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 2, pp. 363–380

A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials

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

Alan Girling
University of Birmingham
Birmingham, UK
a.j.girling@bham.ac.uk

Abstract. This article introduces the Stata menu-driven program `steppedwedge`, which calculates detectable differences and power for stepped-wedge randomized trials. The command permits continuous, binary, and rate outcomes (with normal approximations) for comparisons using two-sided tests. The command allows specification of the number of clusters randomized at each step, the number of steps and the average cluster (cell) size, or an incomplete design in which the user specifies the design pattern (a matrix with one row per cluster, one column per time point, and entries indicating exposure and observable data). Cluster heterogeneity can be parameterized using either the intracluster correlation or the coefficient of variation (of the outcome). The command is illustrated via examples.

Keywords: st0341, steppedwedge, stepped wedge, sample size, cluster-randomized controlled trials, power, detectable difference

1 Introduction

The stepped-wedge design is a modification of the conventional cluster-randomized trial (CRT), in which all clusters initiate as control clusters with sequential but random rollout of the intervention at various points in time so that by the end of the study, all clusters have crossed over to become intervention clusters (Brown and Lilford 2006; Mdege et al. 2011). With increasing frequency, stepped-wedge studies are being used across different settings while evaluating a range of interventions (Mdege et al. 2011).

Calculation of sample size for CRTs is a relatively straightforward modification to that required under individually randomized designs and usually requires inflation of the number of participants by a variance inflation factor (VIF) known as the design effect (Hemming et al. 2011). Currently, there are several options available to Stata users planning a CRT: these include the two-step procedure using the `sampclus` command (Garrett 2001) and the one-step procedure using the `clsampsi` command (Batistatou, Roberts, and Roberts 2014). Lately, these have been supplemented by the command `clustersampsi` (Hemming and Marsh 2013), which will compute power, sample size, and detectable difference under a range of scenarios for CRTs. For basic power and sample-size calculations, see the `power` command.

There is a dearth of literature on the theoretical underpinnings of similar calculations for stepped-wedge studies, with only one well-known publication (Hussey and Hughes 2007). This is compounded by the fact that power calculations for stepped-wedge studies are more complex because they require the use of matrix algebra (Hussey and Hughes 2007). Currently, there are no Stata commands that allow computation of sample size, power, or detectable difference for stepped-wedge studies. Therefore, we have developed a Stata command, `steppedwedge`, that we believe will be very practical for applied health care researchers involved in the design of stepped-wedge studies. This command will compute power and detectable difference for stepped-wedge studies of known sample size.

2 Background

In a conventional, parallel, individually randomized controlled trial (RCT) with continuous outcomes, the power, $1-\beta$, to detect a systematic difference in outcome between two arms of a study is obtained from the equation

$$z_\beta = \frac{\delta}{\sigma}\sqrt{\frac{n}{2}} - z_{\alpha/2}$$

where $\delta = \mu_1 - \mu_2$ is the systematic difference (in means μ_1 and μ_2), σ^2 is the variance of an individual observation such that $\sigma^2/2n$ is the variance of the estimated difference, n is the sample size per arm, and α is the significance level (for a two-sided comparison). In a conventional RCT, all observations are independent of one another. Cluster trials differ because individual observations from the same cluster may be positively correlated with one another (though observations from different clusters are not). The correlation structure is characterized by the intracluster correlation coefficient (ICC), which may be defined either as the correlation between two observations in the same cluster or, equivalently, as the proportion of the individual variance attributable to cluster membership. The individual variance is written as $\sigma^2 = \tau^2 + \sigma_w^2$, where τ^2 is the variance of the cluster means, and σ_w^2 is the within-cluster variance (that is, the conditional variance of an observation given the cluster to which the individual belongs). Then the ICC is

$$\rho = \frac{\tau^2}{\tau^2 + \sigma_w^2}$$

In such a trial with equal numbers of individuals (m) in each cluster, the variance of the estimated treatment-effect estimate is inflated by a factor known as the VIF,

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

compared with an individually randomized trial with the same number of observations. To achieve the same power as an individually randomized study, we must increase the sample size in a cluster trial by the VIF, which in this context is usually referred to as the "design effect". Thus the power for a CRT is given by the equation

$$z_\beta = \frac{\delta}{\sigma}\sqrt{\frac{n}{2\text{VIF}}} - z_{\alpha/2}$$

In Stata, this calculation can be accomplished by the `clustersampsi` command (Hemming and Marsh 2013), which can also determine the minimum sample size at given power and difference and the minimum detectable difference at given power and sample size. These results strictly apply to normally distributed outcomes. Data on counts or proportions are handled by the command using appropriate normal approximations (Armitage, Berry, and Matthews 2002; Hayes and Bennett 1999).

While in CRTs it has become the convention to specify the magnitude of heterogeneity between clusters by the ICC, the coefficient of variation (CV) of the outcome is perhaps a more intuitive method of specification for positive or binary outcomes (Hayes and Bennett 1999). The choice between ICC and CV is largely one of convenience. For example, if the CV (of the outcome) in arm 1 is

$$\text{CV} = \frac{\tau}{\mu_1}$$

the corresponding ICC is

$$\rho = \frac{\text{CV}^2\mu_1^2}{\text{CV}^2\mu_1^2 + \sigma_w^2}$$

3 The steppedwedge command

The new Stata command `steppedwedge` computes power and detectable difference for both complete and incomplete stepped-wedge designs. A complete design is a design in which at each of a fixed number of points in time a block of a fixed number of clusters initiate (by random selection) the intervention. An incomplete design is a design in which the number of clusters randomized to the intervention differs between the time points; or in which at some time points, no clusters are initiated on the intervention; or in which at some time points, no observations of data are taken. These concepts are described in more detail below.

Binary, continuous, and count outcomes are supported with normal approximations made throughout. Between-cluster heterogeneity can be specified either using the ICC or the CV of outcomes (note: this is not to be confused with the CV of cluster sizes, which is a parameter sometimes used in the design of cluster trials but not considered here). An additional option is included to allow the user to specify whether the variance specified is the total variance or the within-cluster variance.

We outline essential formula in the main text, but details have been presented elsewhere (Hussey and Hughes 2007). Of note is that the command computes only power and detectable difference for fixed sample size and does not permit the calculation of sample size for fixed power and detectable difference.

3.1 Stepped-wedge designs

The conventional stepped-wedge design assumes the following: that at each of a fixed number of points in time, a block of clusters are sequentially randomized to receive the

intervention; that they remain exposed to the intervention for all subsequent points in time; and that observations are made at each of these time points and at baseline to form the data for analysis. We refer to the number of observations made on each cluster at each point in time as the cell size. This is a complete design. A design-pattern matrix (of size number of time points by number of clusters) can be used to quantify the design of the stepped-wedge study. Each cell of this design-pattern matrix is an indicator of exposure to the intervention (1 for exposed; 0 for not exposed) and observable data (. for no observable data). The design-pattern matrix is not to be confused with the traditional "design matrix" (see section 3.2).

Figure 1 illustrates the design pattern for a complete design with six time points (including a baseline point) and clusters randomized to the intervention at each of the five steps, with an additional time point for baseline data collection. Cells with a 1 indicate that the clusters within that block at that time point are exposed to the intervention, and cells with a 0 indicate that clusters within that block at that time point are not exposed to the intervention. These blocks may or may not consist of more than one cluster, but it is conventional to assume that an equal number of clusters are contained within each block.

Block	Time					
	0	1	2	3	4	5
1	0	1	1	1	1	1
2	0	0	1	1	1	1
3	0	0	0	1	1	1
4	0	0	0	0	1	1
5	0	0	0	0	0	1

Figure 1. Illustrative example of the stepped-wedge study of a complete design

In practice, the design may not be complete. Figure 2 illustrates an incomplete design in which, at a particular time, a cluster may be exposed (1), not exposed (0), or not contribute to the analysis (·). Here the clusters are randomized sequentially, but observations are made right before the intervention is introduced and at the two time points immediately afterward. In figure 3, we illustrate a design-pattern matrix where data are collected throughout the observation period except for the time in which the intervention is being implemented. We refer to this as a transition period; a design is once again incomplete.

Block	Time					
	0	1	2	3	4	5
1	0	1	1	.	.	.
2	.	0	1	1	.	.
3	.	.	0	1	1	.
4	.	.	.	0	1	1

Figure 2. Illustrative example of the stepped-wedge study of incomplete design with one before and two after measurements

Block	Time					
	0	1	2	3	4	5
1	0	.	1	1	1	1
2	0	0	.	1	1	1
3	0	0	0	.	1	1
4	0	0	0	0	.	1

Figure 3. Illustrative example of the stepped-wedge study of incomplete design with an implementation period

To use the `steppedwedge` command, paste these design patterns into the Stata Data Editor. This can simply be done either by hand or by writing a short code. For example, the design-pattern matrix illustrated in figure 2 could be constructed using the simple Stata commands:

```
. set obs 4
obs was 0, now 4
. forvalues i=0/5 {
  2. generate x`i´ = (_n<=`i´) if (_n<`i´+2) & (_n>=`i´-2)
  3. }
```

(*output omitted*)

3.2 Power for complete stepped-wedge designs

Now consider the complete design with k clusters (assume blocks of size 1 initially) and $(t+1)$ time points (note that for the complete case, $k = t$). Following Hussey and Hughes

(2007), we adopt a linear mixed model to describe the data with time as a fixed factor at $(t+1)$ levels and with a random effect for intercluster variation. Thus there are $(t+2)$ fixed-effects parameters in the model. This linear model may be represented as

$$\begin{aligned} y_{ijs} &= z_{js}\theta_I + \theta_s + \alpha_j + e_{ijs} \\ \alpha_j &\sim N(0, \tau^2) \\ e_{ijs} &\sim N(0, \sigma_w^2) \end{aligned}$$

where y_{ijs} is the outcome for individual i $(i = 1, \ldots, m)$ in cluster j $(j = 1, \ldots, k)$ at time point s $(s = 1, \ldots, t+1)$; z_{js} denotes exposure to the intervention with treatment effect θ_I (the notation I refers to the intervention); θ_s is a fixed effect for time; and α_j represents a random effect for cluster j $(j = 1, \ldots, k)$. Note that here m refers to the cell size, which is the number of observations per cluster at each point in time. Whereas in a CRT, m conventionally refers to the total number of observations within each cluster.

For power calculations, it is convenient to reparameterize this model in matrix format. To this end, the data vector $\mathbf{y}$, of cell means, is ordered by time within clusters—that is according to the rows of the pattern matrix. So the data of cell means (for the t clusters at the $t+1$ time points) are described as a $t \times (t+1)$ by 1 vector $\mathbf{y}$ (one row per cluster per time point) with expected value

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\theta}$$

where $\mathbf{X}$ (of size $t \times [t+1]$ by $[t+2]$) is the conventional design matrix (not to be confused with the design-pattern matrix) that may be derived from the model, and $\boldsymbol{\theta}$ is the $(t+2)$ vector of the parameters. For convenience, we suppose that the intervention effect is the first element of the vector $\boldsymbol{\theta}$. Let $\mathbf{V}$ denote the variance–covariance matrix of $\mathbf{y}$, then $\mathbf{V}$ is a $t(t+1)$ by $t(t+1)$ block-diagonal matrix with identical $(t+1)$ by $(t+1)$ blocks of the form

$$\begin{pmatrix} \frac{\sigma_w^2}{m} + \tau^2 & \frac{\sigma_w^2}{m} & \cdots & \frac{\sigma_w^2}{m} \\ \frac{\sigma_w^2}{m} & \frac{\sigma_w^2}{m} + \tau^2 & \cdots & \frac{\sigma_w^2}{m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_w^2}{m} & \frac{\sigma_w^2}{m} & \cdots & \frac{\sigma_w^2}{m} + \tau^2 \end{pmatrix}$$

The variance of the treatment-effects estimate is the leading element of the variance–covariance matrix for the (fixed-effects) parameter estimates, that is, $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}[1,1]$, where the notation $[1,1]$ refers to the matrix cell in the first column and first row. Hussey and Hughes (2007) use this expression to develop a Wald test for the treatment effect (assuming known variances) with power computed from

$$z_\beta = \frac{\delta}{\sqrt{\{(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\}[1,1]}} - z_{\alpha/2}$$

This formula may be easily rearranged to determine detectable difference for fixed power. This formula can be easily modified to accommodate more than one cluster randomized at each step (that is, blocks of size more than one).

3.3 Power for incomplete stepped-wedge designs

The above formula can be modified for incomplete designs and blocks of varying sizes, simply by taking the appropriate design matrix for the incomplete design and the appropriate block-diagonal matrix, $\mathbf{V}$, which will remain a block-diagonal matrix and contain the same number of blocks as clusters. However, this time, each block will vary depending on the observations taken for cluster k.

3.4 The dialog box

The `steppedwedge` command has been designed to be used both with the command window and with a drop-down dialog box. All the features available within the command (an ado-file) 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 menu tabs.

The **Main** tab allows the user to specify whether the calculation to be performed is a power calculation (default) or a detectable-difference calculation and whether this calculation is for proportions, rates, or means (default). The **Main** tab also allows the user to specify the significance level (default is `0.05`) and the power (default is `0.8`) where appropriate.

On the **Clusters** tab, the user specifies whether the design is complete or incomplete. For complete designs, the user must specify the number of clusters per step (which we refer to as the block size), the average cluster (cell) size (this is the size of each cluster at each step), and the number of steps. For incomplete designs, the user must paste the design pattern into the Stata Data Editor. Under both designs, the user can opt to have the design-pattern matrix printed as part of the output provided (this option additionally prints the design-pattern matrix into the current Data Editor). On the **Clusters** tab, the user also specifies the estimated ICC or the CV.

On the **Values** tab, the user specifies the proportion, rate, or mean (and standard deviation) values for the two arms. Depending on the calculations requested on the **Main** tab (that is, power, detectable difference, binary rates, or continuous outcomes), those values not relevant are shaded out. The user can also indicate whether the variances specified are the total variances (σ^2) or the within-cluster variances (σ_w^2).

4 Examples

4.1 Example 1: Illustration of a complete design

In a real example, a stepped-wedge study is designed to evaluate the effectiveness of an educational package for midwives to promote the practice of sweeping of the membranes in women going postterm in pregnancy. Randomization is carried out at a single point in time, randomizing the order with which the teams of midwives (the clusters) cross over from standard care to the intervention. The trial is carried out within two primary care trusts within Birmingham, UK, and the number of clusters is limited to the 10 midwife teams delivering care within the region. The educational package is implemented irrespective of the evaluation of effectiveness. However, local collaborations for leadership in applied health research and care initiated plans for the evaluation of effectiveness alongside implementation in the form of a stepped-wedge trial.

The primary outcome is the proportion of women accepting a membrane sweep in the period following completion of the training within a team. The denominator population consists of all women who give birth after 39 weeks and 3 days. It is estimated from hospital data that within a week, approximately 12 women will give birth within each team after 39 weeks plus 3 days gestation. It is further anticipated that the sequential rollout of the training intervention will be at the rate of approximately one team per week. So under a stepped-wedge design, there will be 10 steps, with 1 cluster randomized at each step, with approximately 12 observations per cluster per time point. Estimates of ICCs are limited, but to be conservative, we considered a range from 0.01 to 0.1.

A clinically important difference to detect is an increase in the rate of sweeping from about 40% to 50%. Given this fixed sample size (this is a pragmatic study, and we are not at liberty to increase the study population) and this clinically important, detectable difference, we can then determine the level of power available. Using these values, we illustrate how `steppedwedge` can be used to determine the power.

Figure 4 shows a screenshot of the **Main** dialog tab for this calculation, which determines power available for a two-sample comparison of proportions and specifies a significance level of `0.05`. The value for power is shaded out on this dialog tab because this is a power calculation, and power is therefore to be determined.

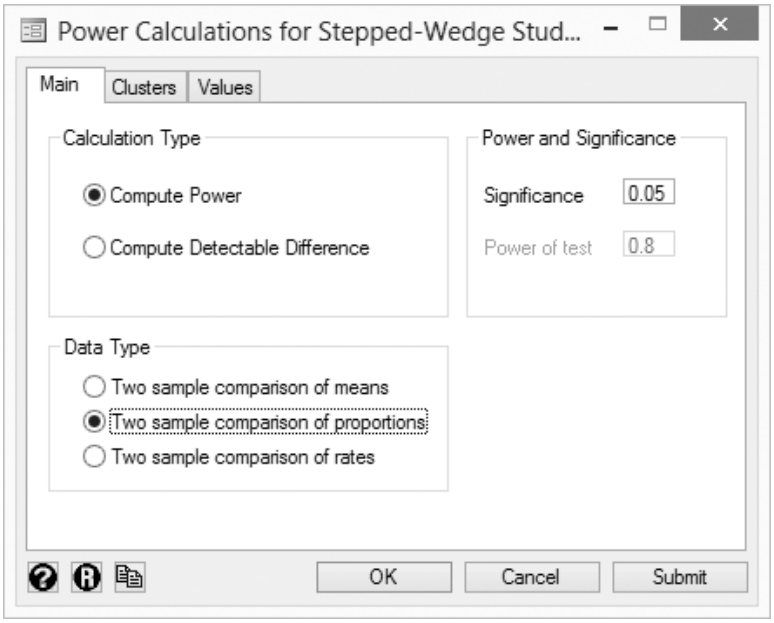

Figure 4. Screenshot of the `steppedwedge` dialog box: **Menu** tab—set up for example 1

Figure 5 shows the corresponding **Clusters** tab with the method set to use a complete design and with the option set to report the design-pattern matrix as part of the command output. On this tab, the number of clusters randomized at each step is set to `1`, the number of steps is set to `10`, and the average cluster size is specified as `12`. The between-cluster heterogeneity is parameterized by the ICC and is set at `0.01`.

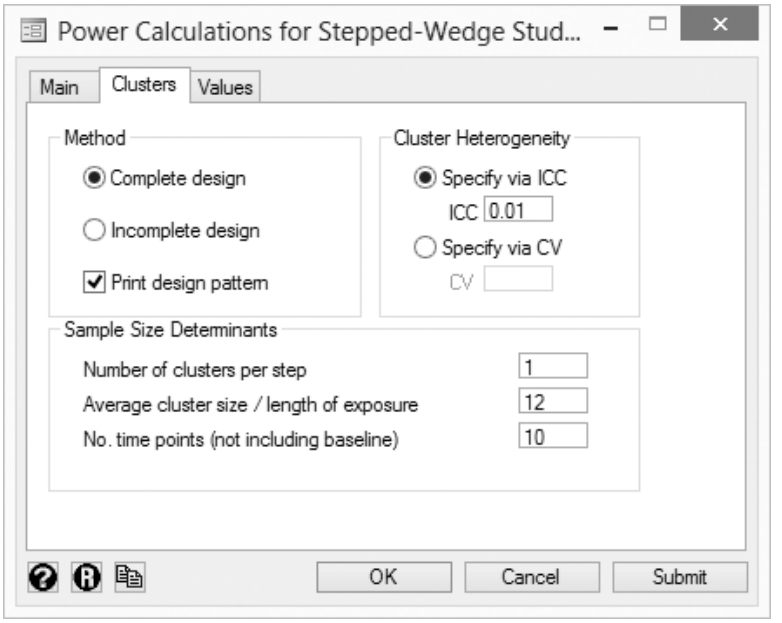

Figure 5. Screenshot of the `steppedwedge` dialog box: **Clusters** tab—set up for example 1

Figure 6 shows the **Values** dialog tab for this calculation. Because this is a comparison of binary proportions, the mean, standard deviation, and rate values are shaded out. The value for *Proportion 1* is set at `0.4` and *Proportion 2* at `0.5`. The variance is specified as the total variance.

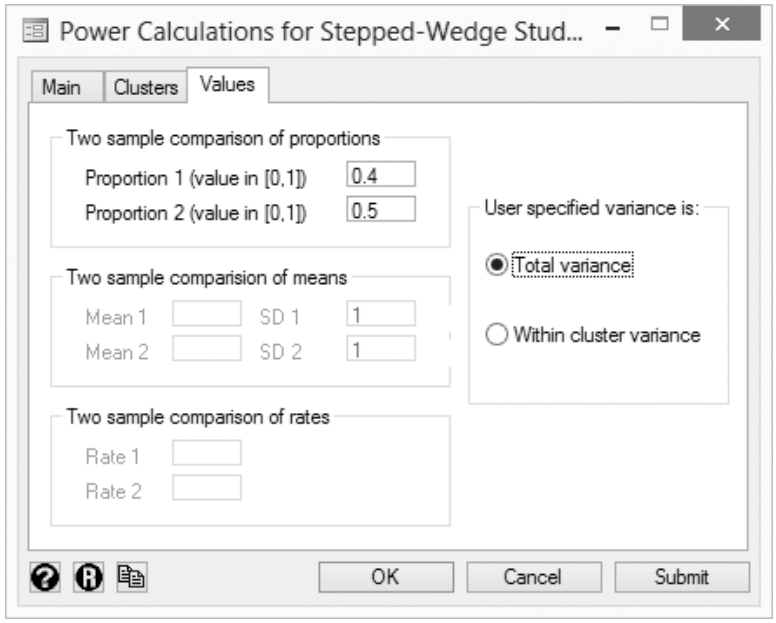

Figure 6. Screenshot of the `steppedwedge` dialog box: **Values** tab—set up for example 1

The Stata output from the command is given below. The output also (for verification) prints out sample-size parameters specified by the user (that is, the average cluster size, number of clusters, number of time points, etc.). The computed power value of 0.70 shows that under the design as specified—detecting a change in proportions from 0.4 to 0.5, at 5% significance level; randomizing 1 cluster at a time to the intervention, each of size 12; and collecting data for each of these 10 time points—the study will have 70% power.

```
. steppedwedge, binomial power complete(1) vartotal(1) p1(0.4) p2(0.5) m(12)
> k(1) rho(0.01) alpha(0.05) steps(10) dm(1)
Power calculation for a stepped wedge study.
For a two sample comparison of proportions (using normal approximations).
For the user specified variables:
Proportion 1:                                                  0.4000
Proportion 2:                                                  0.5000
The variance  has been specified as being the total variance
Significance level:                                            0.05
Coefficient of variation (of clusters):                        0.12
Intra Cluster Correlation (ICC):                               0.0100
Between cluster variation (tau-squared):                       0.0024
Average cluster (cell) size:                                   12
Number of clusters randomised per step:                        1
Number of steps (not including baseline):                      10

Steppedwedge estimated parameters:
Design pattern matrix:

Xtmp[10,11]
```

```
      c1   c2   c3   c4   c5   c6   c7   c8   c9  c10  c11
 r1    0    1    1    1    1    1    1    1    1    1    1
 r2    0    0    1    1    1    1    1    1    1    1    1
 r3    0    0    0    1    1    1    1    1    1    1    1
 r4    0    0    0    0    1    1    1    1    1    1    1
 r5    0    0    0    0    0    1    1    1    1    1    1
 r6    0    0    0    0    0    0    1    1    1    1    1
 r7    0    0    0    0    0    0    0    1    1    1    1
 r8    0    0    0    0    0    0    0    0    1    1    1
 r9    0    0    0    0    0    0    0    0    0    1    1
r10    0    0    0    0    0    0    0    0    0    0    1
number of observations will be reset to 10
Press any key to continue, or Break to abort
obs was 0, now 10
Total number of observations:                          1320
Power:                                                 0.6998
```

4.2 Example 2: Illustration of an incomplete design

In a variation of the example above, we illustrate how `steppedwedge` can be used to compute detectable differences under an incomplete design. To determine efficacy, as opposed to effectiveness, we limit the period of data collected after exposure to the educational intervention to 12 weeks post exposure, thereby circumventing any waning of treatment effect. Under this design variation, data are collected prospectively from a fixed period in time (commencement of study) but limited to 12 weeks following implementation of the training in each team, which will mean staggered data endpoints and, hence, an incomplete design. This design pattern is illustrated in figure 7.

Team	Week																					
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.	.	.
2	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.	.
3	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.	.
4	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.	.
5	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.	.
6	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.	.
7	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.	.
8	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.	.
9	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	.
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1

Figure 7. Illustration of an incomplete design-pattern matrix for example 2

This example also illustrates how `steppedwedge` can be used to determine the detectable difference (for a given level of power).

We have not illustrated the dialog tabs for this example, because they are simply modified from those in example 1. So on the **Menu** tab, it is necessary to select *Compute Detectable Difference* instead of *Compute Power* and set the *Power of test* to `0.8` (the default). On the **Clusters** tab, the design is set as *Incomplete design* rather than *Complete design*, and the *Print design pattern* is unchecked. The user also needs to specify the *Average cluster size* and not the *No. time points* or the *Number of clusters per step* randomized at each step (because these are provided within the design pattern). On the **Values** tab, it is only necessary to provide *Proportion 1* and not *Proportion 2*. Finally, the design pattern (illustrated in figure 7) must be pasted into the Data Editor. The output for this calculation is provided below. Under this modification, with power 80%, it will be possible to detect a difference in proportions from 40% to 51%:

```
. steppedwedge, binomial detectabledifference incomplete(1) vartotal(1) p1(0.4)
> m(12) rho(0.01) alpha(0.05) beta(0.8)

Detectable difference calculation for stepped wedge study.
For a two sample comparison of proportions (using normal approximations)
without continuity correction.

For the user specified variables:

Proportion 1:                                                   0.40
The variance  has been specified as being the total variance
Significance level:                                             0.05
Power:                                                          0.80
Average cluster (cell) size:                                    12
Intra Cluster Correlation (ICC):                                0.0100
Coefficient of variation (of clusters):                         0.12
Between cluster variation (tau-squared):                        0.0024

Steppedwedge estimated parameters:
Total number of observations:                                   2100
Detectable difference:                                          0.1096
with corresponding (decreasing) proportion 2:                   0.2904
or alternatively for (increasing) proportion 2:                 0.5096
```

4.3 Example 3: Illustration of a stepped-wedge study with a transition period

A stepped-wedge study is designed to evaluate the effectiveness of a new educational package in surgical theaters with the aim of reducing in-hospital mortality. It is envisaged that 12 hospitals (clusters) will participate and that the educational package will be rolled out sequentially at a rate of 3 hospitals simultaneously. For each hospital, it is further envisaged that there will be two periods (each three months in length) of data collection, a transition period (again three months in length) in which the educational package is delivered (and no data will be collected), and a further two periods (each three months in length) of data collection post exposure. The design-pattern matrix is illustrated in figure 8.

Hospital	Month							
	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24
1	0	0	.	1	1	.	.	.
2	0	0	.	1	1	.	.	.
3	0	0	.	1	1	.	.	.
4	.	0	0	.	1	1	.	.
5	.	0	0	.	1	1	.	.
6	.	0	0	.	1	1	.	.
7	.	.	0	0	.	1	1	.
8	.	.	0	0	.	1	1	.
9	.	.	0	0	.	1	1	.
10	.	.	.	0	0	.	1	1
11	.	.	.	0	0	.	1	1
12	.	.	.	0	0	.	1	1

Figure 8. Illustration of a design-pattern matrix for example 3

It is expected that there will be about 1,250 surgical procedures per hospital over each 3-month period and that the mortality rate is currently about 12%. The between-hospital variability in this example is parameterized by the CV, and we checked the sensitivity to values between 0.1 and 0.4.

Assuming normal approximations and a binary outcome, we present the output for this detectable-difference calculation below. Simple replications of this calculation show that the difference detectable under this design ranges from a reduction of 9.5% to 10.0% depending on the CV (range 0.1 to 0.4). Output is provided below under the assumption that the CV is 0.3.

```
. steppedwedge, binomial detectabledifference incomplete(1) vartotal(1)
> p1(0.12) m(1250) cluster_cv(0.3) alpha(0.05) beta(0.8) dm(1)

Detectable difference calculation for stepped wedge study.
For a two sample comparison of proportions (using normal approximations)
without continuity correction.

For the user specified variables:

Proportion 1:                                                    0.12
The variance  has been specified as being the total variance
Significance level:                                              0.05
Power:                                                           0.80
Average cluster (cell) size:                                     1250
Intra Cluster Correlation (ICC):                                 0.0123
Coefficient of variation (of clusters):                          0.30
Between cluster variation (tau-squared):                         0.0013

Steppedwedge estimated parameters:
Design pattern matrix:

Xtmp[12,8]
     c1  c2  c3  c4  c5  c6  c7  c8
 r1   0   0   .   1   1   .   .   .
 r2   0   0   .   1   1   .   .   .
 r3   0   0   .   1   1   .   .   .
 r4   .   0   0   .   1   1   .   .
 r5   .   0   0   .   1   1   .   .
 r6   .   0   0   .   1   1   .   .
 r7   .   .   0   0   .   1   1   .
 r8   .   .   0   0   .   1   1   .
 r9   .   .   0   0   .   1   1   .
r10   .   .   .   0   0   .   1   1
r11   .   .   .   0   0   .   1   1
r12   .   .   .   0   0   .   1   1
number of observations will be reset to 12
Press any key to continue, or Break to abort
obs was 0, now 12
Total number of observations:                                    60000
Detectable difference:                                           0.0241
with corresponding (decreasing) proportion 2:                    0.0959
or alternatively for (increasing) proportion 2:                  0.1441
```

4.4 Example 4: Illustration of a stepped-wedge study with a transition period and count outcome

An RCT across 20 wards evaluates the effectiveness of an alternative to the do-not-resuscitate with another approach called the universal form of treatment. The trial will be a CRT with each of the 20 wards acting as a cluster. The trial will run for 28 months, consisting of 12 10-week periods, with 2 wards randomized to the intervention each period (that is, every 10 weeks). In every ward, data will be collected throughout the trial apart from the 10-week period immediately following randomization to allow for a transition period. During every 10-week period and in every ward (apart from transition periods), data relating to harms will be collected on 27 patients. There is also a 10-week period of baseline data collection (during which all wards are on the standard-of-care, do-not-resuscitate arm) and a 10-week period at the end of the study in which all wards are receiving the universal form of treatment intervention. The design pattern is illustrated below.

The effectiveness of the new approach will be evaluated using the outcome of the number of harms. It is expected that in each ward, there will be about 21 harms per 1,000 patient days. Harms will be identified by reviewing patient notes for the duration of their stay, and the average length of stay is about 10 days. The trial has been designed to detect a 25% relative-risk reduction.

This power calculation is therefore used to illustrate the use of the `steppedwedge` command for count outcomes. The current rate of harms is about 0.02 per person per unit of time (here the unit of time is a day). We hope to reduce this by 25%, that is, to 0.015. This is an incomplete design (because there is a transition period) with 20 clusters and an average cluster size of 270 (27 patients each with an average length of stay of 10 days). We have assumed an ICC of 0.01 in illustrative calculations. The output from this example is illustrated below.

```
. steppedwedge, rates power incomplete(1) vartotal(1) r1(0.021) r2(0.015)
> m(270) rho(0.007) alpha(0.05) dm(1)
Power calculation for a stepped wedge study.
For a two sample comparison of rates (using normal approximations).
For the user specified variables:
Rate 1:                                                          0.0210
Rate 2:                                                          0.0150
The variance  has been specified as being the total variance
Significance level:                                              0.05
Coefficient of variation (of clusters):                          0.53
Intra Cluster Correlation (ICC):                                 0.0070
Between cluster variation (tau-squared):                         0.0001
Average cluster (cell) size:                                     270

Steppedwedge estimated parameters:
Design pattern matrix:

Xtmp[20,12]
      c1   c2   c3   c4   c5   c6   c7   c8   c9  c10  c11  c12
 r1    0    .    1    1    1    1    1    1    1    1    1    1
 r2    0    .    1    1    1    1    1    1    1    1    1    1
 r3    0    0    .    1    1    1    1    1    1    1    1    1
 r4    0    0    .    1    1    1    1    1    1    1    1    1
 r5    0    0    0    .    1    1    1    1    1    1    1    1
 r6    0    0    0    .    1    1    1    1    1    1    1    1
 r7    0    0    0    0    .    1    1    1    1    1    1    1
 r8    0    0    0    0    .    1    1    1    1    1    1    1
 r9    0    0    0    0    0    .    1    1    1    1    1    1
r10    0    0    0    0    0    .    1    1    1    1    1    1
r11    0    0    0    0    0    0    .    1    1    1    1    1
r12    0    0    0    0    0    0    .    1    1    1    1    1
r13    0    0    0    0    0    0    0    .    1    1    1    1
r14    0    0    0    0    0    0    0    .    1    1    1    1
r15    0    0    0    0    0    0    0    0    .    1    1    1
r16    0    0    0    0    0    0    0    0    .    1    1    1
r17    0    0    0    0    0    0    0    0    0    .    1    1
r18    0    0    0    0    0    0    0    0    0    .    1    1
r19    0    0    0    0    0    0    0    0    0    0    .    1
r20    0    0    0    0    0    0    0    0    0    0    .    1
number of observations will be reset to 20
Press any key to continue, or Break to abort
obs was 0, now 20
Total number of observations:                                    59400
(This is total length of exposure for rate comparisons)
Power:                                                           0.8237
```

5 Conclusion

Stepped-wedge trials are used with increasing frequency in the evaluation of health care and service delivery interventions. Thus we have developed a Stata command, `steppedwedge`, that will calculate power and detectable difference for stepped-wedge studies, along with a dialog box (which makes this command accessible). A point of particular flexibility is the functionality that allows incomplete designs, a feature that we believe will be of particular use to designers of such pragmatic trials.

The command does, however, have several limitations, including the variance of two proportions, assuming normality, not incorporating continuity corrections, and assuming variance parameters are known. The command also assumes that cluster sizes are equal. In CRTs, approximations to the design effect for unequal cluster sizes have been developed. While not formally shown, we believe that the harmonic mean may be a more useful measure of cluster sizes in the case of unequal clusters. Also, because there is no individual random effect, the command and model described are limited to cross-sectional designs only and cannot be used for cohort designs. It is assumed that at each step in the study, a different cross-section of individuals is included. Perhaps most importantly, the command will not calculate sample-size required for a given level of power to detect a given difference. This, however, is a difficult calculation, and more methodological research is needed before such a command can be programmed.

6 Funding acknowledgments

Karla Hemming was partially funded by a National Institute of Health Research grant for Collaborations for Leadership in Applied Health Research and Care for part of the duration of this work. The views expressed in this article are not necessarily those of the National Institute of Health Research or the Department of Health. Alan Girling is funded by the Engineering and Physical Sciences Research Council Multidisciplinary Assessment of Technology Centre for Healthcare program (Engineering and Physical Sciences Research Council Grant GR/S29874/01).

7 References

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

Batistatou, E., C. Roberts, and S. Roberts. 2014. Sample size and power calculations for trials and quasi-experimental studies with clustering. *Stata Journal* 14: 159–175.

Brown, C. A., and R. J. Lilford. 2006. The stepped wedge trial design: A systematic review. *BMC Medical Research Methodology* 6: 54.

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.

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.

Hemming, K., and J. Marsh. 2013. A menu-driven facility for sample-size calculations in cluster randomized controlled trials. *Stata Journal* 13: 114–135.

Hussey, M. A., and J. P. Hughes. 2007. Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials* 28: 182–191.

Mdege, N. D., M. S. Man, C. A. Taylor Nee Brown, and D. J. Torgerson. 2011. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *Journal of Clinical Epidemiology* 64: 936–948.

About the authors

Karla Hemming is a senior lecturer in biostatistics in the department of Public Health, Epidemiology and Biostatistics at the University of Birmingham.

Alan Girling is a reader in medical statistics in the department of Public Health, Epidemiology and Biostatistics at the University of Birmingham.