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 university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
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 (2013)
13, Number 1, pp. 114–135

A menu-driven facility for sample-size calculations in cluster randomized controlled trials

Karla Hemming
University of Birmingham
Birmingham, UK
k.hemming@bham.ac.uk

Jen Marsh
University of Birmingham
Birmingham, UK
j.l.marsh@bham.ac.uk

Abstract. We introduce the Stata menu-driven command `clustersampsi`, which calculates sample sizes, detectable differences, and power for cluster randomized controlled trials. The command permits continuous, binary, and rate outcomes (with normal approximations) for comparisons of two-sided tests in two equal-sized arms. The command allows for specification of the number of clusters available, or the cluster size, or the average cluster size along with an estimate of the variation of cluster sizes. When the number of clusters available is insufficient to detect the required difference at the prespecified power, `clustersampsi` will return the minimum number of clusters required under the prespecified design along with the minimum detectable difference and maximum achievable power (both for the prespecified number of clusters). Cluster heterogeneity can be parameterized by using either the intracluster correlation or the coefficient of variation. The command is illustrated via examples.

Keywords: st0286, clustersampsi, sample size, cluster randomized controlled trials, minimum detectable difference, maximum achievable power

1 Introduction

Sample-size calculations are frequently undertaken for cluster randomized controlled trials (RCTs). This is usually done by prespecifying the average cluster size, obtaining the sample size required under individual randomization, and inflating by the design effect (DE), which is a simple function of the intracluster correlation (ICC) (Donner and Klar 2000). Alternatively, heterogeneity between clusters can be parameterized by the coefficient of variation (standard deviation or mean) of the outcome and similar two-step procedures (Hayes and Bennett 1999). However, these two-step procedures are sometimes not efficient (for example, when many calculations are required) and sometimes not quite so straightforward. The reasons are outlined below.

Cluster sample-size calculations are not completely straightforward in a number of situations. Complexity arises in cases when the user prespecifies the number of clusters available (as opposed to the average cluster size); when the user requires a power or detectable difference calculation (as opposed to a sample-size calculation); and particularly when the calculation involves binary outcomes. This is because the conventional inflation by the DE is only useful when the user specifies the cluster size

and needs to obtain an estimate of the number of clusters needed. When the user specifies the number of clusters available and needs to obtain an estimate of the cluster size, the inflation over that which is required under individual randomization depends on the very quantity the user is trying to compute, the cluster size.

Additionally, because limited precision sets in as the cluster sizes increase, some designs will be infeasible (Guittet, Giraudeau, and Ravaud 2005). That is, irrespective of how large the clusters are made, a fixed number of available clusters might mean there is insufficient power to detect the required difference. When the objective is to calculate power or detectable difference under cluster RCT designs of fixed sample sizes for continuous outcomes, the user can use the simple relationships that exist between those power and detectable differences obtainable under individual randomization and those obtainable under cluster randomization (Hemming et al. 2011). To obtain an estimate of the detectable difference for binary outcomes where the variance depends on the proportion, the user must solve a quadratic equation. This is also the case for the computation of detectable differences for continuous outcomes when the cluster heterogeneity is parameterized by the coefficient of variation.

Currently, several options are available to Stata users planning a cluster RCT. The `sampsi` command may be used to estimate the required number of clusters (for both binary and continuous outcomes) via a two-step procedure that involves calculating the sample size under individual randomization and inflating this by a self-computed DE. To estimate power for continuous outcomes, the user could also use `sampsi` after inflating the estimated standard deviation by the DE. For cluster designs, `sampsi` cannot be used to estimate detectable differences, power for binary outcomes, or the number of clusters required.

Another two-step method consists of using the `sampclus` command (Garrett 2001), which again requires the user to calculate the sample size required under individual randomization immediately before implementing the command. With `sampclus`, the user is permitted to specify either the number of clusters available or the cluster size and the command returns, whichever is not specified. In cases where the number of clusters available is insufficient to detect the required difference at the prespecified power level, the user is alerted and informed of the minimum number of clusters required. `sampclus` does not compute power (to detect a prespecified difference for a fixed sample size) or detectable difference (to detect a prespecified power for a fixed sample size).

The command `clsampsi` (Batistatou and Roberts 2010) was developed primarily for designs with differential clustering between arms. Differential clustering occurs, for example, when the individuals in the intervention arm are grouped (say, group therapy) but there is no grouping in the control arm. While `clsampsi` does offer a single-step procedure that calculates both the power (for a prespecified difference for a fixed sample size) and the sample size (either the number of clusters or the cluster size), it does not compute the detectable difference and does not alert the user to infeasible designs.

Currently, none of these commands allows computation of the detectable difference, nor do they allow specification of heterogeneity parameterized by the coefficient of variation. In addition, none of these commands allows for varying cluster sizes, repeated measures, or adjustment for covariates. All three of these issues can have important implications on power and so should be considered at the design stage.

In summary, while estimation of sample-size variables for cluster RCTs is not excessively complex, it would be useful to directly compute these quantities in Stata. Currently, there are two options available for Stata users: using Stata's built-in commands in two-step routines where the user modifies either the sample-size computed by Stata or modifies the input variables (say, standard deviation) to account for the clustering; or using a user-written Stata command (`clsampsi`), which is limited to a very specific study design. We have therefore developed a Stata command, `clustersampsi`, that we believe will be very practical for applied health care researchers involved in the design on cluster RCTs.

2 The clustersampsi command

The new Stata command `clustersampsi` computes power, sample size (both the number of clusters and the cluster size), and detectable difference (for both fixed and varying cluster sizes), and it alerts the user to infeasible designs (due to an insufficient number of clusters).

When the design is infeasible, `clustersampsi` computes the minimum number of clusters required (for the prespecified difference and power); the minimum detectable difference (for the prespecified number of clusters and power); and the maximum achievable power (for the prespecified number of clusters and the difference to be detected).

Binary, continuous, and rate outcomes are supported, with normal approximations made throughout. Between-cluster heterogeneity can be specified using either the ICC coefficient or the coefficient of variation of outcomes (cv_{clusters}). An additional option is included to allow downward adjustment of the standard deviations, for example, when baseline measurements are taken.

We outline essential formulas (in the main text and appendix), but details have been presented elsewhere (Hemming et al. 2011; Hayes and Bennett 1999).

2.1 Background

Suppose a trial will test the null hypothesis $H_0 : \mu_1 = \mu_2$, where μ_1 and μ_2 represent the means of two populations, by using a two-sample t test and assuming that $\text{var}(\mu_1) = \sigma_1^2$ and $\text{var}(\mu_2) = \sigma_2^2$. Suppose further that an equal number of individuals will be randomized to both arms, letting d denote the difference to be detected such that $d = \mu_1 - \mu_2$, $1 - \beta$ denote the power, and α denote the significance level. Alternatively, we may be interested in comparing two proportions, p_1 and p_2, or two rates, λ_1 and λ_2. We limit our consideration to trials with two equal-sized parallel arms (two-sided t tests).

Then we assume normality of outcomes and approximate the variance of the difference of the two proportions or two rates (Hemming et al. 2011). The approximations made for binomial proportions (Armitage, Berry, and Matthews 2002) are slightly different from those made in the `sampsi` command (details in appendix).

2.2 Sample-size calculations

When a trial randomizes an intervention over a number of clusters each of size m, then by standard results (Murray 1998), the required sample size n_C is that required under individual randomization (n_I) inflated by the DE,

$$\text{DE} = 1 + (m-1)\rho$$

where ρ is the ICC coefficient. This DE is modified for varying cluster sizes by a function that depends on the coefficient of variation of the cluster sizes, cv_{sizes} (Eldridge, Ashby, and Kerry 2006; this term is not to be confused with the coefficient of variation of outcomes, cv_{clusters}, described above).

From this total sample size, the number of clusters (k) required per arm can be calculated. We round up the number of clusters so that the total sample size is a multiple of the cluster size (using the ceiling function). Additionally, we add one extra cluster to each arm to allow for the use of the t distribution (Hayes and Bennett 1999). If the user instead specifies the average cluster size and needs to determine the number of clusters required per arm (m), the formula can be rearranged to determine the number of clusters as a function of the sample size required under individual randomization, the ICC, and the average cluster size (and also cv_{sizes}). More detailed mathematical formulas are provided in the appendix.

The between-cluster heterogeneity may be parameterized using either the ICC coefficient or cv_{clusters}; `clustersampsi` permits specification of either parameter. The sample-size formula for the cv_{clusters} method is outlined below (Hayes and Bennett 1999). The number of clusters k required is

$$k = 1 + \frac{n_I}{m} + \text{CVIF} \tag{1}$$

where the coefficient of variation inflation factor (CVIF) is

$$\text{CVIF} = \frac{cv^2_{\text{clusters}}(\mu_1^2 + \mu_2^2)(z_{\alpha/2} + z_\beta)^2}{d^2}$$

where $z_{\alpha/2}$ denotes the upper $100\alpha/2$ standard normal centile.

2.3 Power and detectable difference

Cluster RCTs of fixed size have both a fixed number of clusters, each with a fixed cluster size (but possibly varying between clusters), and a prespecified difference to detect. For such clusters, it may be of interest to compute available power. It turns out

that when you parameterize the heterogeneity by using the ICC, the power for cluster RCTs is the power available under individual randomization for a standardized effect size that is deflated by the square root of the DE. Similarly, for cluster RCTs of fixed sample size and prespecified power, the detectable difference is that of a trial using individual randomization inflated by the square root of the DE (Hemming et al. 2011). When parameterizing the heterogeneity with cv_{clusters}, the power available is obtained by a simple rearrangement of the sample-size formula [(2) above], whereas obtaining the detectable difference involves solving a quadratic formula.

2.4 Infeasible designs

A cluster RCT with a fixed number of clusters will be limited by an upper bound on the maximum available power or a lower bound on the detectable difference. These limits exist because of the diminishing return that sets in when the sample size of each cluster is increased (Donner and Klar 2000). These limiting values are referred to as the maximum achievable power or the minimum detectable difference.

For trials with a fixed number of equal-sized clusters k, the trial will be feasible provided that the number of clusters is greater than the product of the number of individuals required under individual randomization (n_I) and the estimated ICC (ρ). So a simple rule is that the number of clusters k will be sufficient provided that

$$k > (n_I \times \rho) + 1$$

or for clusters of varying sizes,

$$k > \left\{n_I \times \rho(cv_{\text{sizes}}^2 + 1)\right\} + 1$$

These formulas differ slightly from those reported elsewhere because of the addition of one more cluster in each arm (to allow for the use of the t distribution). When you parameterize the heterogeneity by the coefficient of variation, the following inequality must hold for the design to be feasible:

$$k > \text{CVIF} + 1$$

Where these inequalities do not hold, the `clustersampsi` command will determine the maximum available power to detect the prespecified difference, the minimum detectable difference under the prespecified value for power, and the minimum number of clusters required to detect the prespecified difference at the prespecified value of the power (Hemming et al. 2011).

2.5 Baseline adjustment: Variance deflation

Baseline measurements and other covariate adjustments lead to increases in power and are useful to consider when designing studies. The implications that adjustment for baseline measurements and predictive covariates has on sample-size calculations can

be formulated in a single framework by measuring or estimating the correlation r between either the baseline measurements or the predictive covariate and the outcome. For continuous outcomes, once an estimate of the correlation r is obtained, the variance of the estimate of the outcome is deflated by the factor $1 - r^2$. For binary outcomes, this deflation factor has been shown to be a good approximation (Hernández, Steyerberg, and Habbema 2004). To use this functionality, the user is therefore required to specify a value of the correlation between either the covariates and the outcome or the baseline values and the outcome.

2.6 The dialog box

The `clustersampsi` command is designed to be used both through the Command window and through a dialog box. All the features available within the command have been programmed into the dialog box (a `.dlg` file), and the computations are carried out using the corresponding ado-file. The dialog box includes three tabs:

1. The **Main** tab allows users to specify whether the calculation to be performed is a sample-size calculation (default), a power calculation, or a detectable difference calculation and whether this calculation is for binary, rates, or continuous (default) outcomes. If users specify a sample-size calculation, then they must also specify whether they desire to prespecify the average cluster size (the default, in which case the command computes the number of clusters required) or whether they wish to prespecify the number of clusters available (in which case the command computes the average cluster size needed). On this tab, the user also specifies the estimated ICC coefficient or the coefficient of variation.

2. The **Options** tab allows the user to specify the significance level (default 0.05), the power (default 0.8), the number of clusters per arm, the cluster size (or average cluster size), and cv_{sizes} (default 0, indicating all the clusters are the same size). Variables required to be specified on the **Options** tab are dependent on those specified on the **Main** tab, and the user will only be able to input the variables relevant to the calculation specified on the **Main** tab. For example, if the user specifies a power calculation on the **Main** tab, the power option on the **Options** tab will be shaded out. If the user specifies a sample-size calculation, then the user must also specify only one of either the number of clusters or the cluster sizes.

3. The **Values** tab allows the user to specify the proportion, rate, or mean (and standard deviation) values for the two arms, along with an estimate of correlation between any before-and-after measurements or the correlation between any covariates and the outcome (default value of 0). The command is limited to a maximum of one before and one after measurement (that is, it cannot accommodate additional repeated measurements). Once again, depending on the calculations requested on the **Main** tab (that is, sample size, power, detectable difference, and binary or continuous outcomes), those values not relevant are shaded out.

3 Examples

3.1 Example 1: Illustration of infeasible designs

In a real example, a cluster RCT will be designed to evaluate the effectiveness of support to promote breastfeeding. Randomization will be carried out at a single point in time, randomizing teams of midwives (the clusters) to either the intervention arm or the standard care arm. The trial will be carried out within a single primary care trust, so the number of clusters is limited to the 40 midwifery teams delivering care within the region. A clinically important difference to detect is an increase in the rate of breastfeeding from about 40% to 50%. Estimates of ICC range from 0.005 to 0.07 in similar trials (MacArthur et al. 2003; MacArthur et al. 2009). Using these values, we illustrate how `clustersampsi` can be used to determine the required cluster size.

Figure 1 shows a screenshot of the **Main** tab for this calculation to determine the sample size for a *Two sample comparison of proportions* with an ICC of 0.005 (the lower of the two ICC estimates).

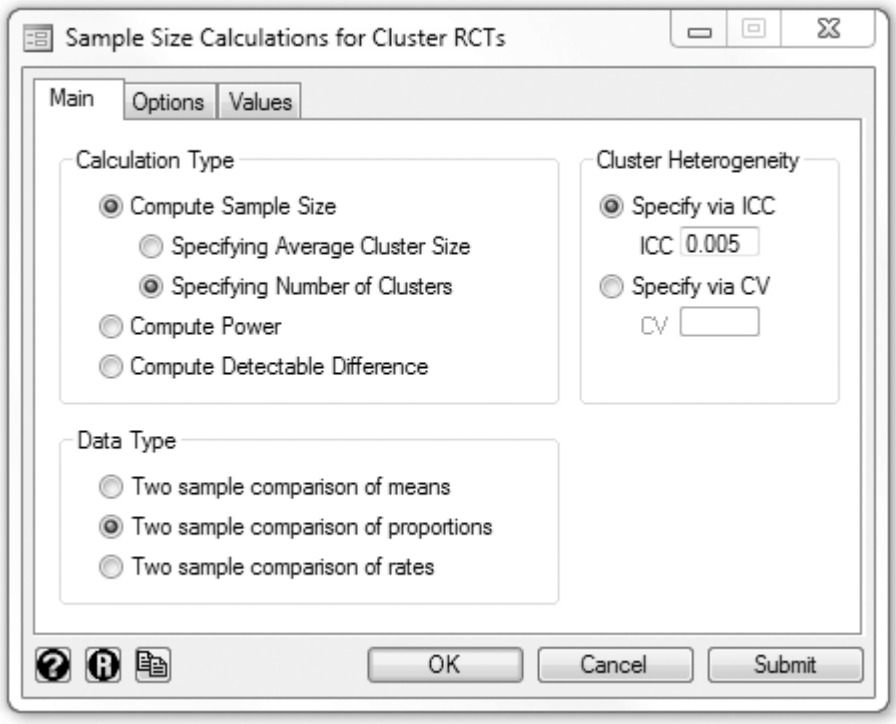

Figure 1. Screenshot of `clustersampsi` dialog box: **Main** tab—set up for example 1

Figure 2 shows the corresponding **Options** tab specifying a *Significance level* of 0.05 and 80% power. On this **Options** tab, the *Number of clusters per arm* is set at 20. The *Average cluster size* is shaded out because this is a sample-size calculation specifying the number of clusters and obtaining an estimate of the average cluster size required. The *Coefficient of variation of cluster sizes* is left at the default value of 0 and so assumes the cluster sizes are equal.

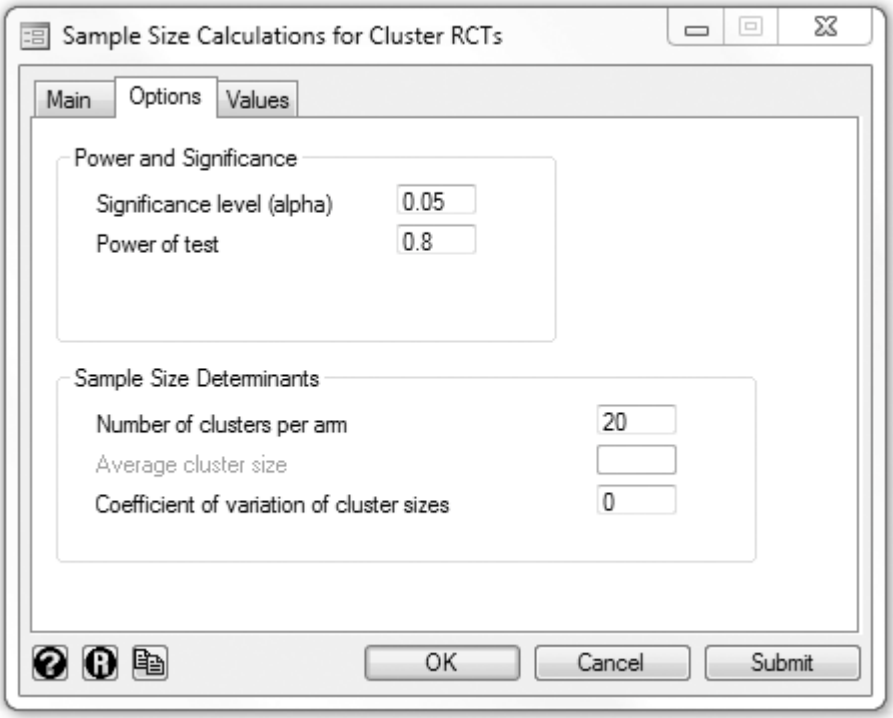

Figure 2. Screenshot of `clustersampsi` dialog box: **Options** tab—set up for example 1

Figure 3 shows the **Values** tab for this calculation. Because this is a comparison of binary proportions, the mean, standard deviation, and rate values are shaded out. *Proportion 1* is set at 0.4 and *Proportion 2* at 0.5. The correlation between before-and-after measurements is set at 0 because no baseline measurements are anticipated in this cross-sectional study.

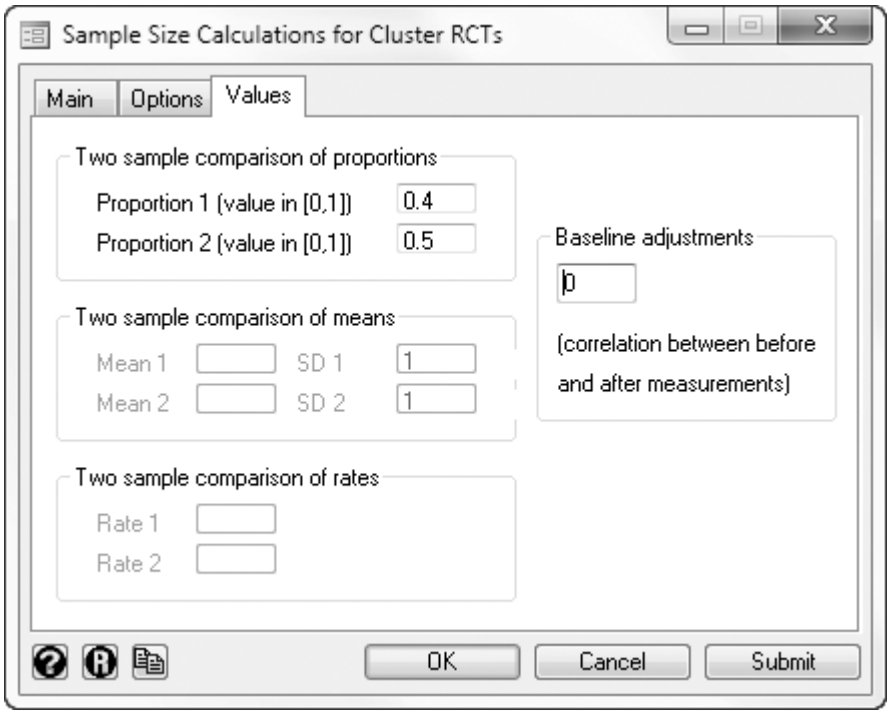

Figure 3. Screenshot of `clustersampsi` dialog box: **Values** tab—set up for example 1

The Stata output from the command is shown below. The output shows that under individual randomization, 385 individuals would be required per arm to detect a change in proportions from 0.4 to 0.5 at 80% power and a 5% significance level. Allowing for cluster randomization with 20 clusters per arm, a total of 23 individuals would be required per cluster, equating to a total sample size of 460 per arm.

```
. clustersampsi, binomial samplesize p1(0.4) p2(0.5) k(20) rho(0.005)
> size_cv(0) alpha(0.05) beta(0.8) base_correl(0)

Sample size calculation to determine number of observations required per cluster,
for a two sample comparison of proportions (using normal approximations)
without continuity correction.

For the user specified parameters:

p1:                                                          0.4000
p2:                                                          0.5000
significance level:                                          0.05
power:                                                       0.80
baseline measures adjustment (correlation):                  0.00
number of clusters available:                                20
intra cluster correlation (ICC):                             0.0050
coefficient of variation (of cluster sizes):                 0.00

clustersampsi estimated parameters:

Firstly, assuming individual randomisation:
sample size per arm:                                         385

Then, allowing for cluster randomisation:
average cluster size required:                               23
sample size per arm:                                         460

Note: sample size per arm required under cluster randomisation is rounded
up to a multiple of average cluster size.
```

In a variation of this example, the ICC is replaced by the higher of the two estimates of 0.07. The output for this computation is provided below. Under this estimate of the ICC, the design becomes infeasible; that is, however many individuals are recruited per cluster, it will not be possible to obtain 80% power to detect a difference between 0.4 and 0.5. In this scenario, the command alerts the user to this fact. The user is told that the minimum number of clusters required to detect a change from 0.4 to 0.5 at 80% power is 28 per arm. Alternatively, the user is told that because of the prespecified number of clusters (here, 20 per arm), the maximum achievable power would be in the region of 65% (that is, with 20 clusters per arm to detect a difference from 0.4 to 0.5, the study would have 65% power), and the minimum detectable difference is 0.12; that is, the design would have 80% power to detect a change from 0.4 to 0.52.

```
. clustersampsi, binomial samplesize p1(0.4) p2(0.5) k(20) rho(0.07) size_cv(0)
> alpha(0.05) beta(0.8) base_correl(0)

Sample size calculation to determine number of observations required per cluster,
for a two sample comparison of proportions (using normal approximations)
without continuity correction.

For the user specified parameters:

p1:                                                          0.4000
p2:                                                          0.5000
significance level:                                          0.05
power:                                                       0.80
baseline measures adjustment (correlation):                  0.00
number of clusters available:                                20
intra cluster correlation (ICC):                             0.0700
coefficient of variation (of cluster sizes):                 0.00

clustersampsi estimated parameters:

The sample size required under individual randomisation is:  385
The specified design is infeasible under cluster randomisation.

You could consider one of the following three options:
(i) Increase the number of clusters per arm to more than:    28
(ii) Decrease the power to:                                  0.65
(iii) Increase the difference to be detected. So,
If, trying to detect an increasing outcome then:
decrease the  difference to be detected to:                  0.1190
with corresponding p2:                                       0.5190
If, trying to detect a decreasing outcome then:
decrease the  difference to be detected to:                  0.1134
with corresponding p2:                                       0.2866
r(198);
```

3.2 Example 2: Illustrating detectable differences

A cluster RCT in Iran to evaluate the effectiveness of a polypill (composed of aspirin, Statin, and a pill that lowers blood pressure) is to be nested within a longitudinal cohort study (Pourshams et al. 2010). The clustered nature of the trial is thought to be crucial because there is a real danger of contamination because of the sharing of medication. A subset of 5,696 individuals, spread over 258 villages, is eligible and has consented to participate in this study. Villages are to be randomized to an intervention arm or a standard care arm. The average size of each village is 22 (after allowing for potential dropout) with a cv_{sizes} of 0.9; that is, there is considerable variation between the sizes of the clusters. The aim of the intervention is to reduce the composite event rate of stroke or myocardial infarction over five years. The event rate in the control group was estimated to be in the region of 0.077 over the five years. Two estimates of the ICC were obtained from previous, similar studies (0.038 and 0.018).

We illustrate how `clustersampsi` can be used to determine the effect sizes detectable at 80% power under both estimates for the ICC for the fixed sample size. Initially, we perform the calculations assuming the ICC is 0.018. The output for this calculation is provided below and illustrates the use of cv_{sizes}. The detectable event rate under the intervention arm is 0.053 (assuming a decreasing event rate), which equates to a relative risk of 0.69, that is, a relative risk reduction of 31%.

```
. clustersampsi, binomial detectabledifference p1(0.077) m(22) k(129)
> rho(0.018) size_cv(0.9) alpha(0.05) beta(0.8) base_correl(0)
Detectable difference calculation for two sample comparison of proportions
> (using normal approximations)
without continuity correction.

For the user specified parameters:

p1:                                                     0.08
significance level:                                     0.05
power:                                                  0.80
baseline measures adjustment (correlation):             0.00
average cluster size:                                   22
number of clusters per arm:                             129
coefficient of variation (of cluster sizes):            0.90
intra cluster correlation (ICC):                        0.0180

clustersampsi estimated parameters:

Firstly, under individual randomisation:
If, trying to detect an increasing outcome then:
detectable difference:                                  0.02
with corresponding p2:                                  0.10
If, trying to detect a decreasing outcome then:
detectable difference:                                  0.02
with corresponding p2:                                  0.06

Then, allowing for cluster randomisation:
design effect:                                          1.70
If, trying to detect an increasing outcome then:
detectable difference:                                  0.03
with corresponding p2:                                  0.10
If, trying to detect a decreasing outcome then:
detectable difference:                                  0.02
with corresponding p2:                                  0.05
```

Because estimation of the ICC is subject to much uncertainty, we have also carried out the calculation assuming the ICC is 0.038. Again the output is provided below. Here the detectable event rate under the intervention arm is 0.049 (again assuming a decreasing event rate), which equates to a relative risk of 0.63, that is, a 37% relative risk reduction.

```
. clustersampsi, binomial detectabledifference p1(0.077) m(22) k(129)
> rho(0.038) size_cv(0.9) alpha(0.05) beta(0.8) base_correl(0)
Detectable difference calculation for two sample comparison of proportions
> (using normal approximations)
without continuity correction.

For the user specified parameters:

p1:                                                          0.08
significance level:                                          0.05
power:                                                       0.80
baseline measures adjustment (correlation):                  0.00
average cluster size:                                        22
number of clusters per arm:                                  129
coefficient of variation (of cluster sizes):                 0.90
intra cluster correlation (ICC):                             0.0380

clustersampsi estimated parameters:

Firstly, under individual randomisation:
If, trying to detect an increasing outcome then:
detectable difference:                                       0.02
with corresponding p2:                                       0.10
If, trying to detect a decreasing outcome then:
detectable difference:                                       0.02
with corresponding p2:                                       0.06

Then, allowing for cluster randomisation:
design effect:                                               2.48
If, trying to detect an increasing outcome then:
detectable difference:                                       0.03
with corresponding p2:                                       0.11
If, trying to detect a decreasing outcome then:
detectable difference:                                       0.03
with corresponding p2:                                       0.05

Warning: Normal approximations used close to boundaries might result in
> proportions out of range
```

A clinically important relative risk is in the region of 0.65, which equates to an event rate in the treatment group of 0.05. If the ICC is as high as 0.038, then the trial will have less than 80% power to detect this difference. We illustrate how `clustersampsi` can be used to determine the power available to detect the clinically important relative risk, assuming the ICC is 0.038:

```
. clustersampsi, binomial power p1(0.077) p2(0.05) m(22) k(129) rho(0.038)
> size_cv(0.9) base_correl(0)
Power calculation for a two sample comparison of proportions (using normal
> approximations)
without continuity correction.

For the user specified parameters:

p1:                                                          0.0770
p2:                                                          0.0500
significance level:                                          0.05
baseline measures adjustment (correlation):                  0.00
average cluster size:                                        22
number of clusters per arm:                                  129
coefficient of variation (of cluster sizes):                 0.90
intra-cluster correlation (ICC):                             0.0380

clustersampsi estimated parameters:

Firstly, assuming individual randomisation:
power:                                                       0.99
Then, allowing for cluster randomisation:
design effect:                                               2.48
power:                                                       0.75
```

The power available to detect this difference is 75%, close to 80%. Thus the trial will almost be sufficiently powered to detect this difference.

3.3 Example 3: Illustrating the coefficient of variation to measure heterogeneity

Hayes and Bennett (1999) show how the coefficient of variation can be used as an alternative to the ICC to describe the variation in outcomes between clusters. In their illustrative cases, they describe an example of a cluster sample-size calculation for a comparison of rates and for measuring cv_{clusters}. We reproduce this example here and illustrate how `clustersampsi` could be used to perform this calculation. The objective is to determine the number of clusters required.

The study is designed to detect a difference between two rates, $\lambda_1 = 0.0148$ and $\lambda_2 = 0.0104$, at 80% power and 5% significance with approximately 424 person-years of observations in each cluster and with a cv_{clusters} of 0.29. `clustersampsi` returns the value of 37 clusters per arm:

```
. clustersampsi, samplesize rates r1(0.0148) r2(0.0104) m(424) cluster_cv(0.29)
> size_cv(0) alpha(0.05) beta(0.8) base_correl(0)
Sample size calculation determining the number of clusters required,
for a two sample comparison of rates (using normal approximations).

For the user specified parameters:

rate 1:                                                    0.0148
rate 2:                                                    0.0104
significance level:                                        0.05
power:                                                     0.80
baseline measures adjustment (correlation):                0.00
average person years per cluster:                          424
cluster coefficient of variation (of outcomes):            0.29

clustersampsi estimated parameters:

Firstly, assuming individual randomisation:
sample size per arm:                                       10217

Then, allowing for cluster randomisation:
sample size per arm:                                       15688
number clusters per arm (m):                               37

Note: sample size per arm required under cluster randomisation is rounded up
to a multiple of average cluster size and includes the addition
of one extra cluster per arm (to allow for t-distribution).
To understand sensitivity to these conservative allowances:
power with m clusters per arm:                             0.81
power with m-1 clusters per arm:                           0.80
```

This is very close to the 36.2 reported by Hayes and Bennett. In the trial, only 28 clusters were recruited. We can therefore use `clustersampsi` to evaluate the power that the trial would have had if limited to 28 clusters:

```
. clustersampsi, rates power r1(0.0148) r2(0.0104) m(424) k(28)
> cluster_cv(0.29) alpha(0.05)
Power calculation for a two sample comparison of rates (using normal
> approximations).


For the user specified parameters:

rate 1:                                                    0.014800
rate 2:                                                    0.010400
significance level:                                        0.05
baseline measures adjustment (correlation):                0.00
average person years per cluster:                          424
number of clusters per arm:                                28
cluster coefficient of variation (of outcomes):            0.29

clustersampsi estimated parameters:

Firstly, assuming individual randomisation:
power:                                                     0.86
Then, allowing for cluster randomisation:
power:                                                     0.69
```

`clustersampsi` estimates the power to be about 69%, again similar to that reported by Hayes and Bennett.

4 Conclusion

While cluster sample-size calculations are, for the most part, simple extensions of those required under individual randomization, specific commands in Stata for this class of problems should prove very useful. Some commands are currently available in Stata to perform these calculations, but one is very basic and requires a two-step approach, and the other is specifically designed for trials in which there is no clustering in the control arm.

The command outlined here, `clustersampsi`, allows not only for clustering but also for varying cluster sizes, for baseline measurements, or for adjustment for predictive covariates. It also incorporates calculations of samples sizes, power, and detectable differences. It will alert the user to infeasible designs and suggest possible options. The user can parameterize cluster heterogeneity by using either the ICC coefficient or the coefficient of variation. The dialog box for `clustersampsi` should allow straightforward implementation for the most common types of cluster RCTs.

When we compare the output of `clustersampsi` with that of `sampclus`, the estimates from `clustersampsi` tend to result in slightly higher sample sizes because it rounds up to a multiple of the average cluster size and because it adds one to the number of clusters. On the other hand, compared with the estimates from `clsampsi`, the estimates from `clustersampsi` tend to be more conservative (that is, a slightly lower estimated sample size or slightly higher estimated power) because of the noncentral F distribution used by `clsampsi`. These differences are more marked at the parameter boundaries (such as small proportions or few clusters).

We have used a number of approximations here. First, we have approximated the variance of proportions and rates, we have assumed normality, and we have not made continuity corrections. Continuity-corrected sample-size calculations are more conservative but are not considered optimal by everyone (Royston and Babiker 2002). More importantly, we have also approximated the variance reduction due to correlation between any baseline measurements for binary outcomes by using normality approximations. For continuous outcome measurements in RCTs, adjustment for baseline measurements will always lead to a reduction in the standard deviation by a factor that depends on the correlation between the before-and-after measurements (Robinson and Jewell 1991). For binary outcomes (as opposed to continuous outcomes), although adjustment for baseline measures will lead to an increase in power, this is not necessarily by the same factor. However, it has been shown by others to provide a good approximation (Hernández, Steyerberg, and Habbema 2004).

5 Appendix: Formulas

The formulas follow those already published (Hemming et al. 2011; Hayes and Bennett 1999), with some minor modifications. When the heterogeneity between clusters is specified by the ICC, then the formulas in Hemming et al. (2011) are used but with the addition of one to the number of clusters in each arm to account for the t distribution

rather than the normal distribution (as recommended by Hayes and Bennett [1999]). When the heterogeneity between clusters is specified by the coefficient of variation, then the formulas follow those in Hayes and Bennett (1999). The essential formulas for both methods are described below.

5.1 Formulas using the ICC

The required sample size per arm for a trial at prespecified power $1-\beta$ to detect a prespecified difference of $d = \mu_1 - \mu_2$ is n_I, where

$$n_I = (\sigma_1^2 + \sigma_2^2)\left\{\frac{(z_{\alpha/2} + z_\beta)^2}{d^2}\right\}$$

Baseline adjustment or adjustment for other covariates will deflate the standard deviation by a factor we call $B = (1 - r^2)$. The formula above can be simply modified by replacing σ_1^2 with $B \times \sigma_1^2$ and similarly for σ_2^2. For clusters of average size $\overline{m}$ with cv_{sizes}, the required number of clusters is k, where

$$k = 1 + \frac{n_I \text{VIF}}{\overline{m}} \tag{2}$$

where the variance inflation factor (VIF) is

$$\text{VIF} = 1 + \left\{\left(cv_{\text{sizes}}^2 + 1\right)\overline{m} - 1\right\}\rho \tag{3}$$

For clusters of equal size, this simplifies to

$$\text{VIF} = 1 + (m-1)\rho$$

For binary variables p_1 and p_2, we approximate $sd_1^2 = p_1(1-p_1)$ and similarly for sd_2^2. For rates λ_1 and λ_2, we approximate the variances $sd_1^2 = \lambda_1$ and $sd_2^2 = \lambda_2$.

The above formulas may be simply rearranged to compute power and detectable differences for mean values. For detectable differences for binary outcomes, it is necessary to solve the following quadratic to find the detectable difference p_2:

$$0 = ap_2^2 + bp_2 + c \tag{4}$$

where

$$\begin{aligned} a &= -1 - a_1 \\ b &= 1 + 2a_1 p_1 \\ c &= p_1(1-p_1) - a_1 p_1^2 \end{aligned}$$

and where

$$a_1 = \frac{(k-1)m}{B \times \text{VIF}(z_{\alpha/2} + z_\beta)^2}$$

This provides two values for p_2 that correspond to increasing and decreasing values.

If the user is limited to a fixed number of clusters and needs to determine the number of observations per cluster, then (5) can be rearranged to give the number of observations required for each cluster. So, where the clusters are of fixed size, the number of observations per cluster is

$$m = \frac{n_I(1-\rho)}{k-1-\rho n_I}$$

so that the number of clusters required to make this design feasible is greater than $\rho n_I + 1$. If the clusters are of varying size, then using the alternative VIF in (6) gives the number of observations required per cluster as

$$m = \frac{n_I(1-\rho)}{k-1-\rho n_I(cv^2_{\text{sizes}}+1)}$$

and, in this case, the minimum number of clusters required to make this design feasible is $\rho(cv^2_{\text{sizes}}+1)n_I + 1$.

As well as computing the minimum number of clusters required under a design that is infeasible, `clustersampsi` computes the maximum power value and the minimum detectable difference available with the limited number of clusters. These values are obtained by finding the maximum value for z_β or the minimum value for d^2, which would result in $k-1-n_I\rho(cv^2_{\text{sizes}}+1) > 0$. So for example, the maximum available power for fixed m is

$$z_\beta = \sqrt{\frac{(k-1)d^2}{\rho(cv^2_{\text{sizes}}+1)(\sigma_1^2+\sigma_2^2)}} - z_{\alpha/2}$$

and the minimum detectable difference for continuous outcomes is

$$d = \sqrt{\frac{\rho(cv^2_{\text{sizes}}+1)(\sigma_1^2+\sigma_2^2)(z_{\alpha/2}+z_\beta)^2}{k-1}}$$

For binary outcomes, the minimum detectable difference is given by (4) except that a_1 is replaced by

$$a_1 = \frac{(k-1)}{(z_{\alpha/2}+z_\beta)^2(Bcv^2_{\text{sizes}}+1)\rho}$$

5.2 Formulas using the coefficient of variation

The required sample size per arm for a trial at prespecified power $1-\beta$ to detect a prespecified difference of $d = \mu_1 - \mu_2$ is again n_I, where

$$n_I = (\sigma_1^2+\sigma_2^2)\left\{\frac{(z_{\alpha/2}+z_\beta)^2}{d^2}\right\}$$

When each of the clusters is size m, the number of clusters required is k so that

$$k = 1 + \frac{Bn_I}{m} + B \times \text{CVIF}$$

where the CVIF is

$$\text{CVIF} = \frac{cv^2_{\text{clusters}}(\mu_1^2 + \mu_2^2)(z_{\alpha/2} + z_\beta)^2}{d^2}$$

and where cv_{clusters} is the coefficient of variation of the outcome across the clusters.

Power and detectable difference are simply obtained by rearranging the above formulas and solving the resulting quadratic where necessary. For proportions, this amounts to solving

$$0 = ap_2^2 + bp_2 + c$$

where

$$\begin{aligned} a &= cv^2_{\text{clusters}} - a_2 - \frac{1}{m} \\ b &= \frac{1}{m} + 2a_2p_1 \\ c &= \frac{p_1}{m} + \frac{p_1^2}{m} + cv^2_{\text{clusters}}p_1^2 - a_2p_1^2 \end{aligned}$$

and where

$$a_2 = \frac{k-1}{B(z_{\alpha/2} + z_\beta)^2}$$

For continuous outcomes, this is such that

$$\begin{aligned} 0 &= a\mu_2^2 + b\mu_2 + c \\ a &= cv^2 - a_2 \\ b &= 2a_2\mu_1 \\ c &= \frac{(\sigma_1^2 + \sigma_2^2)}{m} - a_2\mu_1^2 + cv^2\mu_1^2 \end{aligned}$$

where a_2 is as in the binary case above.

Again, if the user is limited to a prespecified number of clusters, then it is possible to determine the required average cluster size:

$$m = \frac{n_I}{k - 1 - \text{CVIF}}$$

Certain designs will be infeasible; for a feasible design, the number of clusters required is greater than CVIF + 1. Alternatively, limited to this number of clusters, the design will become feasible on either lowering the power or increasing the difference to be detected. The maximum available power and minimum detectable difference are obtained by determining the maximum value for z_β or minimum value for d^2, which results in $k - 1 - \text{CVIF} > 0$.

The maximum available power for both continuous and binary outcomes is

$$z_\beta = \sqrt{\frac{(k-1)d^2}{Bcv^2_{\text{clusters}}(\mu_1^2 + \mu_2^2)}} - z_{\alpha/2}$$

The minimum detectable difference for both continuous and binary outcomes again involves solving a quadratic whose coefficients are

$$a = 1 - a_3$$
$$b = 2a_3\mu_1$$
$$c = \mu_1^2 - a_3\mu_1^2$$

and where

$$a_3 = \frac{(k-1)}{B \times (z_{\alpha/2} + z_\beta)^2 cv^2_{\text{clusters}}}$$

All functions use ceiling values throughout, so for example, if the number of clusters is estimated to be 7.1, this will be rounded up to 8.

`clustersampsi` will not give identical results to `sampsi` for the sample size under individual randomization with binary data (hence, any cluster sample sizes calculated via a two-step approach from results of `sampsi` will not tally with results from `clustersampsi`). This is due to an approximation in the case of equal allocation to treatment group: `sampsi` uses no approximation (equation 3.2 in Machin et al. [1997]) but `clustersampsi` does (equation 3.8 in Machin et al. [1997]). Practically speaking, the difference in sample sizes is only large (more than 10% of the exact sample size required) where small sample sizes (fewer than about 50) are called for. In such situations, the more pressing issue is the use of a cluster design with small samples rather than the precise size of said sample. Power will also differ for comparisons of proportions because of the use of this approximation. Generally, this difference is negligible but may be of concern when looking for particularly large effects.

6 Funding acknowledgment

Karla Hemming was partially funded by a National Institute of Health Research (NIHR) grant for Collaborations for Leadership in Applied Health Research and Care (CLAHRC) for the duration of this work. The views expressed in this publication are not necessarily those of the NIHR or the Department of Health.

7 References

Armitage, P., G. Berry, and J. N. S. Matthews. 2002. *Statistical Methods in Medical Research*. 4th ed. Oxford: Blackwell.

Batistatou, E., and C. Roberts. 2010. clsampsi Stata command. http://www.medicine.manchester.ac.uk/healthmethodology/research/biostatistics/data/clsampsi/.

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

Eldridge, S. M., D. Ashby, and S. Kerry. 2006. Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* 35: 1292–1300.

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., B. Giraudeau, and P. Ravaud. 2005. A priori postulated and real power in cluster randomized trials: Mind the gap. *BMC Medical Research Methodology* 5: 25.

Hayes, R. J., and S. Bennett. 1999. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 28: 319–326.

Hemming, K., A. J. Girling, A. J. Sitch, J. Marsh, and R. J. Lilford. 2011. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology* 11: 102.

Hernández, A. V., E. W. Steyerberg, and J. D. Habbema. 2004. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of Clinical Epidemiology* 57: 454–460.

MacArthur, C., K. Jolly, L. Ingram, N. Freemantle, C. L. Dennis, R. Hamburger, J. Brown, J. Chambers, and K. Khan. 2009. Antenatal peer support workers and initiation of breast feeding: Cluster randomised controlled trial. *British Medical Journal* 338: b131.

MacArthur, C., H. R. Winter, D. E. Bick, R. J. Lilford, R. J. Lancashire, H. Knowles, D. A. Braunholtz, C. Henderson, C. Belfield, and H. Gee. 2003. Redesigning postnatal care: A randomised controlled trial of protocol-based midwifery-led care focused on individual women's physical and psychological health needs. *Health Technology Assessment* 7: 1–98.

Machin, D., M. J. Campbell, P. M. Fayers, and A. Pinol. 1997. *Sample Size Tables for Clinical Studies*. 2nd ed. Oxford: Blackwell Science.

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

Pourshams, A., H. Khademi, A. F. Malekshah, F. Islami, M. Nouraei, A. R. Sadjadi, E. Jafari, N. Rakhshani, R. Salahi, S. Semnani, F. Kamangar, C. C. Abnet, B. Ponder, N. Day, S. M. Dawsey, P. Boffetta, and R. Malekzadeh. 2010. Cohort Profile: The Golestan Cohort Study—A prospective study of oesophageal cancer in northern Iran. *International Journal of Epidemiology* 39: 52–59.

Robinson, L. D., and N. P. Jewell. 1991. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 58: 227–240.

Royston, P., and A. Babiker. 2002. A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome. *Stata Journal* 2: 151–163.

About the authors

Karla Hemming and Jen Marsh are both lecturers at the University of Birmingham in the Department of Public Health, Epidemiology and Biostatistics.