Institute for Fiscal Studies

Going beyond simple sample size calculations: a practitioner's guide

IFS Working Paper W15/17

Brendon McConnell
Marcos Vera-Hernández

Going Beyond Simple Sample Size Calculations: a Practioner's Guide

Brendon McConnell and Marcos Vera-Hernández*

June 2015

Abstract

Basic methods to compute the required sample size are well understood and supported by widely available software. However, the sophistication of the sample size methods commonly used has not kept pace with the complexity of the experimental designs most often employed in practice. In this paper, we compile available methods for sample size calculations for continuous and binary outcomes with and without covariates, for both clustered and non-clustered RCTs. Formulae for panel data and for unbalanced designs (where there are different numbers of treatment and control observations) are also provided. The paper includes three extensions: (1) methods to optimize the sample when costs constraints are binding, (2) simulation methods to compute the power of a complex design, and (3) methods to consider in the sample size calculation adjustments for multiple testing. The paper is provided together with spreadsheets and STATA code to implement the methods discussed.

Keywords: Power analysis, Sample size calculations, Randomised Control Trials, Cluster Randomised Control Trials, Covariates, Cost Minimisation, Multiple outcomes, Simulation
JEL Codes: C8, C9

*We gratefully acknowledge the ESRC-NCRM Node 'Programme Evaluation for Policy Analysis' (PEPA), ESRC Grant reference ES/I02574X/1 for funding this project. We benefited from comments received at a workshop of The Centre for the Evaluation of Development Policies at The Institute for Fiscal Studies. All errors are our own responsibility. Author affiliations and contacts: McConnell (IFS, brendon.mcconnell@gmail.com); Vera-Hernández (UCL and IFS, m.vera@ucl.ac.uk)

1 Introduction

One of the big challenges in economics has been to estimate causal relationships between economic variables and policy instruments. Randomized Controlled Experiments (RCT) have become one of the main tools that researchers use to accomplish this objective[1](Hausman and Wise, 1985; Burtless, 1995; Heckman and Smith, 1995; Duflo et al., 2007). More simple RCTs are usually set up with the objective of estimating the impact of a certain policy or intervention, while more complex RCTs can be implemented to test the competing hypotheses that explain a phenomenon (also known as field experiments, see Duflo (2006) and Levitt and List (2009)).

When setting up a RCT, one of the first important tasks is to calculate the sample size that will be used for the experiment. This is to ensure that the planned sample is large enough to detect expected differences in outcomes between the treatment and control group. A sample size that is too small leads to a underpowered study, which will have a high probability of overlooking an effect that is real. The implications of small sample sizes go beyond that, low power also means that statistically significant effects are likely to be false positives[2]. Studies with samples larger than required also have their drawbacks: they will expose a larger pool of individuals to an untested treatment, they will be more logistically complex and will be more expensive than necessary.

Basic methods to compute the required sample size are well understood and supported by widely available software. However, the sophistication of the sample size formulae commonly used has not kept pace with the complexity of the experimental designs most often used in practice. RCTs are usually analysed using data collected before the intervention started (baseline data) but this is often ignored by the sample size formulae commonly used by researchers, as is the inclusion of covariates in the analysis. Another departure from the basic design is that interventions are commonly assessed not just on a single outcome variable but more than one, creating problems of multiple hypotheses testing that should be taken into account when computing the required sample size. Depending on the context and specific assumptions, taking into consideration some of these departures from the basic design will lead to smaller or larger sample sizes.

The objective of this paper to provide researchers with a practioner's guide, supported by software, to allow them to incorporate in their sample size calculations certain features that are commonly present in RCTs and that are often ignored in practice. Although most of the content is not novel, most of it is dispersedly published in quite diverse notation, making it difficult for the applied researcher to find the right formulae just at the busy time when he/she is writing the research proposal that will fund the RCT. We also note that understanding the sample size implications of different design features can be very useful to design the RCT (what waves of data to collect, what information to collect, etc.)

The article will include sample size calculations for both continuous and binary out-

[1]See Blundell and Costa Dias (2009) and Imbens and Wooldridge (2009) for reviews on non-experimental methods.

[2]An intuitive explanation using a numeric example can be found in The Economist (2013) "Trouble at the lab", 19th October

comes, starting with the simplest case of individual level trials, and then cluster randomised trials. We will also cover how to take into account pre-intervention data, as well as covariates. Along the paper, we favour simplicity in exposition and attempt to keep the language accessible to the applied researcher who does not have previous exposure to sample size calculations. The article has three extensions: cost minimization, simulation methods, and sample size estimation for multiple outcomes. The first extension explains how to allocate the sample in order to minimize costs. It is well known, but little used in practice, that if the budget must cover both intervention and data collection costs, then the same level of power can be achieved at smaller costs if more units are allocated to the control arm than the treatment. The second extension explains how to compute the power using simulation methods, which is useful when there are no existing formulae for the RCT that is being planned. The last extension shows how to adapt the sample size computations when several outcomes are used.

An inherent difficulty in using the sample size formulae that we provide in the paper is that assumptions are needed on some key parameters of the data generating process, which are not required by the basic formulae. Our view is that the widespread trend towards making data publicly available, including the data used in academic publications, will definitively help researchers to find realistic values for the parameters of interest. Moreover, social science journals might follow the trend set by medical journals on making compulsory for authors to report certain key estimates which are commonly used in sample size calculations (Schulz et al., 2010).

The paper is organized as follows. Section 2 provides an overview of an example intervention that will be used throughout the paper, section 3 provides an overview of basic concepts involved in power calculations, section 4 considers power calculations for continuous outcomes, section 5 focusses on discrete outcomes, section 6 outlines several extensions and section 7 concludes. As supplementary material, we include: (1) a training dataset that can be used to calculate the parameters which are relevant for the sample size calculations, (2) spreadsheets to use the methods proposed for continuous outcomes, as well as STATA programs for discrete ones (in Appendix D). We also provide examples of STATA code to estimate key parameters needed to perform sample size calculations in Appendix B and code to compute power through simulations in Appendix C.

2 Overview of an Example Intervention

In this section, we will set up an example that we will use for the rest of the paper. Let's assume that we would like to evaluate APRENDE2, a job-training program that will be implemented by the Colombian government. The Colombian government will run a Randomized Controlled Trial (RCT) to evaluate APRENDE2. Our task is to compute the required sample size for such evaluation. The main outcomes of interest are individual's earnings, and the proportion of individuals that work at least 16 hours a week, the median in the sample.

As it will be clear later on, to be able to compute the sample size requirements,

we will need some basic parameters, such as average earnings, the standard deviation of earnings, and the proportion of individuals that work at least 16 hours a week. We are at the planning stage, so we have not collected the data yet, and hence we do not know for certain what these parameters would be in our target population. We may use grey literature or published studies that report these parameters in our context, or in a context similar to the one that we will be working on.

In this particular study, we have been fortunate that the government previously evaluated APRENDE, a different program to APRENDE2 but with the same beneficiary population. The dataset used to evaluate APRENDE contains the key variables that we need for the sample size calculation of APRENDE2.[3] For instance, for each person in the sample, the dataset contains the earnings, and whether the person is working, as well as some other additional variables that we might use as covariates in the analysis. This information is available for several years, and it also contains an identifier for the town where the person lives, features that will be important when we carry out more complicated analyses.

APRENDE2 might be evaluated in two ways, either as an individual based RCT or a cluster based one. In the former, a few pilot towns will be chosen, and a list compiled of the eligible individuals interested in participating in APRENDE2 in those towns. Within each town, a lottery will be used to decide which individuals are chosen to participate in APRENDE2 in this pilot phase, and which are randomized out. Alternatively, a cluster RCT could be used in which a random mechanism would split the set of towns that are part of the evaluation into treatment and control. Eligible individuals living in treatment towns can apply and participate in APRENDE2. In this case, we will say that the town is the cluster because it is the unit of randomization (but the data used for the evaluation will be collected at a more disaggregated level, i.e., the individual). Other examples of commonly used clusters are schools, job centres, primary care clinics, etc.

One of the main parameters needed to compute the sample size requirements is the effect size, which is the smallest effect of the policy that we want to have enough power to detect. When considering the effect size for an individual based RCT, we must take into account that it refers to the comparison in the outcome levels of individuals initially allocated to treatment versus control. Note that this difference will be diluted by any non-compliance (i.e. individuals initially allocated to treatment that eventually decide not to participate), and hence we must adjust the effect size accordingly. For instance, if we think that APRENDE2 will increase participants' average earnings in 14,000 but 30% of individuals initially allocated to participate in APRENDE2 decide not to take it up, we must plan for a diluted effect size of 9,800 (=14,000*0.7) as this will take into account the non-compliance rate.

In a cluster RCT, the relevant comparison is the differences in outcome levels between the eligible individuals living in treatment towns (irrespective of whether they participated or not) and the eligible individuals living in control ones (also, irrespective

[3]More generally, one could use a general puporse household survey from a similar enviroment if it contains the right variables.

of whether they participated or not). Because not all eligible individuals living in treatment towns will end up participating, the coverage rate of the policy must be taken into account when considering the effect size. Assuming that APRENDE2 increases participants' average earnings by 14,000, we should plan for an effect size of 8,400 (=14,000*0.6) if the coverage rate is expected to 60% (it is expected that 40% of the eligible population living in the treatment towns will not participate in APRENDE2, either because of capacity constraints or because they are not interested). Of course, the effect size would have to be even smaller if we think that individuals in control towns can travel to treatment towns and participate in APRENDE2 (contamination). For instance, if 10% of individuals living in control towns could do that, then the planned effect size would have to be 7,000 (= 14,000*(0.6-0.1)).

3 Basic Concepts

When computing the required sample size for an experiment, one of the most important questions that the researcher must answer is what is the smallest difference in the average of the outcome variable between treatment and control that she would like the study to be able to detect. The answer to this crucial question is the the effect size (sometimes referred to as the minimum detectable effect or MDE in the literature), denoted below as δ.

For those unfamiliar with sample size calculations, this may be a slightly strange concept, as in order to calculate the sample size for a trial, we need to input the impact we expect the trial to have. It is common to refer to existing literature in order to get a sense of this effect size. Of course, the results from previous literature must be contextualized to the study that is being planned. For instance, the researcher might think that APRENDE2 should be less effective than existing studies, maybe because it targets all ages, rather than the youth. Differences in expected non-compliance and contamination between APRENDE2 and other existing studies will also modify the effect size that we will plan for. Nothing precludes the researcher from conducting sample size calculations with several different values of the effect size to gauge the sensitivity of the results.

Assessing whether the intervention being tested had an actual effect on the average of the outcome variable is challenging because, usually, we will not have data on the entire population of treatment and control individuals and clusters. In most cases, we will simply have data on a sample of them. Because of sampling variation, the sample average of the outcome variable of the treatment group individuals will most likely be different from the control group one. What needs to be assessed is whether this difference is large enough as to indicate that it is due to actual differences on the population level values of the outcome variable, which will be attributed to the intervention, or whether it is small enough so that it could be due to sampling noise. This is where hypothesis testing is being called for. The null hypothesis (H_0) will usually be that the population mean of the outcome variable in the treatment group will be the same as in the control group. In other words, the null hypothesis is that the intervention was on average ineffective,

and the alternative hypothesis that the effect of the intervention is δ (the difference in the population mean of the outcome variable between treatment and control, which we call the effect size).

When conducting the hypothesis test, two possible errors are likely to happen. On the basis of the sample at hand, and the test carried out, the researcher could reject a true null hypothesis, that is, to conclude that the intervention was effective when it was not. This type of "false positive" error is usually called a Type I error (see Figure 1). The other possible error is to conclude that that the intervention had no effect when one exists (fail to reject the null hypothesis if it is false). This type of "false negative" error is called a Type II error.

The researcher will never be able to know whether a Type I or Type II error is being committed, because the truth is never fully revealed. But the researcher can design the study as to control the probability of committing each type of error. Significance, usually denoted by α, is the probability of committing a Type I error (Prob[reject $H_0|H_0$ true] $= \alpha$). Commonly α is set to equal .05[4]. This means that when the null is true, we won't reject it in 95% of cases. The probability of a Type II error, denoted by β, is the chance of finding no intervention effect, when one exists (Prob[fail reject $H_0|H_1$ true] $= \beta$). Common values of β are between 0.1 and 0.2.

Power is defined as $1-\beta$, that is, Prob[reject $H_0|H_1$ true]. In our context, Power is the probability that the intervention is found to have an effect on the mean of the outcome variable when there is a genuine effect. Put more bluntly, power is the probability that a study has of uncovering a true effect. The researcher would like the power to be as high as possible; otherwise it has a high chance of overlooking an effect that is real. Usually, Power of 0.8 or 0.9 are considered high enough (consistent with values of β between 0.1 and 0.2).

For the continuous outcome case, we will need to know the variance[5] of the outcome, σ^2. Again, one can get a sense of this from previous studies or from a pilot study if one has been conducted. The parameters mentioned above are the minimum set of parameters for which we need to have estimates to calculate the required sample size for the experiment.

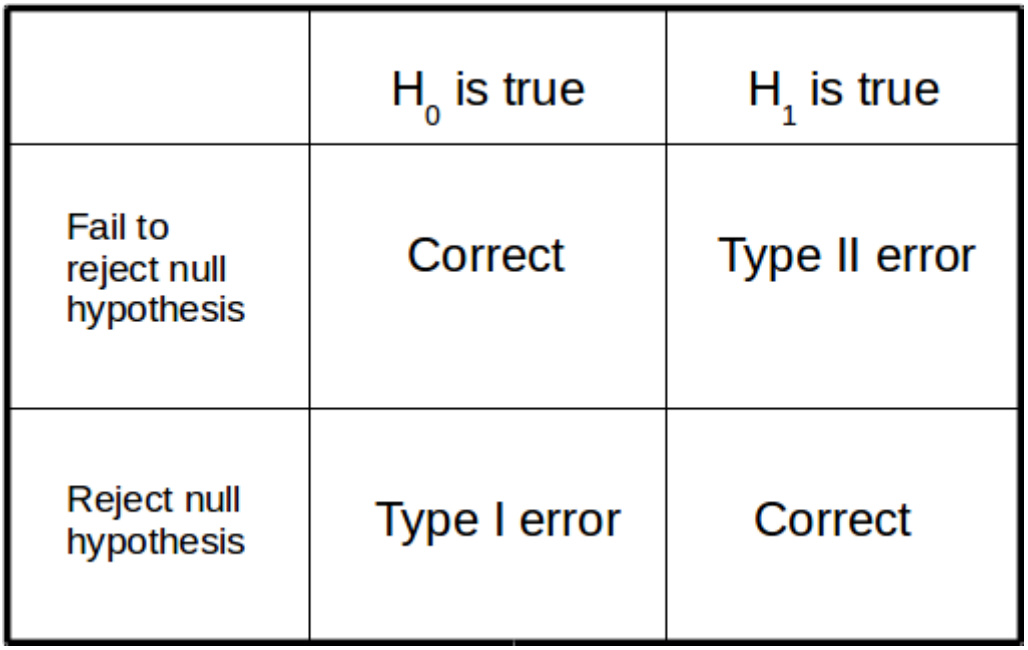
There is an additional input to take into consideration when calculating power for cluster randomised trials. This is the intracluster correlation (ICC), which is a measure of how correlated the outcomes are within clusters. This parameter, denoted here as ρ and defined below, can be estimated from a pilot survey or based on measures found in the exisiting literature. This parameter plays an important role in sample size calculations for cluster randomised trials, and can lead to one requiring much larger sample sizes than in the individual level randomisation case[6]. The reason for this is that the larger is the correlation of outcomes amongst individuals within clusters, the less informative an extra individual sampled within the cluster is. Adding an extra cluster of k individuals

[4]Later in the paper, we will discuss testing for multiple outcomces, which will affect the value chosen for α

[5]In the binary case, the variance of the outcome is a function of the mean

[6]Where covariates are included, it is the conditional ICC that will be used in the calculations below. This may be harder to obtain from previous studies.

Figure 1: Type I and Type II Errors



will result in greater power rather than including k more individuals across existing clusters.

4 Continuous Outcomes

Here we derive the sample size calculation for the simple case of a RCT in which the treatment, T, is randomized at the individual level, and the outcome variable, Y, is continuous. This simple case allows us to focus on the main steps that are necessary to derive the sample size formulae, and it is useful to give a sense of how the other formuale used throughout this paper are derived.[7] Usually, we test whether T had an effect on Y by testing whether the population means of Y are different in the treatment than in the control group. More formally, if we denote the population means in the treatment and control groups by μ_1 and μ_0, respectively, the null hypothesis is H_0: μ_1-μ_0=0; and the alternative hypothesis, that the difference in the population means equals the MDE, by H_1:μ_1-μ_0=δ.

Assume that we have a sample of n_0 individuals in the control group, and a sample of n_1 individuals in the treatment group. We denote by $T_i = 0$ or $T_i = 1$ if individual i is part of the control or treatment group respectively. To test H_0 against H_1, we would

[7]The material in this section is standard of statistical textbooks. In this section, we follow Liu (2013) closely.

estimate the following OLS regression[8]:

$$Y_i = \alpha + \beta T_i + \epsilon_i,$$

where Y_i is the value of the outcome variable (say earnings in the case of APRENDE2) and ε_i is an error term with zero mean and variance σ^2, which for the time being we assume it is known. The z-statistic associated with β is given by the OLS estimate of β divided by its standard error, that is:

$$Z = \frac{\bar{Y_1} - \bar{Y_0}}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}},$$

where $\bar{Y_1}$ and $\bar{Y_0}$ are the sample averages of Y_i for individuals in the treatment and control group respectively, and σ is the standard deviation of Y_i. If the null hypothesis is true, then μ_1=μ_0, and Z follows a Normal distribution with zero mean and variance of one. Hence, the null hypothesis will be rejected at a significance level of α if $Z\geq$ $z_{\alpha/2}$ or $Z\leq -z_{\alpha/2}$, where the cumulative distribution function of the standard Normal distribution evaluated at $z_{\alpha/2}$ is $1-\alpha/2$.

As mentioned above, power, denoted by $1-\beta$, is the probability of rejecting the null hypothesis when the alternative is correct, that is,

$$1-\beta = Pr(Z \leq -z_{\alpha/2} \cup\ Z \geq z_{\alpha/2}|H_1) = Pr(Z \leq -z_{\alpha/2}|H_1) + Pr(Z \geq z_{\alpha/2}|H_1)$$

Because the altenative hypothesis is correct, μ_1-μ_0 is no longer zero, but δ. Hence, the mean of Z is no longer zero but $\delta/(\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}})$. In this case, $Pr(Z \leq -z_{\alpha/2}|H_1)$ is approximately zero, and hence we have that[9]

$$1-\beta = Pr(Z \geq z_{\alpha/2}|H_1) = 1 - Pr(Z < z_{\alpha/2}|H_1)$$

By substracting the mean of Z under the altenative hypothesis from both sides of the inequality, we obtain

$$\beta = Pr(Z - \frac{\delta}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}} < z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}})$$

Because the left hand side of the inequality now follows a Normal distribution with zero mean and unit variance, it is the case that

$$z_{1-\beta} = z_{\alpha/2} - \frac{\delta}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}},$$

which implies that[10]

[8]We use a regression framework to keep the parallelism with forthcoming sections, but a t-test for two independent samples is equivalent

[9]See for instance Liu (2013). Note, however, that Liu (2013) defines $z_{\alpha/2}$ such that the cumulative distribution function of the standard Normal distribution evaluated at $z_{\alpha/2}$ is $\alpha/2$ instead of $1-\alpha/2$.

[10]Note that $z_\beta = -z_{1-\beta}$.

$$z_\beta + z_{\alpha/2} = \frac{\delta}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$$

In the case in which σ is unknown and is estimated using the standard deviation in the sample, then a t distribution with $v = n_0 + n_1 - 2$ degrees of freedom must be used instead of the Normal distribution. In this case, we have that

$$t_\beta + t_{\alpha/2} = \frac{\delta}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$$

Solving for δ, we obtain the expression for the MDE that can be detected with $1 - \beta$ power at significance level α:

$$\delta = (t_\beta + t_{\alpha/2})\sigma\sqrt{\frac{1}{n_0} + \frac{1}{n_1}} \tag{1}$$

Alternatively, assuming that the sample size in the treatment and control groups are the same, $n_0 = n_1 = n^*$, the expression for the sample size of each arm is given by:

$$n^* = 2(t_\beta + t_{\alpha/2})^2\frac{\sigma^2}{\delta^2} \tag{2}$$

It should be noted from equation 2 that for the results of a power calculation to be meaningful, one must have accurate estimates of both the minimum detectable effect and the variance of outcomes, as both are key inputs.

Although subsequent equations will be more complicated, the derivation of these all follows a similar approach to that above.

Finally for this section, we outline the case where variances are unequal, following List et al. (2011). This case is uncommon in practice, as it is difficult *a priori* to consider how the treatment will affect not just the mean of the outcomes, but the variance too. On example could be the provision of weather-linked insurance to farmers. Here one would expect the variance of consumption to decline for the treated individuals. Total sample size is defined as $N = n_0 + n_1$ and

$$\delta = (t_\beta + t_{\alpha/2})\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$$

After working through a series of equations (see Appendix A for the full derivation), we derive a formula for N^* and the optimal allocations of n_0 and n_1:

$$N^* = (t_\beta + t_{\alpha/2})^2\frac{1}{\delta^2}\left(\frac{\sigma_0^2}{\pi_0^*} + \frac{\sigma_1^2}{\pi_1^*}\right), \tag{3}$$

where $\pi_0^* = \frac{\sigma_0}{\sigma_0+\sigma_1}$ and $\pi_1^* = \frac{\sigma_1}{\sigma_0+\sigma_1}$, and $n_0^* = \pi_0^* N^*$ and $n_1^* = \pi_1^* N^*$. From this it can be seen that the group with the larger variance is allocated a greater proportion of the sample.

4.1 Clustered

In many cases, the outcome variable is measured at the individual level, but the randomisation takes place at the cluster level (school, village, firm)[11]. This may be driven by concerns over spillovers within a cluster, whereby individual level randomisation would lead to control members outcomes being contaminated by those of treated individuals. In this case the sample size formula must be adjusted to reflect that observations from individuals of the same cluster are not independent, as they may share some unobserved characteristics.

The estimating equation will take the form of

$$Y_{ij} = \alpha + \beta T_j + v_j + \epsilon_{ij}, \tag{4}$$

where i denotes individual, and j denotes the cluster. T_j is the treatment indicator, v_j and ϵ_{ij} are eror terms at the cluster and individual level. The variances of v_j and ϵ_{ij} are given by $var(v_j) = \sigma_c^2$ and $var(\epsilon_{ij}) = \sigma_p^2$, and $\sigma_c^2 + \sigma_p^2 = \sigma^2$.

To carry out the sample size calculation in the presence of clustering, we require an additional input; the intracluster corrrelation or ICC, denoted here as ρ:

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_p^2}$$

The ICC thus gives a measure of the proportion of the total variance accounted for by the between variance component. The intuition behind the ICC is that the larger the fraction of the total variance accounted for by the between cluster variance component (σ_c^2), the more similar are outcomes within the cluster, and the less information is gained from adding an extra individual within the cluster. Proceeding as in the simple case above, we derive the following equation[12]:

$$\delta^2 = (t_{\alpha/2} + t_\beta)^2 2 \left(\frac{m\sigma_c^2 + \sigma_p^2}{mk} \right), \tag{5}$$

where there are k clusters per treatment arm and m individuals per cluster[13]. Using the definition of the ICC, and rearranging, we arrive at the formula for the total sample per

[11] We use the term individual to denote the level at which the observation is measured. In most cases it will be people or households but it could also be firms (if the cluster is a town and the outcome variable is the profits of small businesses).

[12] With clustering, and assuming equal variances for the two groups, the standard error of $\hat{\beta}$ takes the form:

$$\sqrt{\left(\frac{\sigma_c^2}{k} + \frac{\sigma_p^2}{mk} \right) + \left(\frac{\sigma_c^2}{k} + \frac{\sigma_p^2}{mk} \right)} = \sqrt{2 \frac{m\sigma_c^2 + \sigma_p^2}{mk}}$$

[13] In the clustered case, the degrees of freedom in the t distribution are $2(k-1)$.

treatment arm[14]:

$$n^* = m^* k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}(1 + (m-1)\rho) \tag{6}$$

Comparing equations 2 and 6, the only difference is the term $(1 + (m-1)\rho)$, which is commonly referred to in the literature as either the design effect or the variance inflation factor (VIF). This term is a consequence of the clustered treatment allocation, and will lead to larger required sample sizes .

In order to get a sense of the interplay between the ICC and the number per cluster, Table 1 presents required sample sizes for two different values of δ. We use the APRENDE data here in order to get a standard deviation value, as well as reasonable values for δ. The actual ICC for this data is .042 - within the ranges of ICC values presented in Table 1[15]. Consider first the upper left quadrant. The case where ICC=0 represents individual level randomisation. As the ICC increases, so too does the sample size. The extent of the increases depends also on m, the other key term in the VIF. One can see that for a ρ=.03, a rather small value for the ICC, and m=60, a cluster randomised trial requires almost triple the sample size to that of an individual level equivalent (7083 compared to 2557).

Another way to see this is to consider the upper right quadrant. At low levels of the intracluster correlation, there is a marked decline in the required number of clusters as we increase m (the number of individuals per cluster). For $\rho = .01$, k drops from 279 to 51, 18% of the initial value as we move move from left to right. As the ICC increases, this decline is much shallower. For $\rho = .2$ the right-hand value for k is 74% of the initial value. It should be clear from this table that its very important to get accurate measures of key input parameters. Small differences in these values, such as moving from $\rho = .01$ to .03, can have significant impacts on the required sample size, particularly when m is large.

Finally, by comparing the upper and lower sections of Table 1, we see the impact of the size of the MDE - the larger is the value of δ, the smaller is the sample size required to detect a statistically significant effect.

4.1.1 Unequal numbers of clusters

It is useful to know what the equations look like for uneven allocations of both the number of clusters per treatment arm, k, and the number of individuals within these clusters, m. This may be due to restrictions imposed on the size of one of the treatment arms, for example because of logistical constraints. It should be noted that departing from an equal split for the two groups leads to a larger total required sample size, so

[14]To operationalise this formula one can either solve for m as a function of k or solve for k as a function of m. In the latter case (due to the fact that the degrees of freedom of the t distribution are a function of the number of clusters $(2(k-1)$ in the absence of covariates)), it neccessary to use an iterative process to ensure that the correct degrees of freedom $(2(k^* - 1))$ are used to calculate the optimal number of clusters, k*. This issue will be more pronounced when the number of clusters is small.

[15]In Appendix B we show how to compute the ICC using STATA.

this should be done only when restrictions require so, or when this decreases the overall cost (for instance when control clusters are cheaper than treatment clusters - see section 6.1). That said, total sample size only rises markedly for highly unbalanced allocations. The number of clusters in the treatment arm (k_1) as a function of the number of clusters in the control arm (k_0) is given by:

$$k_1 = \frac{(t_{\alpha/2} + t_\beta)^2 \left(\frac{m\sigma_c^2 + \sigma_p^2}{m}\right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2 \left(\frac{m\sigma_c^2 + \sigma_p^2}{mk_0}\right)} \tag{7}$$

This can also be written in terms of the design effect as:

$$k_1 = \frac{(t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{1+(m-1)\rho}{m}\right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{1+(m-1)\rho}{mk_0}\right)} \tag{8}$$

The formula for the number of individuals per treatment cluster (m_1) as a function of the number of individuals per control cluster (m_0) is given by:

$$m_1 = \frac{(t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma_p^2}{k}\right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2 \left(\frac{2\sigma_c^2}{k} + \frac{\sigma_p^2}{m_0 k}\right)} \tag{9}$$

Rewriting in terms of ρ and σ yields:

$$m_1 = \frac{(t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{(1-\rho)}{k}\right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{1+(2m_0-1)\rho}{m_0 k}\right)} \tag{10}$$

4.2 The Role of Covariates

Although, due to randomisation, covariates are not used to partial out differences between treatment and control, they can be very useful in reducing the residual variance of the outcome variable, and subsequently leading to lower required sample sizes.

There are several different ways of representing the power calculation formula with covariates, which will be presented for completeness, and due to the fact that in different situations, one may only have the required inputs suited to using a single formula.

The simplest or most intuitive version is as follows:

$$n^* = m^* k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma_x^2}{\delta^2}(1 + (m-1)\rho_x), \tag{11}$$

where σ_x^2 is the conditional variance (that is the residual variance once the covariates have been controlled for), and $\rho_x = \frac{\sigma_{x,c}^2}{\sigma_{x,c}^2 + \sigma_{x,p}^2}$ the conditional ICC[16]. The form of equation

[16]In the case with covariates, the number of degrees of freedom of the t tistribution is $2(k-1) - J$, where J is the number of covariates.

11 mirrors that of the unconditional representation in equation 6. If there is baseline data, or data from a similar context with the relevant variables, it is straightforward to get estimates of these conditional parameters[17]. However, if this is not available, and estimates must be gleaned from the existing literature, it may be that these parameters are not directly obtainable. For this reason we present a different form of the conditional power calculation, which use different parameters. Bloom et al. (2007) present the following formula:

$$n^* = m^* k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}(m\rho(1 - R_c^2) + (1 - \rho)(1 - R_p^2)), \tag{12}$$

where R_c^2 is the proportion of the cluster-level variance component explained by the covariates, and R_p^2 the individual-level equivalent. This formulation is useful to see the differing impact of covariates at different levels of aggregation i.e. if the covariates are at individual or cluster level. For instance, an individual covariate can affect both R_p^2 *and* R_c^2, whilst a cluster level covariate can only increase R_c^2. Equation 12 may be useful if R_p^2 and R_c^2 are reported in existing research, and the parameters in equation 11 are not. To reiterate, with a series of calculations, it is straightforward to move from equation 12 to 11, using R_c^2, R_p^2, σ^2 and ρ to obtain values for σ_x^2 and ρ_x[18].

Finally, Hedges and Rhoads (2010) present the formula for the inclusion of covariates as:

$$n^* = m^* k^* = (t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}\left[(1 + (m - 1)\rho) - (R_p^2 + (mR_c^2 - R_p^2)\rho)\right]$$

This equation may be useful for building intuition into the role of covariates, as the first term in parentheses is the regular design effect, whilst the second shows how covariates impact the overall variance inflation factor.

Table 2 presents how the inclusion of both individual and cluster level covariates impact required sample sizes for six different scenarios (m= 8, 20 and 100, ρ=.01 and .3). Values for the standard deviation come from that of the earnings variable for 2002 in the APRENDE data. As it is clear from equation 12, the larger either R_p^2 or R_c^2 is, the smaller the sample size per arm is. Note from equation 12 that the influence of R_p^2 is larger when ρ is smaller. For example, in Table 2, when $\rho = 0.01$, $m = 100$, and R_c^2=0, the sample size per arm decreases from 1351 to 1048 (a 22% reduction) when R_p^2 increases from 0 to 0.5. However, the same increase in the R_p^2 only translates into a decrease from 19342 to 19123 (a 1% reduction) when $\rho = 0.3$. In this sense, increasing R_p^2 is similar to increasing the number of individuals per cluster, which has little effect on power when ρ is high.

[17] Refer to Appendix B to see how to estimate these parameters.

[18] Using the definition of ρ_x, we note that $\sigma^2(1 - \rho)(1 - R_p^2) = \sigma_{x,p}^2 = (1 - \rho_x)\sigma_x^2$ and $\sigma^2\rho(1 - R_c^2) = \sigma_{x,c}^2 = \rho_x\sigma_x^2$. This allows us to write the R^2 terms as functions of ρ, σ^2, ρ_x and σ_x^2: $(1 - R_c^2) = \frac{\rho_x\sigma_x^2}{\rho\sigma^2}$ and $(1 - R_p^2) = \frac{(1-\rho_x)\sigma_x^2}{(1-\rho)\sigma^2}$. These expression are used in intermediate steps to move from equation 12 to equation 11.

As it is also clear from equation 12, the effect of R_c^2 is mediated by $m\rho$, so the reduction in sample size achieved by increasing R_c^2 will be higher when both m and ρ are large. Again, increasing R_c^2 is analogous to increase the number of clusters. This will have a larger effect when ρ is large and when m is large (because a large m indirectly implies that the number of clusters is small, so we obtain a larger effect when we increase them). This is also clear in Table 2: when $\rho = 0.3$, $m = 100$, and R_p^2=0, the sample size per arm decreases from 19342 to 9940 (a 48% reduction) when R_c^2 increases from 0 to 0.5. However, the same increase in the R_p^2 only translate in a decrease from 679 to 654 (a 3.6% reduction) when $\rho = 0.01$ and $m = 8$.

A final point to note here is an issue raised by Bloom et al. (2007) regarding unconditional versus conditional ICCs. As mentioned by the authors, one should not be concerned with the possibility that an individual level covariate, by reducing the individual level variance component by a larger extent than the cluster level component, may lead to a higher conditional ICC. What matters is that by reducing both components, individual level covariates increase precision and thus lower required sample sizes. This issue is not a concern with cluster level covariates, which can only impact σ_c^2, thus will always lead to conditional variances and ICCs that are smaller than their unconditional counterparts.

4.2.1 Unequal numbers of clusters

As we presented in Section 4.1.1, we can write down the sample size equations where either k or m are unequal.

First, consider the expression for k_1 as a function of k_0 and m, written in the form presented by Bloom et al. (2007)[19]:

$$k_1 = \frac{(t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{m\rho(1-R_c^2)+(1-\rho)(1-R_p^2)}{m} \right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{m\rho(1-R_c^2)+(1-\rho)(1-R_p^2)}{mk_0} \right)} \tag{13}$$

As before, we can also write an expression for m_1 as a function of k and m_0[20]:

$$m_1 = \frac{(t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{(1-\rho)(1-R_p^2)}{k} \right)}{\delta^2 - (t_{\alpha/2} + t_\beta)^2 \sigma^2 \left(\frac{2m_0\rho(1-R_c^2)+(1-\rho)(1-R_p^2)}{m_0 k} \right)} \tag{14}$$

4.3 Difference-in-differences and lagged outcome as a covariate

Where the researcher has not only data on the outcome variable subsequent to treatment, but also prior to treatment (baseline), it is possible to employ a difference-in-differences approach, as well as to include the baseline realisation of the outcome variable as a

[19] We can also write an expression for k_1 in the form of either equation 7, where we replace σ_c and σ_p with $\sigma_{x,c}$ and $\sigma_{x,p}$ or equation 8, where we replace σ and ρ with σ_x and ρ_x.

[20] We can also write an expression for m_0 in the form of either equation 9 or 10, replacing unconditional parameters with their conditional versions.

covariate, a special case of the approach above. Following Teerenstra et al. (2012) the data generating process (which includes the panel component) follows:

$$Y_{ijt} = \beta_0 + \beta_1 T_j + \beta_2 POST_t + \beta_3 (POST_t \times T_j) + v_j + v_{jt} + \epsilon_{ij} + \epsilon_{ijt},$$

where i indexes individuals, j clusters, and t time periods (t=0, the pre intervention period, or t=1, the post intervention period). $POST_t$ takes the value 0 if t=0 and 1 if t=1, and T_j is the treatment indicator.

The error terms are structured as two cluster-level components (v_j and v_{jt}) and two individual-level components (ϵ_{ij} and ϵ_{ijt}), where v_j and ϵ_{ij} are time-invariant. Two autocorrelation terms are required in this case, namely the individual autocorrelation of the outcome over time, ρ_p, and the analagous cluster level term, ρ_c:

$$\rho_p = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{pt}^2} \quad \text{and} \quad \rho_c = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_{ct}^2},$$

where var(v_j)=σ_c^2, var(v_{jt})=σ_{ct}^2, var(ϵ_{ij})=σ_p^2 and var(ϵ_{ijt})=σ_{pt}^2[21]. The ICC in this sitation is expressed as[22]:

$$\rho = \frac{\sigma_c^2 + \sigma_{ct}^2}{\sigma_c^2 + \sigma_{ct}^2 + \sigma_p^2 + \sigma_{pt}^2}$$

Once these parameters are in hand, we can define the key paramater used in sample size calculations, r, the fraction of the total variance composed of by the time invariant components:

$$r = \frac{\sigma_c^2 + \sigma_p^2/m}{\sigma_c^2 + \sigma_{ct}^2 + \sigma_p^2/m + \sigma_{pt}^2/m} = \frac{m\rho}{1 + (m-1)\rho}\rho_c + \frac{1-\rho}{1 + (m-1)\rho}\rho_p$$

The sample size formula for a difference-in-differences estimation can be written as

$$n^* = m^* k^* = 2(1-r)(t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}(1 + (m-1)\rho) \tag{15}$$

and the sample size formula for an estimation using the baseline outcome variable as a covariate as:

$$n^* = m^* k^* = (1-r^2)(t_{\alpha/2} + t_\beta)^2 2\frac{\sigma^2}{\delta^2}(1 + (m-1)\rho) \tag{16}$$

In order to see the benefit of using the panel element, it is instructive to compare equations 15 and 16 with equation 6. The most important message is that the sample size requirement is minimized by including the baseline level of the outcome variable as a covariate (note that $1 - r^2 < 1$ and that $1 - r^2 < 2(1-r)$). Alternatively, given a sample, the highest power is achieved by including the baseline value of the outcome variable as covariate. Hence if baseline data on the outcome variable is available, one

[21] We note the abuse of notation in using t subscripts for the variance terms σ_{ct}^2 and σ_{pt}^2, as these terms are constant across the two time periods.

[22] Appendix C.5 details how to estimate these key panel data parameters.

should always control for it as a covariate rather than doing difference-in-differences or a simple post-treatment comparison (McKenzie, 2012; Teerenstra et al., 2012).

Also, it is useful to see that the largest reduction on sample size requirements when we include the baseline value as covariate takes place when r is close to 1 (hence $1-r^2$ which multiplies the sample size formulae 16 is close to zero). Intuitively, by conditoning on the baseline value of the outcome variable, we are netting out the time invariant component of the variance (which is large when r is close to 1).

Note also that if r is close to zero, given a sample, there might be little difference in power between including the baseline value of the outcome as a covariate, and just post-treatment differences. Hence, from the point of view of power, it might be better to spend the resources devoted to collect the baseline on collecting a larger sample post-treatment or several post-treatment waves (see McKenzie (2012))[23]. Interestingly, in terms of power, including the baseline value of the variable as covariate always dominates over differences-in-differences. Moreover, baseline data is required for both estimators. Hence, there is little reason in terms of power to justify difference-in-differences

In Table 3, we report the sample size requirements for the three estimation strategies for various values of r, calibrating the calculations to the likely effect size and variance of the earnings for 2002 in the APRENDE data. The resulting sample sizes quantifies the intuition above - the higher the time invariant component of the variance, r, the greater the benefit of controlling for baseline differences via covariate or difference-in-differences vis. a vis. single post-treatment difference. For low values of r, a difference-in-differences strategy is highly inefficient. The table also makes clear the dominance over the other two strategies of controlling for the baseline outcome as a covariate, for all values of r.

5 Binary Outcome Case

5.1 Non-clustered

Next, we move on to discussing the case where the outcome variable is binary, for instance whether an individual is working or not or whether a student obtained a certain grade level or not. There is a large literature that focusses on the binary outcome case, with several different approaches (for example Demidenko (2007), Moerbeek and Maas (2005)). Some papers deal with effect sizes measured in differences in log odds, others with differences in probability of success between treatment and controls. We follow Schochet (2013) who measures the effect size in terms of differences in the probability of success. We believe that this is more intuitive for most economists, and that the required inputs might be more easily accessible from published studies[24]. One difference between the continuous and the binary outcome case is that in the latter, we do not need the

[23]There might be other reasons to collect baseline data than gains in power. These vary from checking whether the sample is balanced in the outcome variables, to collect information that allow to stratify the sample, and to have the basis for heterogeneity analysis (see McKenzie (2012)).

[24]An advantage of the approach we follow is that the impact parameter does not depend on whether covariates are included or not. This is not the case when impact is measured in log odds. See Schochet (2013) for a detailed discussion of this.

variance. Binary outcomes follow a Bernoulli distributon, so knowing p, the probability of success, also yields the variance; $p(1-p)$.

Using a logistic model, we can write the probability of success for individual i as:

$$p_i = \text{Prob}(y_i = 1|T_i) = \frac{e^{\beta_0+\beta_1 T_i}}{1+e^{\beta_0+\beta_1 T_i}},$$

where y_i is binary (takes value 1 in case of success and 0 in case of failure) and as before, T_i denotes treatment status. The effect size, δ can thus be written as $p(y_i = 1|T_i = 1) - p(y_i = 1|T_i = 0)$ or $(p_1 - p_0)$, where the subscripts denote treatment and control status respectively.

Following an analogous procedure as in the continuous case, we arrive at a sample size equation for the binary case (Donner and Klar, 2010):

$$N^* = \left(\frac{p_1(1-p_1)}{\pi} + \frac{p_0(1-p_0)}{1-\pi}\right)\frac{(z_\beta + z_{\alpha/2})^2}{(p_1-p_0)^2}, \tag{17}$$

where π is the proportion of the sample that is treated[25], $n_1^* = \pi N^*$ and $n_0^* = (1-\pi)N^*$. Note that equation 17 is equivalent to equation 3, where σ_0^2 and σ_1^2 are replaced with their equivalents in the binary case, $p_0(1-p_0)$ and $p_1(1-p_1)$. In general, these variances will be different, so as we saw in equation 3, the optimal treatment-control split will differ from .5. The optimal allocation to treatment status, π^* can be written as:

$$\pi^* = \frac{\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}}{1+\sqrt{\frac{p_1(1-p_1)}{p_0(1-p_0)}}} \tag{18}$$

Hence, in the binary outcome case, the optimal split would only equal .5 in the special case where p_0=$1-p_1$ for example p_0=.4 and p_1=.6. In that case of an even split between treatment and control status ($\pi = .5$), we can write n^* as

$$n^* = (p_1(1-p_1) + p_0(1-p_0))\frac{(z_\beta + z_{\alpha/2})^2}{\delta^2} \tag{19}$$

5.2 Clustered

Having considered the individual level treatment case, we now move to cluster randomised treatment, still following Schochet (2013), as we do for the rest of section 5. For the cluster randomised case, a Generalised Estimating Equation (GEE) approach is followed, where the clustering is accounted for in the variance-covariance matrix, using the ICC, ρ. As before we can write the probability of success for individual i in cluster j as

$$p_{ij} = \text{Prob}(y_{ij} = 1|T_j) = \frac{e^{\beta_0+\beta_1 T_j}}{1+e^{\beta_0+\beta_1 T_j}},$$

[25] If the null hypothesis of zero impact is tested using a Pearson's chi-square test, and $n_1^* = n_0^*$, then $n^* = \frac{(z_{\alpha/2}\sqrt{2\bar{p}(1-\bar{p})}+z_\beta\sqrt{p_1(1-p_1)+p_0(1-p_0)})^2}{(p_0-p_1)^2}$ where $\bar{p} = \frac{p_1+p_0}{2}$, (see Fleiss et al. (2003) equation 4.14, as well as equation 4.19 for different sample sizes in treatment and control).

For cluster j, the $m \times m$ variance covariance matrix V_j is written as

$$V_j = A_j^{1/2} R(\rho) A_j^{1/2}, \tag{20}$$

where A_j is a diagonal matrix with diagonal elements $p_{ij}(1-p_{ij})$ and $R(\rho)$ is a correlation matrix with diagonal elements taking the value of 1, and off-diagonals the value of ρ. Hence $cov(y_{ij}, y_{km}) = \rho$ when $j = m$ and $=0$ when $j \neq m$. Note the lack of a j subscript on $R(\rho)$ - it is taken as common across clusters, as in the Generalised Least Squares (GLS) approach for a continuous outcome. This means that we no longer specify a random effect for each cluster, and allows us to get closed form solutions for the sample size equations[26].

The sample size equation for the binary outcome case with cluster randomisation can be written as:

$$N^* = \left(\frac{p_1(1-p_1)}{\pi} + \frac{p_0(1-p_0)}{1-\pi} \right) \frac{(z_{\alpha/2} + z_\beta)^2}{\delta^2} (1 + (m-1)\rho), \tag{21}$$

where π is the fraction of clusters randomised to receive treatment, $n_1^* = mk_1^* = \pi N^*$ and $n_0^* = mk_0^* = (1-\pi)N^*$. As above, if the treatment is evenly allocated, we can write this as

$$n^* = mk^* = (p_1(1-p_1) + p_0(1-p_0)) \frac{(z_\beta + z_{\alpha/2})^2}{\delta^2} (1 + (m-1)\rho) \tag{22}$$

As before, the sample size equation for the binary outcome mirrors that of the continuous outcome, with the design effect being the only difference between the individual and cluster randomised sample size equations.

Table 4 presents sample size requirements for three different levels of success probability for the control groups, p_0; 0.1, 0.3 and 0.5. The first thing to notice is that the closer p_0 is to .5, the larger the is sample size required. This is because for a binary variable, variance is largest at p=0.5. For example, for m=30 and ρ=.03, the sample size for p_0=0.5 is double that of p_0=.1. As in the continuous case, we see that higher ICCs and larger cluster sizes, m, lead to larger required total samples. This is due to the design effect.

5.2.1 Unequal numbers of clusters

It might be useful to have a formula for k_1 as a function of m and k_0, that will provide power of $(1-\beta)$ for the given m and k_0:

$$k_1 = \frac{\frac{p_1(1-p_1)}{m}(z_{\alpha/2} + z_\beta)^2(1 + (m-1)\rho)}{\delta - \left(\frac{p_0(1-p_0)}{mk_0}\right)(z_{\alpha/2} + z_\beta)^2(1 + (m-1)\rho)} \tag{23}$$

[26]Results from simulations we ran utilising the GEE approach yielded very similar results to those using a linear probabilty model with random effects. Schochet (2009) finds similar results using GEE and random effects logit models too.

5.3 The Role of Covariates

In this section, we consider the case of individual treatment allocation where one has a single covariate, X_i, that is discrete, but not neccessarily binary. In the case where the X_i is continuous, one can discretise the variable. Here, we write p_i as

$$p_i = \text{Prob}(yi = 1|T_i, X_i) = \frac{e^{\beta_0+\beta_1 T_i+\beta_2 X_i}}{1 + e^{\beta_0+\beta_1 T_i+\beta_2 X_i}}$$

Where covariates are included, we need several extra inputs into the sample size equation, relating to the distribution of the covariates, and how success probabilities change according to the covariate values.

First, assume that X_i can take any of the following Q values, $\{x_1, ..., x_Q\}$. Define $\theta_q = Prob(X_i = x_q)$ for $q \in \{1, ..., Q\}$, with $(0 < \theta_q < 1)$ and $\sum_q \theta_q = 1$. Next we need to specify how success probabilities change across the values of X_i. Define $p_{0q} = Prob(Y_i = 1|T_i = 0, X_i = x_q)$ and $p_{1q} = Prob(Y_i = 1|T_i = 1, X_i = x_q)$. Then we can define an effect size for a specific value of q, $\delta_q = p_{1q} - p_{0q}$, and an overall effect size, $\delta = \sum_q \theta_q \delta_q$. Schochet (2013) notes that covariate inclusion will improve efficiency if at least two of the p_{0q} or p_{1q} probabilities differ across covariate values.

With these inputs at hand, we can now write the sample size equation as:

$$N^* = (\mathbf{g}\mathbf{M}^{-1}\mathbf{g}')\frac{(z_\beta + z_{\alpha/2})^2}{\delta^2}, \tag{24}$$

where

$$\mathbf{M} = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_2 & m_2 & m_4 \\ m_3 & m_4 & m_5 \end{bmatrix},$$

$$\begin{aligned}
m_1 &= \sum_q \{\pi\theta_q p_{1q}(1 - p_{1q}) + (1 - \pi)\theta_q p_{0q}(1 - p_{0q})\} \\
m_2 &= \sum_q \{\pi\theta_q p_{1q}(1 - p_{1q})\} \\
m_3 &= \sum_q x_q\{\pi\theta_q p_{1q}(1 - p_{1q}) + (1 - \pi)\theta_q p_{0q}(1 - p_{0q})\} \\
m_4 &= \sum_q x_q\{\pi\theta_q p_{1q}(1 - p_{1q})\} \\
m_5 &= \sum_q x_q^2\{\pi\theta_q p_{1q}(1 - p_{1q}) + (1 - \pi)\theta_q p_{0q}(1 - p_{0q})\}
\end{aligned}$$

and $\mathbf{g}$ is a 1×3 gradient vector with elements:

$$\mathbf{g}[1,1] = \sum_q \theta_q [p_{1q}(1-p_{1q}) - p_{0q}(1-p_{0q})]$$
$$\mathbf{g}[1,2] = \sum_q \theta_q [p_{1q}(1-p_{1q})]$$
$$\mathbf{g}[1,3] = \sum_q x_q \theta_q [p_{1q}(1-p_{1q}) - p_{0q}(1-p_{0q})]$$

In the Appendix section D we provide a purposefully designed STATA programme to carry out this computation for 5 different values of covariates[27].

5.3.1 Clustered

Finally we consider a cluster randomised treatment in the presence of a single, discrete cluster-level covariate. Candidates for this could be a discrete cluster caracterisitc or a continuous variable, such cluster means of the outcome variable at baseline, which are then discretised. We write the probability of success here as

$$p_{ij} = \text{Prob}(y_{ij} = 1 | T_j, X_j) = \frac{e^{\beta_0 + \beta_1 T_j + \beta_2 X_j}}{1 + e^{\beta_0 + \beta_1 T_j + \beta_2 X_j}},$$

The variance-covariance matrix in this scenario is very similar to that without a covariate (see equation 20) with the exception of the use of the conditional ICC, ρ_x, not the raw ICC (ρ) in the correlation matrix. The sample size calculation for this section can be expressed as

$$N^* = 2m^*k^* = (\mathbf{g}\mathbf{M}^{-1}\mathbf{g}') \frac{(z_\beta + z_{\alpha/2})^2}{\delta^2} (1 + (m-1)\rho_x), \tag{25}$$

where $\mathbf{g}$ and $\mathbf{M}$ are defined as above, and ρ_x is the conditional ICC, as we saw in the continuous outcome case with cluster-randomisation and covariates. Note that the inclusion of a cluster-level covariate can lead to precision gains through decreasing the total residual variance, as well as by decreasing the conditional ICC. Schochet (2013) suggests that the latter will have more impact on lowering the required sample size.

Table 5 presents the number of clusters required in the binary outcome case for two values of ρ; 0.05 and 0.1, and a binary covariate. What we see here is that the greater is the difference between p_{00} and p_{01} (the difference in control group success rates for the two values of the covariate), the greater is the sample size reduction due to the inclusion of the covariate. The number of clusters required (for m=60, ρ=.05, p_0=.5, and a constant effect size of .1 across covariate levels) is 51 in the absence of a covariate (bottom right section of Table 4). In Table 5, this number falls to 49 when p_{00}=.4 and p_{01}=.6, and falls markedly to 32 when p_{00}=.2 and p_{01}=.8. In this example, the spread in impact values across the values of X has a more noticeable effect when the spread in control group success rates across X values is larger.

[27] As supplementary material we supply STATA programmes for 2, 3, 4 and 5 possible values of the covariate. Schochet (2013) provides a set of SAS programmes for sample size calculations for binary outcomes.

6 Extensions

6.1 Optimal sample allocation under cost constraints

In the previous sections, we have described how to compute the sample size that will allow us to reach a desired level of power. However, we did not take into account any cost considerations, which are usually very relevant for researchers. Researchers have a number of degrees of freedom when designing the sample of a study: ratio of the sample size in the treatment versus control group, number of clusters vs. number of individuals within a cluster, etc. Hence, it makes sense to choose the options that minimize costs without decreasing power[28].

In our previous examples, we have assumed the number of treatment and control units (clusters and/or individuals) to be the same. This is a common strategy used by researchers because it maximizes the power of the study given a total sample (this is the case for continuous outcomes. As discussed in section 5.1, for the binary outcome case, an even treatment-control split is unlikely to optimal in general). If the costs of the study depend solely on the total sample, then this approach will also maximize power given a total cost[29]. Hence, the approach will also minimize costs for a given level of power.

However, it is easy to think of cases where the costs do not depend solely on the total sample, but also of other parameters. For instance, if the research budget includes the cost of running the intervention then treatment units will be more expensive than the control ones. In these situations, we can maximize power (given an overall cost) by allocating a larger sample to the control group than the equal split. To understand the intuition, starting from the equal split, allocating units from treatment to control will result in a overall cost reduction but also loss of power because of the imbalance. However, the latter could be more than offset if we use the part of the cost reduction to boost total sample size. But of course, this will have a limit. As the loss of power increases more than linearly with the imbalance, the loss of power might not be offset if the sample is already highly unbalanced.

To allow the user to operationalize the above, below we provide the formulae that allows us to compute C^*, the minimum cost that allow us to detect a given MDE, δ. This will be of use for the researcher that has a clear view on the MDE of the intervention that she is testing, and wants to find out the minimum cost that will allow her to detect it in order to submit a competitive bid. We also provide formulae for δ^* which is the feasible MDE given a total cost C. This will be of more use to the researcher that has a binding cost constraint and is figuring out how effective her intervention must be in order to detect its effect under the cost constraint. The specific formulae depends on

[28]Here, we do not study how to minimize costs as function of the autocorrelation of the outcome variable. For low autocorrelated outcomes, costs might be minimized subject to a given level of power by not having baseline, but multiple post-treatment measures (McKenzie, 2012).

[29]This implicitly assumes that the variance of the outcome variable in treatment and control groups is the same. In situations when this is not true, power is maximized if the group with the higher variance is larger (for example see equation 3).

the specific structure of the cost function. We specify a different one on each of the subsections below[30].

6.1.1 Individually allocated treatment

There are many different variations of cost functions we may specify. Here we present a few alternatives. First consider the situation where treatment allocation is individually allocated and the cost function is given by

$$C = c_0 n_0 + c_1 n_1$$

Once we have specified the cost function, we then solve a constrained optimisation problem, minimising the MDE subject to the cost function. This yields a solution in terms of n_0, n_1 and the cost function parameters[31]:

$$\frac{n_1}{n_0} = \sqrt{\frac{c_0}{c_1}}$$

Hence, the more expensive the treatment units are, the smaller the treatment group would be compared to the control. Using the cost function, we can write n_0 and n_1 as functions of the cost parameters:

$$n_0^* = \frac{C}{c_0 + \sqrt{c_0 c_1}} \quad \text{and} \quad n_1^* = \frac{C}{c_1 + \sqrt{c_0 c_1}} \tag{26}$$

The relations in equation 26 would be combined with equation 1 to obtain the smallest MDE achievable in order to obtain a power of $1 - \beta$ given the budget constraint C:

$$\delta^* = (t_\beta + t_{\alpha/2}) \sqrt{\frac{1}{n_0^*} + \frac{1}{n_1^*}}, \tag{27}$$

which leads to a formula for δ^* as a function of the cost parameters:

$$\delta^* = (t_\beta + t_{\alpha/2}) \frac{\sigma}{\sqrt{C}} \left(\sqrt{c_0} + \sqrt{c_1}\right) \tag{28}$$

This formulation is useful in order to assess whether or not to conduct a trial at a given budget, C. For instance, if the budget for the trial, C, were very small, this would limit the number of individuals in the trial, and would thus require a very large effect size in order to achieve a power of $1 - \beta$. If this effect size is unrealistic, the RCT is under-powered with the given budget. Alternatively we can derive an expression for the minimum total cost, C^*, required in order to achieve a power of $1 - \beta$ with a given value of δ. In order to do so, we use the relations in equations 26 and 27:

$$C^* = (t_\beta + t_{\alpha/2})^2 \frac{\sigma^2}{\delta^2} \left(\sqrt{c_0} + \sqrt{c_1}\right)^2 \tag{29}$$

[30] In what follows we provide formulae for optimal allocations for continuous outcomes. One can follow a similar approach to the one outlined below to derive the equivalent formulae for binary outcome cases.

[31] The equivalent formula for the discrete case is: $\frac{n_1}{n_0} = \sqrt{\frac{c_0}{c_1} \frac{p_1(1-p_1)}{p_0(1-p_0)}}$.

6.1.2 Cluster-level treatment allocation with heterogenous cluster costs

Moving on to the cluster randomisation case, we follow the same method as above, first specifying a cost function and then minimising the MDE subject to this cost function. For instance, if the treatment is the provision of a cluster-level service or amenity (a clean well or improved sanitation amenities at the village level), then the only difference between the costs of treatment and control areas will be the fixed cost of this service provision, yielding a cost function of the form:

$$C = f_0 k_0 + f_1 k_1,$$

where $f_0 = f_0' + vm$ and $f_1 = f_1' + vm$. Here, the cluster size is fixed at a certain m[32], thus we do not consider this dimension of the sample when optimising. Solving a constrained optimisation problem as before (where the objective function is the square of the MDE) gives us the following solution[33]:

$$\frac{k_1}{k_0} = \sqrt{\frac{f_0}{f_1}}$$

As before, we use the cost function to write the optimal values of k_1 and k_0 as functions of the cost parameters:

$$k_0^* = \frac{C}{f_0 + \sqrt{f_0 f_1}} \quad \text{and} \quad k_1^* = \frac{C}{f_1 + \sqrt{f_0 f_1}} \tag{30}$$

The remaining step in this constrained optimisation problem is to use equation 8 with equation 30 to compute the effect size as a function of the optimal values of k_0 and k_1, which will yield the minimum effect size that will need to be found in order to obtain a power of $1 - \beta$, given the budget constraint, C[34]:

$$\delta^* = (t_{\alpha/2} + t_\beta)\sqrt{\sigma^2 \left(\frac{1 + (m-1)\rho}{m k_0^*} + \frac{1 + (m-1)\rho}{m k_1^*} \right)}, \tag{31}$$

which can be written in terms of the cost function parameters as:

$$\delta^* = (t_{\alpha/2} + t_\beta)\sqrt{\sigma^2 (1 + (m-1)\rho) \left(\frac{f_0 + \sqrt{f_0 f_1}}{mC} + \frac{f_1 + \sqrt{f_0 f_1}}{mC} \right)} \tag{32}$$

[32] For example, where the cluster is a school, and the outcome variable is the result of a test taken by all pupils in the school, m.

[33] In this section, we ignore that the degrees of freedom of the t-distribution are a function of the number of clusters, k. This will normally have a minimal effect on the sample size, unless the number of clusters is very small.

[34] We provide the formulae for the case without covariates. One can replace σ^2 with σ_x^2 and ρ with ρ_x to adapt the formulae to the case with covariates.

We can also derive an expression for the minimum total cost, C^*, required in order to achieve a power of $1-\beta$ with a given value of δ. To do so, we combine equations 30 and 31 to arrive at:

$$C^* = (t_{\alpha/2} + t_\beta)^2 \frac{\sigma^2}{\delta^2} \frac{(1+(m-1)\rho)}{m} \left(\sqrt{f_0} + \sqrt{f_1}\right)^2 \tag{33}$$

6.1.3 Cluster-level treatment allocation with heterogenous individual costs

Were the treatment to have an individual-level component, such as a vaccination programme or a job training programme, then the variable costs may be of more relevance. This may give rise to a cost function such as:

$$C = 2fk + v_0 m_0 k + v_1 m_1 k,$$

where k is fixed and f represents the fixed cost of data collection at the cluster. The constrained optimisation problem yields the following ratio:

$$\frac{m_1}{m_0} = \sqrt{\frac{v_0}{v_1}},$$

which we can combine with the cost function in order to get the following expressions for optimal values of m_0 and m_1:

$$m_0^* = \frac{C-2fk}{k(v_0 + \sqrt{v_0 v_1})} \quad \text{and} \quad m_1^* = \frac{C-2fk}{k(v_1 + \sqrt{v_0 v_1})} \tag{34}$$

As we saw in section 6.1.2, the final step now is to use equation 10 with equation 34 to compute the effect size as a function of these optimal values of m_0 and m_1, which will yield the minimum detectable effect size required to achieve a power of $1-\beta$, given the budget constraint, C:

$$\delta^* = (t_{\alpha/2} + t_\beta)\sqrt{\sigma^2 \left(\frac{1+(m_0^*-1)\rho}{m_0^* k} + \frac{1+(m_1^*-1)\rho}{m_1^* k}\right)}, \tag{35}$$

which we can write in terms of the cost function parameters as:

$$\delta^* = (t_{\alpha/2} + t_\beta)\sqrt{\sigma^2 \left(\frac{2\rho}{k} + \frac{(1-\rho)(\sqrt{v_0}+\sqrt{v_1})^2}{C-2fk}\right)}, \tag{36}$$

Finally, we can derive an expression for the minimum total cost, C^*, required in order to achieve a power of $1-\beta$ with a given value of δ. To do so, we combine equations 35 and 34:

$$C^* = 2fk + \frac{(t_{\alpha/2} + t_\beta)^2 \sigma^2 (1-\rho)(\sqrt{v_0}+\sqrt{v_1})^2}{\delta^2 - (t_{\alpha/2} + t_\beta)^2 \sigma^2 (\frac{2\rho}{k})} \tag{37}$$

6.1.4 Cluster-level treatment allocation with homogenous individual and cluster costs

We now consider a cost function with homogenous individual and cluster costs. The aim here is to achieve an optimal allocation of the total sample into number of clusters, k, and cluster sizes,m. The cost function thus takes the form

$$C = k(f + vm),$$

where f is the fixed cost per cluster, and v the variable cost, per individual. Minimising the square of the MDE (as given in equation 5) subject to this cost constraint yields optimal values for m:

$$m^* = \sqrt{\frac{f}{v}\frac{\sigma_p^2}{\sigma_c^2}} \tag{38}$$

Using this formula and the form of the cost function we derive an expression for the optimal k:

$$k^* = \frac{C}{f + v\sqrt{\frac{f}{v}\frac{\sigma_p^2}{\sigma_c^2}}} \tag{39}$$

As Liu (2013) notes, it my be instructive to use the definition of the ICC to rewrite equations 38 and 39 as:

$$m^* = \sqrt{\frac{f}{v}\frac{1-\rho}{\rho}} \quad \text{and} \quad k^* = \frac{C}{f + v\sqrt{\frac{f}{v}\frac{1-\rho}{\rho}}} \tag{40}$$

Here we see that the larger is the ICC, the smaller the optimal m. This is due to the fact that when the ICC is high, outcomes within clusters are highly correlated, and increasing the number within the cluster, m, adds little in precision gains. Resources are better spent by increasing the number of clusters, k.

With the optimal values of m and k in hand, we can compute the minimum effect size in order to achieve power of $1 - \beta$, given the budget constraint, C:

$$\delta^* = (t_{\alpha/2} + t_\beta)\sqrt{2\sigma^2\left(\frac{1 + (m^* - 1)\rho}{m^* k^*}\right)} \tag{41}$$

To complete this section, we derive an expression for the minimum total cost, C^*, required in order to achieve a power of $1 - \beta$ with a given value of δ. In order to do so, we combine equations 41 and 40:

$$C^* = \frac{2\sigma^2}{\delta^2}(t_{\alpha/2} + t_\beta)^2\left[f + v\sqrt{\frac{f}{v}\frac{1-\rho}{\rho}}\right]\left(\frac{1 + (\sqrt{\frac{f}{v}\frac{1-\rho}{\rho}} - 1)\rho}{\sqrt{\frac{f}{v}\frac{1-\rho}{\rho}}}\right) \tag{42}$$

6.1.5 More complex cost functions

The researcher might face more complex design features that the ones discussed above, and the optimal allocation might require different number of treatment and control clusters, and/or different number of individuals sampled across treatment and control clusters. For these situations, it is useful to generalize equation 5 and express the MDE as:

$$\delta^2 = (t_{\alpha/2} + t_\beta)^2 \left(\frac{\sigma_c^2}{k_0} + \frac{\sigma_p^2}{m_0 k_0} + \frac{\sigma_c^2}{k_1} + \frac{\sigma_p^2}{m_1 k_1} \right).$$

In general, the optimal allocation of k_0,k_1,m_0, and m_1 will be obtained by minimizing $\left(\frac{\sigma_c^2}{k_0} + \frac{\sigma_p^2}{m_0 k_0} + \frac{\sigma_c^2}{k_1} + \frac{\sigma_p^2}{m_1 k_1} \right)$ subject to a cost constraint[35]. For specific parameter values, this minimisation can be done using numerical optimization software.

There are some simplified cases in which one can also combine some of the previous results with a grid search on one of the unknown parameters to find the optimal solution using a simple spreadsheet. For instance, consider the case in which treatment and control cluster fixed costs are different but m is not given, so an optimal m must be found. This would be as the case of subsection 6.1.2 but with unknown m. Hence, using that f_0=$f_0^{'}$+vm and that f_1=$f_1^{'}$+vm, different values of f_0 and f_1 can be computed for each possible m. These will lead to different values of k_0 and k_1 using equation 30. The optimal combination of m, k_0 and k_1 will be the one that minimizes $\left(\frac{\sigma_c^2}{k_0} + \frac{\sigma_p^2}{m k_0} + \frac{\sigma_c^2}{k_1} + \frac{\sigma_p^2}{m k_1} \right)$.

Another simplified case which is likely to be of interest is when the treatment and control variable cost per observation are different, but the fixed cost per cluster are the same. This is the case of section 6.1.3 but where k is not given, and an optimal value for it must be found. Again, a grid search for different values of k can be used to find the optimal combination of k,m_0,m_1. For each different value of k, the corresponding values of m_0 and m_1 can be computed using equations 34. The optimal combination of k, m_0 and m_1 will be the one that minimizes $\left(\frac{\sigma_c^2}{k} + \frac{\sigma_p^2}{m_0 k} + \frac{\sigma_c^2}{k} + \frac{\sigma_p^2}{m_1 k} \right)$.

In practice, the cost function might not be as simple as the one used above. For instance, costs could increase discontinuously in the number of individuals per cluster if interviewers must spend an extra night, or a new vehicle must be purchased to be able to cover more than a certain amount of clusters in a given time period. However, the exercise of computing the cost for each combination of clusters and number of individuals per cluster should be feasible. This information can be embedded within an isopower curve, which yields the different combinations of clusters (k) and individuals per cluster (m) that provide the same level of power[36]. The cost information, together with the isopower curve, will allow the researcher to choose the combination that minimizes cost. There may be cases where the data collection is commissioned to a survey firm that

[35] Note that minimizing the MDE is equivalent to maximizing power. Note also that the other components of the MDE formulae are fixed with the sample

[36] using formulae such as equations (8), (10), (13), (14) or (23) depending on the case

is not willing to share the cost function. In these circumstances, the researcher can provide possible combinations of number of clusters and number of individuals (from the isopower curve), and the survey firm can choose the one that minimizes its costs.

6.2 Simulation

A researcher might need to compute the required sample size for an experiment whose features do not conform to the ones indicated in previous sections. The possibilities of variation are endless. They include experiments in which the number individuals per cluster varies across clusters, experiments with more than two treatment arms, or using data from more than two time periods, to say a few. In situations where some features of the experimental design vary significantly with respect to the canonical cases given above, simulation methods can be very useful to estimate the power of a given design, and correspondingly adjust the sample of the design to achieve the desired level of power.

To understand the logic of the simulation approach, it is useful to remember the definition of power: the probability that the intervention is found to have an effect on outcomes when that effect is true. In a hypothetical scenario in which the researcher happened to have 1,000 samples as the ones of her study, and if she could be certain that "the effect is true" in all these samples, then she could estimate such probability (power) by simply counting in how many of these samples she "finds" the effect (the null hypothesis of zero effect is rejected), and dividing it by 1,000.

The simulation approach simply operationalizes the above by providing the researcher with 1,000 (or more) computer-generated samples, hopefully similar to the one of her study (or at least, obtained under the assumptions that that the researcher is planning the study). Because these are computer-generated samples, the researcher can obtain these samples imposing the constraint that the effect is true (and in particular, it will draw the samples assuming that the effect of the intervention is the same as the effect size, δ, for which she wants to estimate the power).

In general, the steps required to estimate the power of a given design through simulation are as follow (see Appendix C for an example)[37]:

Step 1: define the number of simulations that will be used to estimate the power of the design, say S; as well as the significance level for the tests.

Step 2: define a model that will be used to draw computer-generated samples "as those in the study". This model will have a non-stochastic part (sample size, number of clusters, distribution of the sample across clusters, number of time periods, ICC, autocorrelation terms, mean and standard deviation of the outcome variable, effect size, etc) and a stochastic part (error term)[38]. An example of such model could be, for instance, equation 4 but for specific values for the effect size, standard deviation and ICC (in Appendix C.6 these are set as $\delta = 4$, $\sigma = 10$ and $\rho = .3$).

[37] Feiveson (2002) provide insightful examples for Poisson regression, Cox regression, and the rank-sum test.

[38] If a pilot dataset is available, an alternative approach is to bootstrap from this data(see Kleinman and Huang (2014)).

Step 3: using computer routines for pseudo-random numbers, obtain a draw of the error term (or composite of error terms) for each individual in the sample. It is crucial that the error term is drawn taking into account the stochastic structure of our experiment (the correlation of draws amongst different individuals and time periods through the ICC or similar parameters). To draw samples from the error terms, a distribution will need to be assumed. Although assuming Normality is common, the approach allows to assume other distributions that might be more appropriate for the specific experiment.

Step 4: using the model and parameter values indicated in Step 2, and the sample of the error term (or composite of error terms) generated in Step 3, obtain the values of the outcome variable for the sample. Once this is done, the draws of the error term generated in Step 3 can be discarded.

Step 5: using the data on outcomes generated in Step 4, and the model of Step 2, test the null hypothesis of interest (usually, that the intervention has no effect[39]). Keep a record of whether the null hypothesis has been rejected or not.

Step 6: Repeat Steps 3 to 5 for S times

Step 7: the estimated power is the number of times that the null hypothesis was rejected in Step 5 divided by S.

Although using simulation methods to estimate power has a long tradition in statistics, the approach is not so commonly used in practice (Arnold et al. 2011)[40]. We suspect that Step 3 is the most challenging for the applied researchers. In Appendix C, we provide several hints, which could be of some help.

6.3 Adjusting sample size calculations for multiplicity

A common problem with experiments (and more generally in empirical work) is that, more than one null hypothesis is usually tested. For instance, it is common to test the effect of the intervention on more than one outcome variable. This creates a problem because the number of rejected null hypothesis (the number outcome variables for which an effect is found) will increase (independently of whether they are true or not) with the number of null hypotheses (outcome variables) tested if the significance level is kept fixed with the number of hypotheses.

For instance, consider that we are testing the effect of an intervention on three different outcome variables, and that we use an α equal to 0.05 for each test. If we assume that the three outcome variables are independent, then probability that we do not reject any of the three when the three null hypotheses are all true is $(1-0.05)^3$. Hence, the probability that we reject at least one of them if the three are true is $1-(1-0.05)^3 = 0.14$. Why is this a problem? Assume that the intervention will be declared successful it is found that it improves at least one of the outcomes. The numbers above implies that the intervention will be declared successful with a probability of 0.14 (larger than the

[39] We are assuming that the test for the null hypothesis has the correct size. Otherwise see Lloyd (2005)

[40] See Hooper (2013), Kontopantelis et al. (Forthcoming), and Kumagai et al. (2014) for some recent implementations of the simulation approach to estimate power.

normal significance level of 0.05) even if it has no real effect on any of the three outcome variables.

The problem of multiplicity of outcome variables is recognized by regulatory agencies that approve medicines (Food and Drug Administration (1998) and European Medicines Agency (2002)) and has recently become more common also in applied work in economics (Anderson (2008), Carneiro and Ginja (2014)[41]). The standard solution requires performing each individual hypothesis test under an α smaller than the usual 0.05 (Ludbrook 1998, Romano and Wolf 2005) so that the probability that at least one null hypotheses is rejected when all null hypotheses are true ends up being 0.05[42]. Hence, when doing the sample size calculations, the researcher should also use a smaller α than 0.05, which will increase the sample size requirements.

When the outcome variables are independent, the probability that at least one null hypothesis is rejected when all are true, usually called the *Family Wise Error Rate* (FWER) is $1-(1-\alpha)^h$, where α is the level of significance of the individual tests and h is the number of null hypothesis that are tested (i.e. number of outcome variables). Hence, if our study needs a FWER =0.05, then the significance level for each individual test is given by $1-(1-0.05)^{(1/h)}$, which would be 0.0169 in our example of $h=3$[43].

In most experiments, the outcome variables will not be independent. Taking into account this dependency will yield higher values of α, and consequently smaller sample size requirements. If one was willing to assume the degree of dependency amongst the different outcome variables, then a time consuming but feasible approach to compute the required power is to use the simulation methods previously described combined with a method for Step 5 (testing the null hypothesis) that takes into account the multiple tests carried out and the dependence in the data (such as Romano and Wolf (2005) or Westfall and Young (1993)). If this was not available, a rule of thumb is to use $\alpha=1-(1-0.05)^{(1/\sqrt{h})}$, a correction which was popularised by John W. Tukey(Braun, 1994). This will result in an α larger than when independence is assumed, and hence smaller sample size requirements.

7 Conclusion

In this paper, we have reviewed the methods to provide sample size calculations for studies that go beyond the simple comparisons of treatment and control averages. Extensions have included how to maximize power given a cost constraint, adjusting the sample size calculations when multiple outcomes are being tested, and the use of simulation exercises to estimate the power of more complex designs not covered so far.

Researchers will need to make more assumptions when taking advantage of these more complex methods that we provide here than in simpler ones. However, we believe

[41] There is less consensus on whether correcting for multiplicity is necessary when testing multiple treatments (see Wason et al. (2014)).

[42] An alternative way to analyse the data is to test jointly (through an F-test) the null hypotheses that the intervention does not have an impact on any of the outcome variables considered.

[43] A common simplification is to use the Bonferroni correction, which would be $0.05/h$

that the increasing availability of publicly available datasets means that researchers are in a relatively good position to make credible assumptions about the required parameters. Researchers could also help other researchers by reporting basic parameters such as intra cluster correlations, individual and cluster level correlations, and R-squares in their papers. Journal editors could help to speed up the process by coordinating on certain reporting, rules as it has happened in Medicine (Schulz et al., 2010).

Our experience is that researchers more often than not, tend to go for an equal split of sample size between treatment and control, before thinking how an unequal split could decrease the costs (or provide more power at the same cost). The formulae and other information that we have provided here might enable researchers to improve on this practice.

Another issue in which we expect further development in the future is to consider adjustments in the sample size calculations when the RCT considers multiple outcomes. It is starting to make its way on empirical work to adjust the p-values when multiple hypotheses are being tested. Hence, it is only a question of time that this adjustment is not only done ex-post once the data is collected, but considered ex-ante when the study is being designed and the sample planned.

References

Anderson, M. L. (2008): "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 103, 1481–1495.

Bloom, H. S., L. Richburg-Hayes, and A. R. Black (2007): "Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions," *Educational Evaluation and Policy Analysis*, 29, 30–59.

Blundell, R. and M. Costa Dias (2009): "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 44, 565–640.

Braun, H. I. e. (1994): *The collected works of John W Tukey Vol. VIII. Multiple comparisons : 1948-1983*, New York: Chapman & Hall.

Burtless, G. (1995): "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 63–84.

Carneiro, P. and R. Ginja (2014): "Long-Term Impacts of Compensatory Preschool on Health and Behavior: Evidence from Head Start," *American Economic Journal: Economic Policy*, 6, 135–73.

Demidenko, E. (2007): "Sample size determination for logistic regression revisited," *Statistics in Medicine*, 26, 3385–3397.

Donner, A. and N. Klar (2010): *Design and Analysis of Cluster Randomization Trials in Health Research*, Chichester: Wiley, 1st ed.

Duflo, E. (2006): "Field Experiments in Development Economics," Tech. rep., MIT.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Chapter 61 Using Randomization in Development Economics Research: A Toolkit," Elsevier, vol. 4 of *Handbook of Development Economics*, 3895 – 3962.

European Medicines Agency (2002): "Points to Consider on Multiplicity Issues in Clinical Trials," Tech. rep.

Feiveson, A. (2002): "Power by simulation," *Stata Journal*, 2, 107–124.

Fleiss, J. L., B. Levin, M. C. Paik, and J. Fleiss (2003): *Statistical Methods for Rates & Proportions*, Hoboken, NJ: Wiley-Interscience, 3rd ed.

Food and Drug Administration (1998): "Statistical Principles for Clinical Trials. E9." Tech. rep.

Hausman, J. and D. Wise, eds. (1985): *Social Experimentation*, University of Chicago Press.

Heckman, J. J. and J. A. Smith (1995): "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 9, 85–110.

Hedges, L. and C. Rhoads (2010): "Statistical Power Analysis in Education Research," Tech. rep., The Institute of Education Sciences.

Hooper, R. (2013): "Versatile sample-size calculation using simulation," *The Stata Journal*, 13(1), 21–38.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Kleinman, K. and S. S. Huang (2014): "Calculating power by bootstrap, with an application to cluster-randomized trials," *ArXiv e-prints*.

Kontopantelis, E., D. Springate, R. Parisi, and D. Reeves (Forthcoming): "Simulation-based power calculations for mixed effects modelling: ipdpower in Stata," *Journal of Statistical Software*.

Kumagai, N., K. Akazawa, H. Kataoka, Y. Hatakeyama, and Y. Okuhara (2014): "Simulation Program to Determine Sample Size and Power for a Multiple Logistic Regression Model with Unspecified Covariate Distributions." *Health*, 6, 2973–2998.

Levitt, S. D. and J. A. List (2009): "Field experiments in economics: The past, the present, and the future," *European Economic Review*, 53, 1 – 18.

List, J., S. Sadoff, and M. Wagner (2011): "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design," *Experimental Economics*, 14, 439–457.

Liu, X. S. (2013): *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques*, Routledge.

Lloyd, C. J. (2005): "Estimating test power adjusted for size," *Journal of Statistical Computation and Simulation*, 75, 921–933.

McKenzie, D. (2012): "Beyond baseline and follow-up: The case for more T in experiments," *Journal of Development Economics*, 99, 210 – 221.

Moerbeek, M. and C. J. M. Maas (2005): "Optimal Experimental Designs for Multilevel Logistic Models with Two Binary Predictors," *Communications in Statistics - Theory and Methods*, 34, 1151–1167.

Romano, J. P. and M. Wolf (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73, 1237–1282.

Schochet, P. Z. (2009): "Technical Methods Report: The Estimation of Average Treatment Effects for Clustered RCTs of Education Interventions," Tech. rep., The Institute of Education Sciences.

——— (2013): "Statistical Power for School-Based RCTs With Binary Outcomes," *Journal of Research on Educational Effectiveness*, 6, 263–294.

Schulz, K. F., D. G. Altman, and D. Moher (2010): "CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials," *BMJ*, 340.

Teerenstra, S., S. Eldridge, M. Graff, E. de Hoop, and G. F. Borm (2012): "A simple sample size formula for analysis of covariance in cluster randomized trials," *Statistics in Medicine*, 31, 2169–2178.

Wason, J. M., L. Stecher, and A. Mander (2014): "Correcting for multiple-testing in multi-arm trials: is it necessary and is it done?" *Trials*, 15, 364.

Westfall, P. and S. Young (1993): *Resampling-Based Multiple Testing?: Examples and Methods for P-Value Adjustment*, Wiley.

Appendices

A Derivation of sample size formula for individual-randomized case with unequal variances

In this section we derive the optimal sample allocations for the individual-randomized case where variances are unequal, arriving at the expression found in equation 3. We start with the expression for δ:

$$\delta = (t_\beta + t_{\alpha/2})\sqrt{\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}}$$

$$\delta^2 = (t_\beta + t_{\alpha/2})^2 \left(\frac{\sigma_0^2}{N - n_1} + \frac{\sigma_1^2}{n_1}\right)$$

$$\frac{\partial \delta^2}{\partial n_1} = (t_\beta + t_{\alpha/2})^2 \left((-1)(-1)\frac{\sigma_0^2}{(N-n_1)^2} + (-1)\frac{\sigma_1^2}{(n_1)^2}\right) = 0$$

$$\Rightarrow \quad \frac{\sigma_0^2}{(N-n_1)^2} = \frac{\sigma_1^2}{(n_1)^2}$$

$$\Rightarrow \quad \frac{\sigma_0}{(N-n_1)} = \frac{\sigma_1}{(n_1)}$$

$$\Rightarrow \quad n_1\sigma_0 = (N-n_1)\sigma_1$$

$$\Rightarrow \quad N = \frac{n_1\sigma_0}{\sigma_1} + n_1 = \frac{n_1\sigma_0 + n_1\sigma_1}{\sigma_1}$$

$$\Rightarrow \quad N = \frac{n_1\sigma_0 + n_1\sigma_1}{\sigma_1} = \frac{\sigma_0 + \sigma_1}{\sigma_1} n_1$$

Defining $\pi_1 = \frac{\sigma_1}{\sigma_0+\sigma_1}$ leads to $N = \frac{n_1}{\pi_1}$. By symmetry, the F.O.C. for n_0 will result in $N = \frac{n_0}{\pi_0}$, where $\pi_0 = \frac{\sigma_0}{\sigma_0+\sigma_1}$. Plugging these values for n_0 and n_1 back into the equation for δ^2 we see that:

$$N^* = (t_\beta + t_{\alpha/2})^2 \frac{1}{\delta^2}\left(\frac{\sigma_0^2}{\pi_0^*} + \frac{\sigma_1^2}{\pi_1^*}\right),$$

where $\pi_0^* = \frac{\sigma_0}{\sigma_0+\sigma_1}$ and $\pi_1^* = \frac{\sigma_1}{\sigma_0+\sigma_1}$

B Estimating Key Parameters in STATA

On ocassions, we might have a dataset from a similar environment to the one for which we are planning the RCT. In this case, we might use this dataset to estimate the key parameters needed for the sample size calculations. In this section, we show how to do this with the APRENDE dataset. The variables earnings_02 and earnings_01 denote earnings in two different time periods. The variable read is used as a covariate, and the varaible town_id is the cluster identifier.

```
use "APRENDE.dta", clear
drop if read==.
/ * estimate ρ*/
loneway earnings_02 town_id
gen rho=r(rho)
/* estimate σ_p*/
gen sigma_p=r(sd_w)
/* estimate σ_c*/
gen sigma_c=r(sd_b)
/* estimate σ_x, using the varable read as the X variable*/
regr earnings_02 read, cluster(town_id)
predict uhat,resid
sum uhat
gen sigma_x=r(sd)
/ * estimate ρ_x*/
loneway uhat town_id
gen rho_x=r(rho)
/* estimate σ_xp*/
gen sigma_xp=r(sd_w)
/* estimate σ_xc*/
gen sigma_xc=r(sd_b)
/ * estimate R_p^2*/
gen R2_i=(sigma_p^2-sigma_xp^2)/sigma_p^2
/ * estimate R_c^2*/
gen R2_c=(sigma_c^2-sigma_xc^2)/sigma_c^2
/* estimate the panel data parameters ρ_p and ρ_c */
egen earnings_02_c=mean(earnings_02),by(town_id)
egen earnings_01_c=mean(earnings_01),by(town_id)
gen earnings_02_p=earnings_02-earnings_02_c
gen earnings_01_p=earnings_01-earnings_01_c
corr earnings_02_c earnings_01_c
gen rho_c=r(rho)
corr earnings_02_p earnings_01_p
gen rho_p=r(rho)
```

C Simulation Code

The objective here is to provide sections of code that are useful if you would like to run your own simulations. This may be done in order to verify formulae for certain outcomes, or to provide a simulated estimate of sample size if no formulae are available for the specific, likely complex, trial design to be implemented.

C.1 How to create clusters

The code below creates a dataset with 100 clusters ($k = 100$) with 10 observations per cluster ($m = 10$), with equal treatment/control allocation. At the cluster level, a normally distributed cluster-level error term, with a standard deviation of 10, is created (called group below). As shown later, this will be used to create an ICC.

```
/*CLUSTER LEVEL*/
/* create an empty dataset with 100 observations*/
set obs 100
gen cluster=_n
/*draw from a normal distribution, with mean 0, standard deviation
10 */
gen group=rnormal(0,10)
sum
local N=r(N)
/*create the treatment variable indicator*/
gen treat=0
/* allocate half of the cluster to treatment status, the remaining
half to control*/
replace treat=1 if _n<=`N'/2
so cluster
tempfile cluster_error_g
/*create a temporary file for the cluster errors */
save `cluster_error_g',replace
/*INDIVIDUAL LEVEL*/
clear
/*n=mk, so if we require k=100 and m=10, we need n=1000*/
set obs 1000
/*Generate clusters*/
gen u=invnormal(uniform())
/*cut the data into 100 equally sized sections*/
egen cluster = cut(u), g(100)
replace cluster=cluster+1
so cluster
*merge in cluster errors
merge cluster using `cluster_error_g'
```

C.2 How to generate data with a specific ICC

In this section, all of the initial code is the same as before. The addition comes in the bottom lines, when we create the outcome variable, y, as a mix of individual- and cluster-level errors. The code below sets the ICC=.3 . Note too that we set $\sigma = 10$, the average of y in the control group to 10, and $\delta = 4$.

```
/*CLUSTER LEVEL*/
set obs 100
gen cluster=_n
/* set σ = 10 */
gen group=rnormal(0,10)
sum
local N=r(N)
gen treat=0
replace treat=1 if _n<=`N'/2
so cluster
tempfile cluster_error_g
save `cluster_error_g',replace
/*INDIVIDUAL LEVEL*/ clear
set obs 1000
gen u=invnormal(uniform())
egen cluster = cut(u), g(100)
replace cluster=cluster+1
so cluster
merge cluster using `cluster_error_g'
tab _m
drop _m
/* set σ = 10 */
gen individual=rnormal(0,10)
/*create error term as composite of group and individual error terms,
with weights to achieve an ICC=.3*/
gen epsilon = (sqrt(.3))*group + (sqrt(.7))*individual
/* set y=10 for the control group */
gen y=10+ epsilon if treat==0
/* generate a δ = 4 */
replace y=10+4+epsilon if treat==1
```

Once the data has been created, we may want to estimate the ICC. This is done using STATA's *loneway* command. Below we consider the ICC for the control clusters:

```
loneway y cluster if treat==0
```

C.3 How to Introduce Covariates

To add a covariate at the cluster level, we follow the code as in the section above with two additions. In the cluster level code we add:

```
gen x_g=rnormal(0,10)
```

Then, at the bottom of the code, when specifying the data generating process for y, we decide on the R^2 of this covariate. Note that both cluster- and individual-level error terms both have $\sigma = 10$, as does the covariate. The code below generates an R^2 of .2, and a conditional ICC of .3 as well (the conditional variance is .8 of total variance, so to get a conditional ICC of .3, we weight the group component by .3*.8=.24):

```
gen y=10+(sqrt(.24))*group + (sqrt(.2))*x_g + (sqrt(.56))*individual
```

C.4 How to generate data with binary outcomes and specific ICCs

In order to get clustered binary outcomes, we specify a beta-binomial distribution as the data generating process at the cluster level. This means cluster success rates,p_j are draws from a distribution with mean p and variance $\rho p(1-p)$. The beta-binomial distribution has two parameters, α and β, which we can derive using the two expressions $p = \frac{\alpha}{\alpha+\beta}$ and $\rho = \frac{1}{1+\alpha+\beta}$. Rearranging, we get $\alpha = \frac{p(1-\rho)}{\rho}$ and $\beta = \frac{p(1-p)(1-\rho)}{\rho}$. At the individual level, binary outcomes y_{ij} are generated from a bernoulli(p_j) distribution. In this way we can generate individual binary outcomes that are correlated within cluster, with ICC ρ.

```
/*CLUSTER LEVEL*/
set obs 100
gen cluster=_n
/* set ICC = .25*/
local rho=.25
/* set p0=.5 and δ = .05 */
local p0=.5
local p1=.55
/* rbeta is STATA's beta-binomial distribution command*/
gen p0=rbeta( (`p0'*(1-`rho')/`rho'),((1-`rho')*(1-`p0')/`rho'))
gen p1=rbeta( (`p1'*(1-`rho')/`rho'),((1-`rho')*(1-`p1')/`rho'))
local N=r(N)
gen treat=0
replace treat=1 if _n<=`N'/2
gen p=p0
replace p=p1 if treat==1
drop p0 p1
```

```
so cluster
tempfile cluster_p_g
save `cluster_p_g',replace
/*INDIVIDUAL LEVEL*/
clear
set obs 1000
* Generate clusters
gen u=invnormal(uniform())
egen cluster = cut(u), g(100)
replace cluster=cluster+1
so cluster
*merge in cluster errors
merge cluster using `cluster_p_g'
tab _m
drop _m
/*quick way to generate bernoulli(p) distributed data*/
gen y = ( uniform() < p )
```

C.5 How to create panel data

In this section we detail how to create panel data, with specific autocorrelation and ICC terms. Below, the values are set to $\rho_c = .2$, $\rho_p = .7$, the ICC $\rho = .3$, and as before, $\sigma = 10$:

```
/*CLUSTER LEVEL*/
set obs 100
gen cluster=_n
/* set σ = 10 */
gen grp=rnormal(0,10)
gen grp1=rnormal(0,10)
gen grp2=rnormal(0,10)
/* where we determine ρc */
gen group1=sqrt(.2)*grp +sqrt(.8)*grp1
gen group2=sqrt(.2)*grp +sqrt(.8)*grp2
sum
drop grp*
local N=r(N)
di `N'
gen treat=0
replace treat=1 if _n<=`N'/2
so cluster
save cluster_error,replace
/*INDIVIDUAL LEVEL*/
clear
set obs 1000
```

```
* Generate clusters
gen u=invnormal(uniform())
egen cluster = cut(u), g(100)
replace cluster=cluster+1
so cluster
*merge in cluster errors
merge cluster using cluster_error
tab _m
drop _m u
/* set σ = 10 */
gen indv=rnormal(0,10)
gen indv1=rnormal(0,10)
gen indv2=rnormal(0,10)
/* where we determine ρ_p */
gen individual1=sqrt(.7)*indv+sqrt(.3)*indv1
gen individual2=sqrt(.7)*indv+sqrt(.3)*indv2
drop indv*
/* where we set the ICC */
gen y0=10+(sqrt(.3))*group1 + (sqrt(.7))*individual1
/* allow y to increase by 2 in the second period, a common time trend*/
gen y1=12+(sqrt(.3))*group2 + (sqrt(.7))*individual2
drop group* individual*
/* set δ = 4 */
replace y1=y1+4 if treat==1
```

As part of the simulation, we might want to verify that the data is being generated with the correct ρ_p and ρ_c. Below is some simple code to estimate these parameters from the simulation directly above:

```
/* construct cluster means of outcome variable */
egen y1_cluster=mean(y1),by(cluster)
egen y0_cluster=mean(y0),by(cluster)
/*construct individual component of outcome variable */
gen y1_individual=y1-y1_cluster
gen y0_individual=y0-y0_cluster
/* ρ_c estimate */
corr y1_cluster y0_cluster
/* ρ_p estimate*/
corr y1_individual y0_individual
```

C.6 How to compute power through simulation

In this section we present code in order to simulate power as detailed in section 6.2. The basic idea here is to run 1000 simulations, and then count the number of times the

treatment coefficient is statistically significantly different from zero. Counting the number of times we get a statistically significant parameter, and dividing it by 1000 yields our simulated power. In the code below, we refer to the seven steps outlined in section 6.2 in order to outline how one may proceed with simulating power. The example below is for a continuous outcome with cluster-randomized treatment.

```
/*STEP 1*/
local numits=1000
local it=0
/ create a temporary file that will store the output from the simulations
*/
tempname memhold
tempfile montec_results
postfile `memhold' reject_t rho using `montec_results'
/* start quietly */
qui{
/*start iterations here*/
while `it'<=`numits'{
local it=`it'+1
clear
/* STEPS 2 and 3 */
*cluster errors
set obs 100
gen cluster=_n
gen group=rnormal(0,10)
sum
local N=r(N)
di `N'
gen treat=0
replace treat=1 if _n<=`N'/2
so cluster
tempfile cluster_error_g
save `cluster_error_g',replace
clear
set obs 1000
* Generate clusters
gen u=invnormal(uniform())
egen cluster = cut(u), g(100)
replace cluster=cluster+1
so cluster
*merge in cluster errors
merge cluster using `cluster_error_g'
tab _m
```

```
drop _m
gen individual=rnormal(0,10)
/* STEP 4*/
gen y=10+(sqrt(.3))*group + (sqrt(.7))*individual
replace y=y+4 if treat==1
loneway y cluster if treat==0
local rho=r(rho)
regr y treat,cluster(cluster)
/* STEP 5 */
local t_loop=_b[treat]/_se[treat]
local df=100
local critical_u=invttail(`df',.05/2)
local critical_l=invttail(`df',1-.05/2)
local reject_t=(`t_loop'>`critical_u')|(`t_loop'<`critical_l')
di `reject_t'
/*write output from simulation to the temporary file*/
post `memhold' (`reject_t') (`rho')
clear
}
/* STEP 6 */
} /* close quietly*/
postclose `memhold'
use `montec_results',clear
/* STEP 7 */
/* reject=1 if null is rejected, 0 otherwise.  So the mean of reject
from the 1000 simulation draws will yield the simulated power */
sum reject_t rho
```

D STATA Programmes to Compute Power with Binary Outcomes

Below is some code that creates a STATA .ado programme in order to get power for binary outcomes. The first section of code immediately below is for the case where there are no covariates:

```
cap prog drop discretepower
program discretepower
syntax anything [,alpha(real .05) beta(real .8) pi(real .5)]
tokenize "`0'",parse(" ,")
local rho =`1'
local m =`2'
local p_0 =`3'
local impact =`4'
```

```
local z_alpha = invnormal(1-(`alpha'/2))
local z_beta = invnormal(`beta')
local deff = (1+`rho'*(`m'-1))
local p_1 =`p_0'+`impact'
local k =(`deff'/`m')*( (`p_1'*(1-`p_1')/`pi')+(`p_0'*(1-`p_0')/(1-`pi'))
) *((`z_alpha'+`z_beta')^2)/`impact'^2
di
di "Total number of clusters= " `k'
end
```

Here is an example of how to use this .ado file for the case where $p_0 = 0.5$, $\delta = .1$, $\rho = .05$ and $m = 30$. The order in which these parameters must be entered is specified by the positional arguments within the .ado file. For example, see the line `local rho =`1'` . This tells us that the value of ρ should be entered first. Looking at the proceeding lines, we also see that the order of the remaining parameters are to be entered is m, p_0 and finally the design effect δ. Looking at the third line of code above (starting with the word syntax), there are several options that may be changed from the default values - α (set at .05), β (set at .8) and π (set at .5):

```
discretepower .05 30 .5 .1
```

Now for the more complicated code below, where we allow a single discrete covariate, as we saw in section 5.3. The code below allows for a discrete X with five points of support. The code could easily be shortened for a simple binary X, or extended for more points of support. The code for points of support up to 5 is supplied as supplementary material:

```
cap prog drop discretepowerX
program discretepowerX
syntax anything [,alpha(real .05) beta(real .8) pi(real .5)]
tokenize "`0'",parse(" ,")
local rho =`1'
local m =`2'
local x_0 =`3'
local theta_0 =`4'
local p_C0 =`5'
local impact_0 =`6'
local x_1 =`7'
local theta_1 =`8'
local p_C1 =`9'
local impact_1 =`10'
local x_2 =`11'
```

```
local theta_2 =`12'
local p_C2 =`13'
local impact_2 =`14'
local x_3 =`15'
local theta_3 =`16'
local p_C3 =`17'
local impact_3 =`18'
local x_4 =`19'
local theta_4 =`20'
local p_C4 =`21'
local impact_4 =`22'
local p_C =(`theta_0'*`p_C0' + `theta_1'*`p_C1' + `theta_2'*`p_C2' +
`theta_3'*`p_C3' + `theta_4'*`p_C4')
local p_T0 =`p_C0'+`impact_0'
local p_T1 =`p_C1'+`impact_1'
local p_T2 =`p_C2'+`impact_2'
local p_T3 =`p_C3'+`impact_3'
local p_T4 =`p_C4'+`impact_4'
local deff = (1+`rho'*(`m'-1))
local impact =(`theta_0'*`impact_0' + `theta_1'*`impact_1' + `theta_2'*`impact_2
+ `theta_3'*`impact_3 ' + `theta_4'*`impact_4')
local z_alpha = invnormal(1-(`alpha'/2))
local z_beta = invnormal(`beta')
#delimit ;
matrix M=(
(`pi'*`theta_0'*`p_T0'*(1-`p_T0') + (1-`pi')*`theta_0'*`p_C0'*(1-`p_C0'))
+
(`pi'*`theta_1'*`p_T1'*(1-`p_T1') + (1-`pi')*`theta_1'*`p_C1'*(1-`p_C1'))
+
(`pi'*`theta_2'*`p_T2'*(1-`p_T2') + (1-`pi')*`theta_2'*`p_C2'*(1-`p_C2'))
+
(`pi'*`theta_3'*`p_T3'*(1-`p_T3') + (1-`pi')*`theta_3'*`p_C3'*(1-`p_C3'))
+
(`pi'*`theta_4'*`p_T4'*(1-`p_T4') + (1-`pi')*`theta_4'*`p_C4'*(1-`p_C4'))
,
(`pi'*`theta_0'*`p_T0'*(1-`p_T0')) +
(`pi'*`theta_1'*`p_T1'*(1-`p_T1')) +
(`pi'*`theta_2'*`p_T2'*(1-`p_T2')) +
(`pi'*`theta_3'*`p_T3'*(1-`p_T3')) +
(`pi'*`theta_4'*`p_T4'*(1-`p_T4'))
,
`x_0'*(`pi'*`theta_0'*`p_T0'*(1-`p_T0') + (1-`pi')*`theta_0'*`p_C0'*(1-`p_C0'))
+
```

```
`x_1`*(`pi`*`theta_1`*`p_T1`*(1-`p_T1`) + (1-`pi`)*`theta_1`*`p_C1`*(1-`p_C1`))
+
`x_2`*(`pi`*`theta_2`*`p_T2`*(1-`p_T2`) + (1-`pi`)*`theta_2`*`p_C2`*(1-`p_C2`))
+
`x_3`*(`pi`*`theta_3`*`p_T3`*(1-`p_T3`) + (1-`pi`)*`theta_3`*`p_C3`*(1-`p_C3`))
+
`x_4`*(`pi`*`theta_4`*`p_T4`*(1-`p_T4`) + (1-`pi`)*`theta_4`*`p_C4`*(1-`p_C4`))
\
(`pi`*`theta_0`*`p_T0`*(1-`p_T0`)) +
(`pi`*`theta_1`*`p_T1`*(1-`p_T1`)) +
(`pi`*`theta_2`*`p_T2`*(1-`p_T2`)) +
(`pi`*`theta_3`*`p_T3`*(1-`p_T3`)) +
(`pi`*`theta_4`*`p_T4`*(1-`p_T4`))
,
(`pi`*`theta_0`*`p_T0`*(1-`p_T0`)) +
(`pi`*`theta_1`*`p_T1`*(1-`p_T1`)) +
(`pi`*`theta_2`*`p_T2`*(1-`p_T2`)) +
(`pi`*`theta_3`*`p_T3`*(1-`p_T3`)) +
(`pi`*`theta_4`*`p_T4`*(1-`p_T4`))
,
`x_0`*(`pi`*`theta_0`*`p_T0`*(1-`p_T0`)) +
`x_1`*(`pi`*`theta_1`*`p_T1`*(1-`p_T1`)) +
`x_2`*(`pi`*`theta_2`*`p_T2`*(1-`p_T2`)) +
`x_3`*(`pi`*`theta_3`*`p_T3`*(1-`p_T3`)) +
`x_4`*(`pi`*`theta_4`*`p_T4`*(1-`p_T4`))
\
`x_0`*(`pi`*`theta_0`*`p_T0`*(1-`p_T0`) + (1-`pi`)*`theta_0`*`p_C0`*(1-`p_C0`))
+
`x_1`*(`pi`*`theta_1`*`p_T1`*(1-`p_T1`) + (1-`pi`)*`theta_1`*`p_C1`*(1-`p_C1`))
+
`x_2`*(`pi`*`theta_2`*`p_T2`*(1-`p_T2`) + (1-`pi`)*`theta_2`*`p_C2`*(1-`p_C2`))
+
`x_3`*(`pi`*`theta_3`*`p_T3`*(1-`p_T3`) + (1-`pi`)*`theta_3`*`p_C3`*(1-`p_C3`))
+
`x_4`*(`pi`*`theta_4`*`p_T4`*(1-`p_T4`) + (1-`pi`)*`theta_4`*`p_C4`*(1-`p_C4`))
,
`x_0`*(`pi`*`theta_0`*`p_T0`*(1-`p_T0`)) +
`x_1`*(`pi`*`theta_1`*`p_T1`*(1-`p_T1`)) +
`x_2`*(`pi`*`theta_2`*`p_T2`*(1-`p_T2`)) +
`x_3`*(`pi`*`theta_3`*`p_T3`*(1-`p_T3`)) +
`x_4`*(`pi`*`theta_4`*`p_T4`*(1-`p_T4`))
,
```

```
`x_0'^2*(`pi'*`theta_0'*`p_T0'*(1-`p_T0') + (1-`pi')*`theta_0'*`p_C0'*(1-`p_C0')
+
`x_1'^2*(`pi'*`theta_1'*`p_T1'*(1-`p_T1') + (1-`pi')*`theta_1'*`p_C1'*(1-`p_C1')
+
`x_2'^2*(`pi'*`theta_2'*`p_T2'*(1-`p_T2') + (1-`pi')*`theta_2'*`p_C2'*(1-`p_C2')
+
`x_3'^2*(`pi'*`theta_3'*`p_T3'*(1-`p_T3') + (1-`pi')*`theta_3'*`p_C3'*(1-`p_C3')
+
`x_4'^2*(`pi'*`theta_4'*`p_T4'*(1-`p_T4') + (1-`pi')*`theta_4'*`p_C4'*(1-`p_C4')
);
mat invM=invsym(M);
matrix g=(
`theta_0'*(`p_T0'*(1-`p_T0') - `p_C0'*(1-`p_C0')) +
`theta_1'*(`p_T1'*(1-`p_T1') - `p_C1'*(1-`p_C1')) +
`theta_2'*(`p_T2'*(1-`p_T2') - `p_C2'*(1-`p_C2')) +
`theta_3'*(`p_T3'*(1-`p_T3') - `p_C3'*(1-`p_C3')) +
`theta_4'*(`p_T4'*(1-`p_T4') - `p_C4'*(1-`p_C4'))
,
`theta_0'*(`p_T0'*(1-`p_T0')) +
`theta_1'*(`p_T1'*(1-`p_T1')) +
`theta_2'*(`p_T2'*(1-`p_T2')) +
`theta_3'*(`p_T3'*(1-`p_T3')) +
`theta_4'*(`p_T4'*(1-`p_T4'))
,
`theta_0'*`x_0'*(`p_T0'*(1-`p_T0') - `p_C0'*(1-`p_C0')) +
`theta_1'*`x_1'*(`p_T1'*(1-`p_T1') - `p_C1'*(1-`p_C1')) +
`theta_2'*`x_2'*(`p_T2'*(1-`p_T2') - `p_C2'*(1-`p_C2')) +
`theta_3'*`x_3'*(`p_T3'*(1-`p_T3') - `p_C3'*(1-`p_C3')) +
`theta_4'*`x_4'*(`p_T4'*(1-`p_T4') - `p_C4'*(1-`p_C4'))
);
matrix gprime=g';
matrix A=g*invM*gprime;
# delimit cr;
local A=A[1,1]
local k = (`deff'/`m')*(`A') * (`z_alpha'+`z_beta')^2) / `impact'^2
di "k==" `k'
end
```

As in the simpler case above, the order in which one must enter the paramters is defined by the positional arguments in the .ado - in this case complicated case, 22 parameters are required. Here is an example of how to use the programme for a case of $\rho = .05$ and $m = 30$. We know how to order the parameters by referring to the positional arguments at the beginning of this code. So, we see to start, the order is ρ first, then

m, then x_0, followed by θ_0 and so on for the remaining 18 parameters:

```
discretepowerX 0.05 30 -2 .2 .1 0 -1 .2 .4 .05 0 .2 .5 .1 1 .2 .6
.15 2 .2 .9 .2
```

To clarify, the full syntax for this prgramme is:
`discretepowerX` ρ m x_0 θ_0 p_{C1} δ_0 x_1 θ_1 p_{C1} δ_1 x_2 θ_2 p_{C2} δ_2 x_3 θ_3 p_{C3} δ_3 x_4 θ_4 p_{C4} δ_4

Table 1: Total Sample Size Requirements for Continuous Outcomes under Cluster Randomisation

		Total Sample Size Requirements (n*)				Number of Clusters, k*			
		Effect size =10000							
		numbers of individuals per cluster (m)				*numbers of individuals per cluster (m)*			
		10	30	60	100	10	30	60	100
ICC (ρ)	0	2508	2508	2508	2508	251	84	42	25
	0.01	2743	3264	4046	5089	274	109	67	51
	0.03	3194	4718	7004	10053	319	157	117	101
	0.05	3646	6173	9963	15017	365	206	166	150
	0.1	4774	9808	17360	27428	477	327	289	274
	0.2	7030	17079	32153	52251	703	569	536	523
		Effect size =20000							
		numbers of individuals per cluster (m)				*numbers of individuals per cluster (m)*			
		10	30	60	100	10	30	60	100
ICC (ρ)	0	628	628	628	628	63	21	10	6
	0.01	693	839	1058	1351	69	28	18	14
	0.03	806	1202	1796	2589	81	40	30	26
	0.05	919	1565	2536	3829	92	52	42	38
	0.1	1201	2474	4384	6931	120	82	73	69
	0.2	1765	4292	8083	13136	177	143	135	131

The cells in the left panels report the total sample size per treatment arm (m*k) and the right panels report the number of clusters per treatment arm (k) required to achieve 80% power at 5% significance if the effect size is either 10,000 (top panel) or 20,000 (bottom panel) and the standard deviation is 126383.5. The intra cluster correlation (ρ) is given in the first column of the Table.

Table 2: Total Sample Size Requirements for Continuous Outcomes under Cluster Randomisation with a Covariate

		numbers of individuals per cluster (m) = 100					*numbers of individuals per cluster (m) = 20*					*numbers of individuals per cluster (m) = 8*				
		R_p^2					R_p^2					R_p^2				
ICC (ρ) = 0.01	R_c^2	0	0.1	0.2	0.4	0.5	0	0.1	0.2	0.4	0.5	0	0.1	0.2	0.4	0.5
	0	1351	1294	1232	1110	1048	766	704	642	518	456	679	617	555	430	368
	0.1	1293	1237	1176	1054	993	754	692	630	506	444	674	612	550	426	364
	0.2	1231	1175	1114	993	933	741	679	617	494	432	669	607	545	421	359
	0.4	1107	1052	992	871	812	716	654	592	469	407	659	597	535	411	349
	0.5	1045	991	931	811	752	704	642	580	456	394	654	592	530	406	344
		R_p^2					R_p^2					R_p^2				
ICC (ρ) = 0.3	R_c^2	0	0.1	0.2	0.4	0.5	0	0.1	0.2	0.4	0.5	0	0.1	0.2	0.4	0.5
	0	19342	19298	19254	19167	19123	4219	4176	4132	4044	4000	1951	1907	1863	1776	1732
	0.1	17462	17418	17374	17287	17243	3843	3800	3756	3668	3624	1801	1757	1713	1625	1581
	0.2	15581	15538	15494	15406	15362	3467	3423	3380	3292	3248	1650	1606	1562	1475	1431
	0.4	11821	11777	11733	11645	11602	2715	2671	2627	2540	2496	1349	1305	1262	1174	1130
	0.5	9940	9897	9853	9765	9721	2339	2295	2251	2164	2120	1199	1155	1111	1023	980

Each cell reports the total sample size (m*k) required to achieve 80% power at 5% significance if the effect size is 20,000 and the standard deviation is 126383.5. The number of individuals per cluster (m)is 100 in the left panel, 20 in the middle panel and 8 in the right panel. The intra cluster correlation (ρ) is 0.01 in the top panel and 0.3 in the bottom panel. Rc2 is the proportion of the cluster-level variance component explained by the covariate, and Rp2 is its individual-level equivalent.

Table 3: Sample size Requirements for Continuous Outcomes in Panel Data Models

		Difference	Difference-in-differences	Lagged Outcome
r	0.1	4909	8820	4860
	0.25	4909	7354	4603
	0.5	4909	4909	3687
	0.75	4909	2464	2159
	0.9	4909	998	949

The ICC is .05 and the number of individuals per cluster , m, is set to 20. Effect size is equal to 10000 and the standard deviation is 126383.5

Table 4: Sample Size Requirements for Discrete Outcomes Under Cluster Randomisation

		Total Sample Size Requirements (N*)					Number of Clusters (2k*)			
		Control Group Success Rate (p0):				0.1				
		numbers of individuals per cluster (m)					*numbers of individuals per cluster (m)*			
		10	30	60	100		10	30	60	100
ICC	0	392	392	392	392		39	13	7	4
	0.01	428	506	624	781		43	17	10	8
	0.03	498	734	1087	1558		50	24	18	16
	0.05	569	961	1550	2335		57	32	26	23
	0.1	746	1531	2708	4278		75	51	45	43
	0.2	1099	2669	5023	8163		110	89	84	82
		Control Group Success Rate (p0):				0.3				
		numbers of individuals per cluster (m)					*numbers of individuals per cluster (m)*			
		10	30	60	100		10	30	60	100
ICC	0	706	706	706	706		71	24	12	7
	0.01	770	911	1123	1406		77	30	19	14
	0.03	897	1321	1957	2804		90	44	33	28
	0.05	1024	1731	2790	4203		102	58	47	42
	0.1	1342	2755	4874	7700		134	92	81	77
	0.2	1978	4804	9042	14693		198	160	151	147
		Control Group Success Rate (p0):				0.5				
		numbers of individuals per cluster (m)					*numbers of individuals per cluster (m)*			
		10	30	60	100		10	30	60	100
ICC	0	769	769	769	769		77	26	13	8
	0.01	838	992	1223	1531		84	33	20	15
	0.03	977	1438	2131	3054		98	48	36	31
	0.05	1115	1885	3038	4577		112	63	51	46
	0.1	1461	3000	5307	8384		146	100	88	84
	0.2	2154	5230	9846	15999		215	174	164	160

Effect size is set to .1 and treatment is evenly allocated (π=.5).

Table 5: Number of Clusters Required for Discrete Outcomes Under Cluster Randomisation With A Binary Covariate

Control group success rates for $X_j=0/X_j=1$	ICC=.05			ICC=.1		
	Impacts for $X_j=0/X_j=1$			Impacts for $X_j=0/X_j=1$		
	.1/.1	.05/.15	.03/.17	.1/.1	.05/.15	.03/.17
.45/.55	50	49	49	88	86	85
.4/.6	49	47	47	85	83	81
.3/.7	42	40	39	74	70	68
.2/.8	32	29	27	56	50	47

Number of individuals per cluster,m ,is set at 60. The overall base rate in this table is set to .5, with the overall impact set to .1. Treatment is evenly allocated (π=.5), and θ= P(Xj=1)=.5.