The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2014)
14, Number 1, pp. 159–175

Sample size and power calculations for trials and quasi-experimental studies with clustering

Evridiki Batistatou
Institute of Population Health
Centre for Epidemiology
University of Manchester
Manchester, UK
evridiki.batistatou@manchester.ac.uk

Chris Roberts
Institute of Population Health
Centre for Biostatistics
University of Manchester
Manchester, UK
chris.roberts@manchester.ac.uk

Steve Roberts
Institute of Population Health
Centre for Biostatistics
University of Manchester
Manchester, UK
steve.roberts@manchester.ac.uk

Abstract. This article considers the estimation of power and sample size in experimental and quasi-experimental intervention studies, where there is clustering of subjects within one or both intervention arms, for both continuous and binary outcomes. A new command, `clsampsi`, which has a wide range of options, calculates the power and sample size needed (that is, the number of clusters and cluster size) by using the noncentral F distribution as described by Moser, Stevens, and Watts (1989, *Communications in Statistics—Theory and Methods* 18: 3963–3975). For comparative purposes, this command can also produce power and sample-size estimates on the basis of existing methods that use a normal approximation.

Keywords: st0329, clsampsi, sample size, power calculation, intervention studies

1 Introduction

Clusters are often naturally occurring groups, including, for example, schools, work sites, hospitals, or general practices. Trials involving randomizing clusters of individuals to interventions are commonly called cluster randomized trials. The implication of cluster randomization for the design and analysis of clinical trials is widely recognized (Donner and Klar 2000; Hayes and Moulton 2009). Similarity among responses for subjects in the same cluster contradicts the assumption of independence upon which standard methods of statistical analysis are based; the resulting dependency leads to loss of precision and reduced power when estimating treatment effects (Murray 1998; Donner and Klar 2000; Hayes and Moulton 2009). Clusters may also be formed for the purpose of treatment. Examples of such clusters include exercise classes for the treatment of musculoskeletal disorders (Hayden et al. 2005), group therapies for psychological problems (Hunot et al. 2007; Bisson and Andrew 2007), and self-help groups for smoking cessation (Stead and Lancaster 2005).

As mentioned earlier, an important feature of trials with clustering is that the outcome of interest for individuals within a cluster tends to be more alike because they share the same delivery of treatment and because of interaction between individuals. The degree to which responses within the same cluster are similar can be expressed by the intracluster correlation coefficient (ICC), which is based on the relation of the between- to within-cluster variance of outcome. The magnitude of the clustering effect—often called the design effect (DE)—depends on both the cluster size and the ICC (Donner and Klar 2000). Application of a conventional sample-size estimation formula that does not account for the DE may lead to sample-size estimates that are too small, resulting in trials that are inadequately powered.

The implications of clustering effects due to the organization of care are now receiving greater recognition. Recent guidance for reporting trials on nonpharmacological interventions has highlighted the need to consider the effects of clustering by care provider, because variation in outcome by care provider has the same implications for estimating sample size and power as cluster randomized trials (Boutron et al. 2008). In nonpharmacological trials involving a health professional activity, such as surgery, counseling, or physical therapy, the success of the intervention could depend upon the characteristics of the therapist.[1] Where treatment and care depend on substantial skill or training, it is realistic to assume variation in average outcome between therapists. This variation, which implies lack of independence of patient outcomes within therapist and, hence, intracluster correlation by therapist, has implications for the design of trials of such interventions. Clustering may also be an issue in quasi-experimental studies (Cook and Campbell 1979) used across a wide range of situations, including health and social policy, in which comparisons are made between interventions delivered to units using observational data.

In trials of group-administered interventions in both trial arms (fully nested design), both the cluster size and the ICC may differ between trial arms (Roberts and Roberts 2005). Alternatively, a group-administered therapy may also be compared with an individual therapy, resulting in no clustering effect in one arm; here the trial will consist of partially nested data. In either design, the outcome variation in the two trial arms may be unequal, often leading to substantial heteroskedasticity.

1. The term "therapist" is used to refer to the health professional providing therapy, irrespective of discipline or specialty.

Power and sample size in trials with clustering are generally based on a summary analysis that estimates the number of higher-level units or a calculation of a DE as a multiplier of the sample size of subjects calculated ignoring clustering (Donner and Klar 2000). While the standard sample size and power formula implemented in the `power` command (`sampsi` for versions 12 and earlier) and `sampncti` command (Harrison and Brady 2004) could be used by considering cluster level analysis, this requires estimating the summary level variance or the DE from the posited estimate of the ICC and the cluster size. Only the `sampclus` command (Garrett 2001)—acting as an adjustment to `sampsi` (but not straightforward)—takes the DE into account. However, this command is applicable only in situations where the clustering effect, in terms of both the cluster size and the intracluster correlation, is assumed to be the same for both treatments.

This article presents `clsampsi`—a power and sample-size calculator designed for cluster randomized trials with continuous or binary outcomes—which is suitable for studies with homogeneous and heterogeneous patterns of clustering (with either fully or partially nested data). The command calculates the power for a given set of design parameters. It can also calculate the sample size of a prespecified number of clusters or number of subjects per cluster for a given power. Options for calculating sample size and power for design parameters other than those prespecified are also available.

2 Statistical analysis

Consider a fully nested data model for a continuous outcome y for the jth observation pertaining to the jth subject in the ith cluster,

$$y_{ij} = \mu + \delta I_T + \gamma x_{ij} + u_i + \epsilon_{ij} \tag{1}$$

where μ is the mean outcome, x is a matrix of baseline covariates, γ is the vector effect of baseline covariate x, I_T is an indicator variable for the treatment arm ($I_T = 1$ for the group-therapy arm; $I_T = 0$ for the control-therapy arm), δ is the treatment effect, u_i is the random effect for cluster variation with $u_i \sim N(0, \sigma_u^2)$, and ϵ_{ij} is the random effect for subject variation within the cluster with $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$. Intracluster variability is measured by the ICC (see introduction) defined by $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$. This model assumes that all individuals are nested within clusters in both therapy arms.

A special case of (1) is the partially nested data model. Consider, for example, a partially nested individually randomized trial, in which a group therapy is compared with a control therapy with no clustering effect in the control arm, following the regression model

$$y_{ij} = \mu + \delta I_T + \gamma x_{ij} + I_T u_i + I_T \epsilon_{ij} + (1 - I_T) r_{ij} \tag{2}$$

where r_{ij} is the random effect for subject variation in the control therapy, with $r_{ij} \sim N(0, \sigma_r^2)$; σ_u^2, σ_ϵ^2, and σ_r^2 are the variances of u, ϵ, and r, respectively. In the control-therapy arm of (2), each subject could be treated as a cluster of size one. Also the total variance $\sigma_u^2 + \sigma_\epsilon^2$ in the group-therapy arm is not necessarily equal to the variance σ_r^2 in the control arm, which results in between-arm heteroskedasticity.

Models (1) and (2) can be fit in standard multilevel software using maximum likelihood or restricted maximum-likelihood (REML) estimation methods such as `mixed` or `gllamm`. Roberts and Roberts (2005) showed that using REML for the analysis of partially nested data models may work well for test size (power was not explored in their article); however, the Satterthwaite t test is the best choice for unadjusted analysis. They also noted the need to consider heteroskedasticity—now possible in Stata (with the `mixed, residuals()` option).

3 Power and sample-size calculations

3.1 Asymptotic methods

For a known number of clusters and cluster size in each trial arm, the `clsampsi` command calculates the power for a given set of design parameters. In this article, the effect on power when a covariate is in the analysis model [see (1) and (2)] is not addressed. The large-sample formula for computing the power level $1-\beta$ to detect an unstandardized treatment effect δ in a two-sided test with significance level α is given by

$$1-\beta = \Phi\left(\frac{\delta}{\left[\frac{\sigma_G^2}{N_G}\{1+(m_G-1)\rho_G\}+\frac{\sigma_C^2}{N_C}\{1+(m_C-1)\rho_C\}\right]^{1/2}} - z_{\alpha/2}\right) \tag{3}$$

where σ_G^2, N_G, ρ_G and σ_C^2, N_C, ρ_C are the variance outcome, the total sample size, and the ICC in each trial arm (G: group, C: control), respectively, and $z_{\alpha/2}$ is the $100\times(\alpha/2)$ percent standard normal variate. The variance of the treatment effect $\widehat{\delta} = \overline{y}_G - \overline{y}_C$, where $\overline{y}_G$ and $\overline{y}_C$ are the mean outcomes in the two trial arms, is given by

$$\text{var}\left(\widehat{\delta}\right) = \frac{\sigma_G^2}{N_G}\{1+(m_G-1)\rho_G\}+\frac{\sigma_C^2}{N_C}\{1+(m_C-1)\rho_C\} \tag{4}$$

For a known group size m and power, Donner and Klar (2000) give the following asymptotic formula derived for the required number of clusters k in each trial arm,

$$k = \frac{2(z_{\alpha/2}+z_\beta)^2}{\delta^2} \times \frac{\sigma^2\{1+(m-1)\rho\}}{m}$$

where σ^2, m, and ρ are the variance outcome, the cluster size, and the ICC in each trial arm, and $z_{\alpha/2}$ and z_β are the $100\times(\alpha/2)$ and $100\times\beta$ percent standard normal variates.

For a given sample size, power can be maximized by changing the allocation ratio between trial arms (Roberts and Roberts 2005):

$$R = \frac{k_G m_G}{k_C m_C} = \sqrt{\frac{\sigma_G^2\{1+(m_G-1)\rho_G\}}{\sigma_C^2\{1+(m_C-1)\rho_C\}}} \tag{5}$$

For an unknown number of clusters or cluster sizes, `clsampsi`—using the sample-size ratio N_C/N_G and either the ratio of m_C/m_G or the ratio of k_C/k_G, respectively—calculates the starting values of the unknown parameters via the asymptotic approximation. These values are then modified (that is, increased) until the desired power is reached. Now a noninteger solution is achieved. In unclustered sample-size calculations, one can usually calculate the sample size that gives power close to that required for the trial. In cluster randomized trials, however, this becomes more approximate as a calculation: fixing either the cluster size or the number of clusters before estimating the other may give only a rough approximate to the desired power. In trials with a more heterogeneous pattern, the problem becomes yet more complex.

3.2 Small-sample size and unequal variance methods

The above approximation assumes asymptotic normality, which is suitable only with large numbers of clusters. Otherwise, sample size will be underestimated, and power will be overestimated in small trials. One option, which is implemented in the user-written command `sampncti` (Harrison and Brady 2004), is to replace normal deviates by the corresponding centiles of the t and noncentral t distribution and then use degrees of freedom based on the Satterthwaite (1946) approximation, which in this circumstance is given by the expression

$$\upsilon = \frac{\left[\dfrac{\sigma_G^2\{1+(m_G-1)\rho_G\}}{k_G m_G} + \dfrac{\sigma_C^2\{1+(m_C-1)\rho_C\}}{k_C m_C}\right]^2}{\dfrac{\left[\dfrac{\sigma_G^2\{1+(m_G-1)\rho_G\}}{k_G m_G}\right]^2}{(k_G-1)} + \dfrac{\left[\dfrac{\sigma_C^2\{1+(m_C-1)\rho_C\}}{k_C m_C}\right]^2}{(k_C-1)}}$$

A further complication is the presence of heteroskedasticity due to differences in intracluster correlation or cluster size between arms, an extreme case of which occurs in a partially nested design. Considering a summary measures analysis, Hoover (2002) suggested an approximate method based on the central and noncentral t distribution (Disantostefano and Muller 1995). Unfortunately, this approximation performs poorly when a smaller group size with smaller variance is compared with a larger group size with a larger variance, which is the situation when group therapy is compared with an individual therapy in a summary measures analysis of a partially nested design. Moser, Stevens, and Watts (1989) described an exact numerical method for the estimation of power for this test involving integration of the noncentral F distribution. We have therefore implemented the exact method described by Moser, Stevens, and Watts (1989). This method is exact only when all clusters within each arm are of equal size; otherwise, the noncentral F itself is only an approximation. Also this method slightly overestimates or underestimates power depending on both the sample-size ratio and the outcome variance ratio between arms (more details are shown in Moser, Stevens, and Watts [1989]).

As an option, the command we present here also allows an asymptotic approximation using the `sampsi` command and the noncentral t distribution method implemented in the user-written command `sampncti` (Harrison and Brady 2004); table 1 compares the three options.

Table 1. Power calculation for continuous data using `sampsi`, `sampncti`, and `clsampsi` for varying allocation ratio N_C/N_G (N_C: total sample size in individual-therapy arm and N_G: total sample size in group-therapy arm), number of clusters k, and $\rho = 0.05$ and 0.2. N_G was held constant at 100; for each N_C/N_G value, 1st row: $k = 5$, $m = 20$; 2nd row: $k = 10$, $m = 10$; 3rd row: $k = 20$, $m = 5$. Here the treatment effect (δ) is equal to 0.5, and the outcome variance is equal to 1 in both trial arms (no heteroskedasticity).

N_C/N_G	Integration of noncentral F (default)		Noncentral t (`sampncti` option)		Normal (`sampsi` option)	
	$\rho = 0.05$	$\rho = 0.2$	$\rho = 0.05$	$\rho = 0.2$	$\rho = 0.05$	$\rho = 0.2$
0.25	0.494	0.343	0.500	0.342	0.536	0.392
	0.546	0.456	0.547	0.459	0.572	0.483
	0.569	0.526	0.569	0.527	0.592	0.546
1	0.730	0.426	0.737	0.410	0.829	0.546
	0.864	0.674	0.866	0.675	0.891	0.727
	0.912	0.831	0.912	0.832	0.921	0.848
4	0.767	0.425	0.771	0.424	0.921	0.605
	0.939	0.738	0.941	0.739	0.970	0.817
	0.979	0.918	0.980	0.918	0.986	0.937

As table 1 shows, even in the case of no heteroskedasticity, power calculated by `clsampsi` is always lower—especially for a small number of clusters—than the one using the normal approximation (assumed by `sampsi`) and is equal to or slightly lower than the one using the noncentral t distribution (assumed by `sampncti`); only for a small number of clusters ($k = 5$) and high ICC ($\rho = 0.2$) is power slightly higher by `clsampsi` when compared with `sampncti`.

To estimate the sample size of a cluster, the command uses the estimated number of clusters based on the normal approximation as a starting value. The number of clusters is then increased until the required power is obtained. Numbers of clusters in each arm are increased to maintain the desired ratio of subjects between treatment arms, which will be 1:1 unless the ratio option is specified. To estimate cluster size, the command obtains starting values based on the normal approximation, which generally gives a noninteger value for the cluster size. The command therefore rounds down the value before increasing the cluster size until a combination with the required power is obtained. Cluster sizes in each arm are increased to maintain the desired ratio of subjects between treatment arms. The option `minimum()` uses the value given by (5) as a starting value before carrying out a numerical search to estimate the combination of cluster sizes in each arm that give the minimum number of subjects for the specified cluster sizes.

3.3 Unequal cluster size

In all the above formulas, $m_i = m$ is assumed for all clusters $i = 1, \ldots k$. However, equal cluster sizes are rarely encountered in practice, and because an imbalance in cluster size can influence the power of the trial (Kerry and Bland 2001), it should be taken into account for the sample-size calculation. The high influence of severe imbalance in cluster size on power when a small number of clusters and high ICC are present was also shown by Guittet, Ravaud, and Giraudeau (2006), who investigated the effect of cluster-size variation on power in trials with continuous outcomes. A number of authors have given expression for the DE where subjects are weighted equally. To deal with a potential imbalance in cluster size, Kerry and Bland (2001) proposed adjusting the DE, assuming a cluster-level analysis on the linear scale weighting by cluster size; this adjustment is applicable to both continuous and binary outcomes.

Following this proposal, `clsampsi` implements the DE as (see also a formula for a distribution-based correction given by Guittet, Ravaud, and Giraudeau [2006])

$$\text{DE} = 1 + \left(\overline{m} + \frac{\sigma_m^2}{\overline{m}} - 1 \right) \rho$$

where σ_m^2 (`varm()` in `clsampsi`) is the variation in cluster size. If σ is the individual-level standard deviation, then

$$\sigma \sqrt{\frac{\text{DE}}{\overline{m}}}$$

is used in `clsampsi` to estimate the variance of the cluster level required to determine power or sample size using the Moser, Stevens, and Watts (1989) method or a normal or t approximation. Note that power determined using an analysis based on summary measures taking account of variance in cluster size, assuming clusters are weighted according to their size, will tend to be conservative when compared with REML, maximum likelihood, and generalized estimating equations estimations that implicitly weight clusters using minimum variance weight (van Breukelen, Candel, and Berger 2007).

4 Binary outcomes

`clsampsi` implements the formula discussed earlier for both continuous and binary outcomes. For example, with binary data, (3) becomes

$$1 - \beta = \Phi \left(\frac{\delta}{\left[\frac{p_G(1-p_G)}{N_G} \{1 + (m_G - 1)\rho_G\} + \frac{p_C(1-p_C)}{N_C} \{1 + (m_C - 1)\rho_C\} \right]^{1/2}} - z_{\alpha/2} \right)$$

where p_G and p_C are the probabilities of an outcome $y_{ij} = 1$ in the grouping and control arm, respectively. Expression (4) is given by

$$\text{var}\left(\widehat{\delta}\right) = \frac{p_G(1-p_G)\{1 + (m_G - 1)\rho_G\}}{m_G k_G} + \frac{p_C(1-p_C)\{1 + (m_C - 1)\rho_C\}}{m_C k_C}$$

For partially nested data, this expression simplifies to the following (Moerbeek, van Breukelen, and Berger 2000):

$$\text{var}\left(\widehat{\delta}\right) = \frac{p_G(1-p_G)\{1+(m_G-1)\rho_G\}}{m_G k_G} + \frac{p_C(1-p_C)}{k_C}$$

For continuous data, power using `clsampsi` is determined assuming a model analysis based on summary measures; for binary data, cluster proportions are used.

4.1 Coefficient of variation: Another measure of the between-cluster variability

Earlier, we discussed the ICC as a measure of the between-cluster variability. Another measure of this variability is the coefficient of variation (CV) (Hayes and Moulton 2009). For continuous outcomes, the CV is given by $\text{CV} = \sigma_u/\mu$, where σ_u and μ represent the between-cluster standard deviation and the mean of the continuous parameter, respectively. When the outcome is continuous, the ICC ρ may be a more appropriate measure to use to correct for clustering because it includes the within-cluster variability while the CV does not. For binary data, the CV is given by $\text{CV} = \sigma_u/p$, where p is the probability of an outcome equal to 1. In the latter case, Thomson, Hayes, and Cousens (2009) show that a simple expression relating ρ and CV exists:

$$\rho = \frac{\text{CV}^2 p}{(1-p)} \tag{6}$$

Thus, when the $\text{CV} = 0$, then $\rho = 0$. Also because $\rho < 1$, an upper bound for CV is

$$\text{CV} < \sqrt{\frac{(1-p)}{p}}$$

For continuous outcomes, there is no simple relationship, similar to (6), between ρ and the CV. As noted by Thomson, Hayes, and Cousens (2009), if we assume a constant CV across trial arms, a different estimate of the between-cluster variance in the intervention arm is taken from that found by assuming a constant ρ across trial arms. As a result, different sample-size estimates will be estimated. The `clsampsi` command implements the straightforward relation (6) for binary data only, given a fixed ρ or CV.

4.2 arcsin (angular) transformations for binary outcomes

When analyzing a clinical trial with binary outcomes, say, p_G and p_C (definitions were given earlier), we frequently use the identity link to measure the treatment effect $\delta = p_G - p_C$. However, other links, such as the `arcsin` link (Cochran and Cox 1957), may transform the above binary outcomes in a way that improves the stability of the variance of the outcome. The power of a test of difference between p_G and p_C using the `arcsin` approximation is given by

$$z_\beta = z_{1-\alpha} - \sqrt{2N}(\texttt{arcsin}\sqrt{p_G} - \texttt{arcsin}\sqrt{p_C})$$

where $1-\beta$ is the desired power level, α is the significance level, and N is the sample size (here we assume that the sample size of each trial arm is N).

This transformation stretches out both tails of the distribution of proportions and compresses the middle. For this reason, one could argue that there is no clear benefit of using this transformation for proportions, say, 0.3–0.7, that may be more likely to occur in practice. Moreover, as suggested by Dobson and Gebski (1986), the `arcsin` transformation underestimates the sample size or, equivalently, overestimates the power of a test. The `arcsin` transformation has been implemented in `clsampsi` for estimating power and sample sizes via the `arcsin` option.

4.3 Using the generalized estimating equation or the logistic random-intercept model

Analysis using the generalized estimating equation or the logistic random-intercept model estimates the log odds-ratio. Suppose $\psi_1 = \log_e\{p_1/(1-p_1)\}$ and $\psi_2 = \log_e\{p_2/(1-p_2)\}$ are the log odds of the binary outcomes p_1 and p_2, respectively. One might estimate power (implemented by the `logodds` option in `clsampsi`) by

$$1-\beta = 1-\Phi\left\{z_{(1-\alpha/2)} - \frac{\psi_G-\psi_C}{\mathrm{SE}(\psi_G-\psi_C)}\right\}$$

Using the delta method approximation for the standard error (SE) of the log odds gives

$$\mathrm{SE}(\psi_2-\psi_1) = \sqrt{\frac{1+(m_1-1)\rho_1}{N_1p_1(1-p_1)} + \frac{1+(m_2-1)\rho_2}{N_2p_2(1-p_2)}}$$

5 The clsampsi command

5.1 Syntax

`clsampsi` *#1 #2* [`, sd1(#) sd2(#) sd(#) k1(#) k2(#) k(#) m1(#) m2(#) m(#) varm1(#) varm2(#) varm(#) rho1(#) rho2(#) rho(#) cv1(#) cv2(#) cv(#) ni(#) alpha(#) power(#) onesided ratio(#) rangek1(#) rangek2(#) rangek(#) minimum(#) maxm1(#) maxm2(#) maxm(#) sampsi sampncti arcsin logodds`]

When continuous data are used, the means *#1* and *#2* of sample 1 and sample 2 have to be specified with their standard deviations. When binary data are used, proportions *#1* and *#2* have to be between 0 and 1; `sd1(#)` and `sd2(#)` have to be omitted.

5.2 Options

`sd1(#)` specifies the standard deviation of sample 1. If `sd1()` is specified without `sd2()`, the standard deviation of sample 2 is set equal to the standard deviation of sample 1.

`sd2(#)` specifies the standard deviation of sample 2. If `sd2()` is specified without `sd1()`, the standard deviation of sample 1 is set equal to the standard deviation of sample 2.

`sd(#)` specifies the standard deviation of sample 1 and sample 2 assuming equal standard deviations in both samples.

`k1(#)` specifies the number of clusters in sample 1 (it has to be at least 2; otherwise, the program cannot run). If `k1()` is specified without `k2()`, the number of clusters in sample 2 is set equal to the number of clusters in sample 1.

`k2(#)` specifies the number of clusters in sample 2 (it has to be at least 2; otherwise, the program cannot run). If `k2()` is specified without `k1()`, the number of clusters in sample 1 is set equal to the number of clusters in sample 2.

`k(#)` specifies the number of clusters in sample 1 and sample 2 assuming equal number of clusters in both samples.

`m1(#)` specifies the cluster size in sample 1. If `m1()` is specified without `m2()`, the cluster size in sample 2 is set equal to the cluster size in sample 1.

`m2(#)` specifies the cluster size in sample 2. If `m2()` is specified without `m1()`, the cluster size in sample 1 is set equal to the cluster size in sample 2.

`m(#)` specifies the cluster size in sample 1 and sample 2 assuming equal cluster size in both samples.

`varm1(#)` specifies the cluster-size variation in sample 1. If neither `varm1()` nor `varm2()` is specified, the default is `varm1(0)`. If `varm1()` is specified without `varm2()`, the cluster-size variation in sample 2 is set equal to the cluster-size variation in sample 1.

`varm2(#)` specifies the cluster-size variation in sample 2. If neither `varm1()` nor `varm2()` is specified, the default is `varm2(0)`. If `varm2()` is specified without `varm1()`, the cluster-size variation in sample 1 is set equal to the cluster-size variation in sample 2.

`varm(#)` specifies the cluster-size variation in sample 1 and sample 2 assuming equal cluster-size variation in both samples. To have a noticeable effect on power when using the `varm(#)` option, `varm(#)` needs to be larger than `m(#)` (and `rho(#)` > 0).

`rho1(#)` specifies the intraclass correlation coefficient in sample 1. If neither `rho1()` nor `rho2()` is specified, the default is `rho1(0)`. If `rho1()` is specified without `rho2()` and the cluster size of sample 2 equals 1, then `clsampsi` assumes the intraclass correlation coefficient of sample 2 is 0.

`rho2(#)` specifies the intraclass correlation coefficient in sample 2. If neither `rho1()` nor `rho2()` is specified, the default is `rho2(0)`. If `rho2()` is specified without `rho1()` and the cluster size of sample 1 equals 1, then `clsampsi` assumes the intraclass correlation coefficient of sample 1 is 0.

`rho(#)` specifies the intraclass correlation coefficient in sample 1 and sample 2 assuming an equal intraclass correlation coefficient in both samples.

`cv1(#)` specifies the coefficient of variation of outcome in sample 1. If neither `cv1()` nor `cv2()` is specified, the default is `cv1(0)`. If `cv1()` is specified without `cv2()`, the coefficient of variation in sample 2 is set equal to the coefficient of variation in sample 1.

`cv2(#)` specifies the coefficient of variation of outcome in sample 2. If neither `cv1()` nor `cv2()` is specified, the default is `cv2(0)`. If `cv2()` is specified without `cv1()`, the coefficient of variation in sample 1 is set equal to the coefficient of variation in sample 2.

`cv(#)` specifies the coefficient of variation of outcome in sample 1 and sample 2 assuming an equal coefficient of variation of outcome in both samples. The `cv(#)` option is available here only for testing proportions; there is no simple relation between `rho(#)` and `cv(#)` for testing means.

`ni(#)` specifies the sample size for integrating the noncentral F distribution. The default is `ni(10000)`.

`alpha(#)` specifies the significance level of the test. The default is `alpha(.05)`.

`power(#)` specifies the power of the test. The default is `power(.90)`.

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

`ratio(#)` specifies the allocation ratio between sample 2 and sample 1 ($= N2/N1$, where $N1$ and $N2$ are the total sample sizes of sample 1 and sample 2, respectively).

`rangek1(#)` adds $(\# - 1)$ clusters to the prespecified `k1(#)` number of clusters.

`rangek2(#)` adds $(\# - 1)$ clusters to the prespecified `k2(#)` number of clusters.

`rangek(#)` adds $(\# - 1)$ clusters to the prespecified `k(#)` number of clusters.

`minimum(#)` determines the minimum sample size of subjects required to achieve the specified `power(#)` for given `m1()` and `m2()`.

`maxm1(#)` specifies the maximum cluster size desired in sample 1 when estimating cluster sizes. The default is `maxm1(8/rho(#))`.

`maxm2(#)` specifies the maximum cluster size desired in sample 2 when estimating cluster sizes. The default is `maxm2(8/rho(#))`.

`maxm(#)` specifies the maximum cluster size desired in sample 1 and sample 2 when estimating cluster sizes assuming equal cluster size in both samples; the default is `maxm(8/rho(#))`.

`sampsi` determines the power of a summary comparison of means or proportions using the standard `sampsi` command.

`sampncti` determines the power using the `sampncti` command (Harrison and Brady 2004) and the `nct2` program (Steichen 2000).

`arcsin` determines the power for the comparison of two proportions (only) with angular transformation, which improves the stability of the variance of the outcome. When the `arcsin` option is used, power is estimated using the Satterthwaite approximate F test (default). The `arcsin` plus `sampsi` options and `arcsin` plus `sampncti` options are also available using the z approximation (`sampsi`) and the noncentral t distribution (`sampncti`) for comparison of proportions, respectively.

`logodds` determines the power for the comparison of two proportions (only) using the generalized estimating equation or the logistic random-intercept model estimate of the log odds-ratio. When the `logodds` option is used, power is estimated using the Satterthwaite approximate F test (default). The `logodds` plus `sampsi` options and `logodds` plus `sampncti` options are also available using the z approximation (`sampsi`) and the noncentral t distribution (`sampncti`) for comparison of proportions, respectively.

5.3 Stored results

`clsampsi` stores the following in `r()`:

Scalars

`r(k1)`	number of clusters in sample 1
`r(k2)`	number of clusters in sample 2
`r(m1)`	cluster size in sample 1
`r(m2)`	cluster size in sample 2
`r(N1)`	total sample size in sample 1
`r(N2)`	total sample size in sample 2
`r(N)`	total sample size in both samples
`r(rho1)`	ICC in sample 1
`r(rho2)`	ICC in sample 2
`r(varm1)`	cluster-size variation in sample 1
`r(varm2)`	cluster-size variation in sample 2
`r(estpower)`	estimated power
`r(delta)`	difference between the two (continuous or binary) outcomes
`r(ratiom)`	ratio of `m2()/m1()`
`r(ratiok)`	ratio of `k2()/k1()`
`r(ratioN)`	ratio of $N2/N1$

6 Low back pain trial: Example

A randomized controlled trial was planned to compare two interventions for the treatment of chronic lower back pain. A group-administered treatment comprising active exercise and education delivered by physiotherapists was compared with information on self-management; a detailed description of the study is given by Johnson et al. (2007). Based on previous data among persons consulting with back pain, a within-treatment standard deviation of six Roland–Morris disability questionnaire points was assumed.

Assuming that the sample size in the group-therapy arm is a multiple of group size $m = 5$, we wish to calculate the number of clusters required to have 90% power to detect a minimum clinically important difference of three points in the Roland–Morris disability questionnaire (Roland and Fairbank 2000) using a two-tailed test and 5% significance level. There is no prior information on the ICC for the group-therapy arm; we use $\rho = 0.05$ as a prior guess.

`clsampsi` suggests that 19 groups of 5 people in the group-therapy arm and 98 people in the individual-therapy arm are required to achieve 90% power to detect the prespecified treatment effect.

```
. clsampsi 3 0, sd1(6) sd2(6)  m1(5) m2(1) rho1(0.05) rho2(0)
Calculating Number of Clusters for Specified Cluster Size(s) and Power

 Design with N2/N1 approximately 1

Estimated power/sample size using the Satterthwaite approximate F test for
two-sample comparison of means with clustering
Test Ho: mu1 = mu2, where mu1 is the mean in population 1
                     and mu2 is the mean in population 2
Assumptions:       alpha =   0.0500  (two-sided)
                             Sample 1      Sample 2
               Mean (mu)         3             0
      Total St. Dev.(sd)         6             6
   Number of Clusters (k)        19            98
         Cluster Size (m)         5             1
  Cluster Size Var.(varm)         0             0
          Sample Size (N)        95            98
Intra-Cluster Corr. (rho)       .05             0
       SD (summary level)   2.93939             6
                 Total Sample Size:   193
           Allocation ratio (N2/N1): 1.03
Ratio of Number of clusters (k2/k1): 5.16
     Ratio of Cluster sizes (m2/m1):    .2
 Satterthwaite´s degrees of freedom:  52.47
   Sample size (ni) for integration:  10000
                  Estimated power: 0.9006
```

In this example, we assumed that all the groups in the group-therapy arm are of equal size. For unequal groups, however, as the variation of the mean group size increases, more groups will be needed depending on the ICC. Suppose that the distribution of cluster size has Poisson distribution with mean μ_m; then the variance in cluster size is also μ_m. Assuming variance in cluster size is equal to 5 in the above example (`varm1(5)`),

the total sample size in both trial arms increases from 193 to 200 subjects, as shown below. Assuming `varm1(25)`, the total sample size required in both trial arms increases from 193 to 214 (results not shown here).

```
. clsampsi 3 0, sd1(6) sd2(6) m1(5) varm1(5) m2(1) rho1(0.05) rho2(0)
varm2 must be specified; varm2 is set equal to varm1
Calculating Number of Clusters for Specified Cluster Size(s) and Power

 Design with N2/N1 approximately 1

Estimated power/sample size using the Satterthwaite approximate F test for
two-sample comparison of means with clustering
Test Ho: mu1 = mu2, where mu1 is the mean in population 1
                     and mu2 is the mean in population 2
Assumptions:       alpha =   0.0500  (two-sided)
                             Sample 1      Sample 2
                Mean (mu)         3             0
       Total St. Dev.(sd)         6             6
   Number of Clusters (k)        20           100
         Cluster Size (m)         5             1
  Cluster Size Var.(varm)         5             5
          Sample Size (N)       100           100
Intra-Cluster Corr. (rho)       .05             0
       SD (summary level)     2.997             6
                 Total Sample Size:   200
           Allocation ratio (N2/N1):    1
Ratio of Number of clusters (k2/k1):    5
     Ratio of Cluster sizes (m2/m1):    .2
 Satterthwaite´s degrees of freedom:  54.90
   Sample size (ni) for integration:  10000
                  Estimated power: 0.9056
```

Assuming clusters of equal size (`varm(0)`), increasing the 19 groups by two additional groups will also increase power as suggested below by `clsampsi`.

```
. clsampsi 3 0, sd1(6) sd2(6) k1(19) k2(98) m1(5) m2(1) rho1(0.05) rho2(0)
> rangek1(3)

calculate power for specified number of clusters k and cluster sizes m

Estimated power/sample size using the Satterthwaite approximate F test for
two-sample comparison of means with clustering

Test Ho: mu1 = mu2, where mu1 is the mean in population 1
                    and mu2 is the mean in population 2

Assumptions:       alpha =   0.0500  (two-sided)

                             Sample 1      Sample 2
               Mean (mu)          3             0
      Total St. Dev.(sd)          6             6
   Number of Clusters (k)        19            98
         Cluster Size (m)         5             1
  Cluster Size Var.(varm)         0             0
          Sample Size (N)        95            98
Intra-Cluster Corr. (rho)       .05             0
       SD (summary level)   2.93939             6

                 Total Sample Size:   193
           Allocation ratio (N2/N1): 1.03
Ratio of Number of clusters (k2/k1): 5.16
     Ratio of Cluster sizes (m2/m1):    .2
 Satterthwaite´s degrees of freedom:  52.47
   Sample size (ni) for integration:  10000

                 Estimated power: 0.9006

                 Power for increasing combinations of k1 and k2

  k1     k2     k2/k1  N1     N2     N2/N1  N      Power
  19     98     5.158  95     98     1.032  193    0.9006
  20     98     4.900  100    98     0.980  198    0.9093
  21     98     4.667  105    98     0.933  203    0.9168
```

7 Acknowledgment

The authors' research was supported by a UK Medical Research Council Methodology Research Grant (G0800606).

8 References

Bisson, J., and M. Andrew. 2007. Psychological treatment of post-traumatic stress disorder (PTSD). *Cochrane Database of Systematic Reviews* 18: CD003388.

Boutron, I., D. Moher, D. G. Altman, K. F. Schulz, and P. Ravaud. 2008. Methods and processes of the CONSORT Group: Example of an extension for trials assessing nonpharmacologic treatments. *Annals of Internal Medicine* 148: W60–W66.

Cochran, W. G., and G. M. Cox. 1957. *Experimental Designs*. 2nd ed. New York: Wiley.

Cook, T. D., and D. T. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Disantostefano, R. L., and K. E. Muller. 1995. A comparison of power approximations for Satterthwaite's test. *Communications in Statistics—Simulation and Computation* 24: 583–593.

Dobson, A. J., and V. J. Gebski. 1986. Sample sizes for comparing two independent proportions using the continuity-corrected arc sine transformation. *Journal of the Royal Statistical Society, Series D* 35: 51–53.

Donner, A., and N. Klar. 2000. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Garrett, J. M. 2001. sxd4: Sample size estimation for cluster designed samples. *Stata Technical Bulletin* 60: 41–45. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 387–393. College Station, TX: Stata Press.

Guittet, L., P. Ravaud, and B. Giraudeau. 2006. Planning a cluster randomized trial with unequal cluster sizes: Practical issues involving continuous outcomes. *BMC Medical Research Methodology* 6: 17.

Harrison, D. A., and A. R. Brady. 2004. Sample size and power calculations using the noncentral t-distribution. *Stata Journal* 4: 142–153.

Hayden, J., M. W. van Tulder, A. Malmivaara, and B. W. Koes. 2005. Exercise therapy for treatment of non-specific low back pain. *Cochrane Database of Systematic Reviews* 3: CD000335.

Hayes, R. J., and L. H. Moulton. 2009. *Cluster Randomised Trials*. Boca Raton, FL: Chapman & Hall/CRC.

Hoover, D. R. 2002. Clinical trials of behavioural interventions with heterogeneous teaching subgroup effects. *Statistics in Medicine* 21: 1351–1364.

Hunot, V., R. Churchill, V. Teixeira, and M. Silva de Lima. 2007. Psychological therapies for generalised anxiety disorder. *Cochrane Database of Systematic Reviews* 1: CD001848.

Johnson, R. E., G. T. Jones, N. J. Wiles, C. Chaddock, R. G. Potter, C. Roberts, D. P. Symmons, P. J. Watson, D. J. Torgerson, and G. J. Macfarlane. 2007. Active exercise, education, and cognitive behavioral therapy for persistent disabling low back pain: A randomized controlled trial. *Spine* 32: 1578–1585.

Kerry, S. M., and J. M. Bland. 2001. Unequal cluster sizes for trials in English and Welsh general practice: Implications for sample size calculations. *Statistics in Medicine* 15: 377–390.

Moerbeek, M., G. J. P. van Breukelen, and M. P. F. Berger. 2000. Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics* 25: 271–284.

Moser, B. K., G. R. Stevens, and C. L. Watts. 1989. The two-sample t test versus Satterthwaite's approximate F test. *Communications in Statistics—Theory and Methods* 18: 3963–3975.

Murray, D. M. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.

Roberts, C., and S. A. Roberts. 2005. Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials* 2: 152–162.

Roland, M., and J. Fairbank. 2000. The Roland–Morris disability questionnaire and the Oswestry disability questionnaire. *Spine* 15: 3115–3124.

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

Stead, L. F., and T. Lancaster. 2005. Group behaviour therapy programmes for smoking cessation. *Cochrane Database of Systematic Reviews* 2: CD001007.

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

Thomson, A., R. Hayes, and S. Cousens. 2009. Measures of between-cluster variability in cluster randomized trials with binary outcomes. *Statistics in Medicine* 28: 1739–1751.

van Breukelen, G. J. P., M. J. J. M. Candel, and M. P. F. Berger. 2007. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine* 26: 2589–2603.

About the authors

Eva Batistatou is a lecturer in epidemiology and biostatistics in the Centre for Epidemiology, Institute of Population and Health at the University of Manchester, Manchester, UK.

Chris Roberts is a professor of biostatistics in the Centre for Biostatistics, Institute of Population and Health at the University of Manchester, Manchester, UK.

Stephen Roberts is a senior lecturer in medical statistics in the Centre for Biostatistics, Institute of Population and Health at the University of Manchester, Manchester, UK.