereas the true power increases with increasing numbers of clusters.

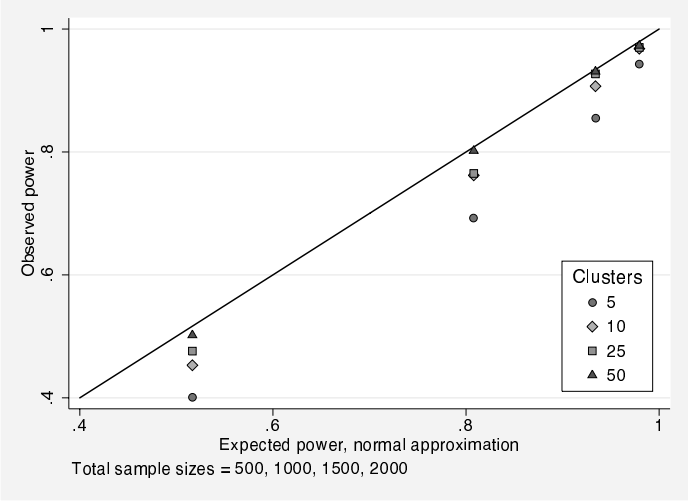

Figure 1: Observed power from simulation study plotted against expected power using the normal approximation

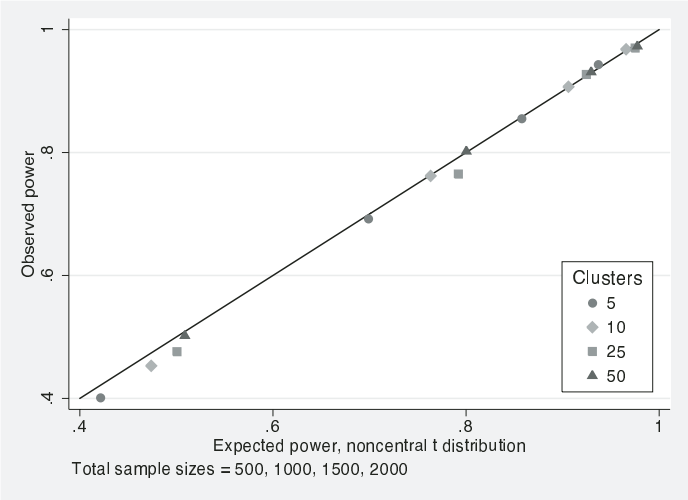

Figure 2: Observed power from simulation study plotted against expected power using the noncentral t-distribution

4 Stata implementation

4.1 Syntax

`sampncti` $\#_1$ $[\#_2]$, `sd1(#)` `[sd2(#) alpha(#) power(#) n1(#) n2(#) ratio(#) onesample onesided welch]`

4.2 Options

`sd1(#)` and `sd2(#)` are the standard deviations of population 1 and population 2, respectively. When the `onesample` option is used, `sd1(#)` is the standard deviation of the single sample (note that it can be abbreviated as `sd(#)`). If `sd2(#)` is not specified, `sampncti` assumes that `sd2` = `sd1`.

`alpha(#)` is the significance level of the test. The default is `alpha(0.05)` unless `set level` has been used to reset the default significance level for confidence intervals. If a `set level` $\#_{\text{lev}}$ command has been issued, the default value is `alpha(`$1-\#_{\text{lev}}/100$`)`. See [R] **level**.

`power(#)`$= 1-\beta$ is the power of the test. The default is `power(0.90)`.

`n1(#)` and `n2(#)` are the sizes of sample 1 and sample 2, respectively. One or both must be specified when computing power. If neither `n1(#)` nor `n2(#)` is specified, then `sampncti` computes sample size. When the `onesample` option is used, `n1(#)`

is the size of the single sample (note that it can be abbreviated as `n(#)`). If only one of `n1(#)` or `n2(#)` is specified, then the unspecified one is computed using the formula: `ratio` = `n2`/`n1`.

`ratio(#)` is the ratio of sample sizes for two-sample tests: `ratio` = `n2`/`n1`. The default is `ratio(1)`.

`onesample` indicates a one-sample test. The default is a two-sample test.

`onesided` indicates a one-sided test. The default is a two-sided test.

`welch` indicates that the approximate degrees of freedom for the t-distribution should be obtained from Welch's formula rather than from Satterthwaite's approximation formula, which is the default when `sd1` $\neq$ `sd2`. This option is not appropriate if `sd1` = `sd2` or if `onesample` is specified.

4.3 Remarks

`sampncti` estimates required sample size or power of tests for studies comparing means using the noncentral t-distribution, as described in section 2. If `n1(#)` and `n2(#)` are specified, `sampncti` computes power; otherwise, it computes sample size. The syntax and options for `sampncti` follow those of `sampsi` (see [R] **sampsi**) for the comparison of means. `sampncti` is an immediate command; all its arguments are numbers.

`sampncti` computes sample size or power for two types of tests:

1. Two-sample comparison of mean μ_1 of population 1 with mean μ_2 of population 2. The null hypothesis is $\mu_1 = \mu_2$, and normality is assumed. Either the postulated values of the means are specified as $\mu_1 = \#_1$ and $\mu_2 = \#_2$, or the difference in means is specified as $\delta = \mu_2 - \mu_1 = \#_1$ (and $\#_2$ is not given). The postulated standard deviations are `sd1(#)` and `sd2(#)`.

2. One-sample comparison of the mean μ of a population with a hypothesized value of μ_0. The null hypothesis is $\mu = \mu_0$, and normality is assumed. If two arguments are given to `sampncti`, the first argument $\#_1$ is μ_0, and the second argument $\#_2$ is the postulated value of μ; i.e., the null hypothesis is $\mu = \#_1$, and the alternative hypothesis is $\mu = \#_2$. If a single argument $\#_1$ is given, then this is the postulated deviation from the hypothesized value, $\delta = \mu - \mu_0$; i.e., the null hypothesis is $\delta = 0$, and the alternative hypothesis is $\delta = \#_1$. The postulated standard deviation is `sd1(#)`. To get this test, the `onesample` option must be given.

`sampncti` requires the package `nct` (Steichen 2000).

5 Examples

Two examples from the *Stata Base Reference Manual* for `sampsi` (see [R] **sampsi**) are repeated using `sampncti`.

5.1 Two-sample test of equality of means

Example

We calculate the required sample size to detect a difference between a mean of 132.86 with standard deviation 15.34 and a mean of 127.44 with standard deviation 18.23, using a ratio of 2:1, for $\alpha = 0.05$ (two-sided) and power of 0.80, as follows:

```
. sampncti 132.86 127.44, sd1(15.34) sd2(18.23) p(0.8) r(2)
Estimated sample size from noncentral t-distribution for
  two-sample comparison of means
Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:
         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =   132.86
            m2 =   127.44
           sd1 =    15.34
           sd2 =    18.23
         n2/n1 =     2.00
Satterthwaite's degrees of freedom:  251.8726
Estimated required sample size:
            n1 =      109
            n2 =      218
```

The sample sizes have increased by 1 and 2 from those obtained using `sampsi`.

To calculate the power for the same study in the event that each sample contains 100 subjects, we type

```
. sampncti 5.42, sd1(15.34) sd2(18.23) n1(100)
Estimated power from noncentral t-distribution for
  two-sample comparison of means
Test Ho: delta = 0, where delta is the difference in means
                      between the two arms
Assumptions:
         alpha =   0.0500  (two-sided)
         delta =     5.42
           sd1 =    15.34
           sd2 =    18.23
sample size n1 =      100
            n2 =      100
         n2/n1 =     1.00
Satterthwaite's degrees of freedom:  192.3805
Estimated power:
         power =   0.6193
```

Note that we have also used the alternative feature of `sampncti` of expressing the problem in terms of the difference between the means $\delta = 132.86 - 127.44 = 5.42$. The results of `sampncti` indicate that the power is slightly less than the figure of 0.6236 reported by `sampsi`.

◁

5.2 One-sample test of mean

▷ Example

We wish to calculate the sample size for a one-sample comparison to give 95% power to detect a reduced mean of −10 compared with the hypothesized value of 0. We use a one-sided test with $\alpha = 0.025$ and estimate that the standard deviation will be about 20. We type

```
. sampncti -10, sd(20) onesam a(0.025) onesided p(0.95)
Estimated sample size from noncentral t-distribution for
  one-sample comparison of mean to hypothesized value
Test Ho: delta = 0, where delta is the deviation from the
                      hypothesized value
Assumptions:
         alpha =   0.0250  (one-sided)
         power =   0.9500
         delta =      -10
            sd =       20
Estimated required sample size:
             n =       54
```

Using the noncentral t-distribution rather than a normal assumption has increased the required sample size from 52 to 54.

We now compute the power with a sample size of $n = 60$ and a one-sided significance level of $\alpha = 0.01$.

```
. sampncti 0 -10, sd(20) onesam a(0.01) onesided n(60)
Estimated power from noncentral t-distribution for
  one-sample comparison of mean to hypothesized value
Test Ho: m =      0, where m is the mean in the population
Assumptions:
         alpha =   0.0100  (one-sided)
 alternative m =      -10
            sd =       20
 sample size n =       60
Estimated power:
         power =   0.9274
```

`sampncti` reports a power of 0.9274 compared with 0.9390 from `sampsi`.

◁

The results from these two examples are very similar to those obtained from `sampsi`, as the sample sizes are sufficiently large for the normal approximation to be reliable. In situations such as these, it would be reasonable to use the results from `sampsi`. More extreme differences in power are observed with smaller sample sizes, as illustrated in section 3.

6 Saved Results

`sampncti` saves in `r()`:

Scalars

`r(N_1)`	sample size n_1	`r(power)`	power of the test
`r(N_2)`	sample size n_2	`r(df_t)`	degrees of freedom

7 References

Chow, S. C., J. Shao, and H. Wang. 2002. A note on sample size calculation for mean comparisons based on noncentral t-statistics. *Journal of Biopharmaceutical Statistics* 12(4): 441–456.

Department of Health. 1998. *Medical device and equipment management for hospital and community-based organisations*. London: Medical Device Agency. Bulletin MDA DB 9801.

Lachin, J. M. 1981. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 2(2): 93–113.

Rowan, K. and N. Black. 2000. A bottom-up approach to performance indicators through clinician networks. *Health Care UK* Spring 2000: 42–46.

Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2: 110–114.

Snedecor, G. W. and W. G. Cochran. 1989. *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press.

Steichen, T. 2000. *NCT: Stata modules related to the noncentral t distribution*. Boston College Department of Economics, Statistical Software Components S411901. Downloadable from *http://ideas.repec.org/c/boc/bocode/s411901.html*.

Welch, B. L. 1947. The generalization of "Student's" problem when several different population variances are involved. *Biometrika* 34: 28–35.

About the Authors

David Harrison is Statistician at ICNARC (Intensive Care National Audit & Research Centre).

Tony Brady is Senior Statistician at ICNARC.