Understanding the principles of power calculations with non-continuous outcomes

Andres Martinez, University of Michigan
Elisa Rothenbühler, World Bank

Cluster-randomized trials (CRTs) have become increasingly common in the social sciences as a means for evaluating the effectiveness of interventions. The natural clustering that arises in many contexts, i.e. the fact that certain objects (e.g. villages) are naturally assigned to specific groups (e.g. health facilities) because of similarities across those objects (e.g. same distance to the health facility), and the fact that programs are often implemented at the group level makes CRTs a logical choice (Bloom, 2005a, 2005b; Boruch et al. 2000; Cook, 2005). When feasible, CRTs are the ideal approach for establishing causal relationships (Boruch, 2005).

To be effective, CRTs need to be well designed and implemented. The sheer existence of a CRT to evaluate a program or policy is not enough to generate rigorous evidence of the effectiveness (or lack thereof) of a program. The focus of this example is on the statistical power to detect a treatment effect—one of the multiple key study design elements. Power in this context is the probability to statistically find an effect given that such effect exists. Naturally, we want to design studies with high statistical power as under-powered studies will fail to (statistically) detect an effect, even if that effect exists. In conducting a power analysis for a program evaluation study a common approach is to provide the sample size required to detect a given treatment effect, although in some contexts a more realistic approach is to provide the smallest treatment effect that is statistically detectable given a certain sample size. The former is sometimes called the sample size approach or the ex-ante approach, while the

latter is sometimes called the minimum detectable effect (MDE) approach or the ex-post approach. The calculations that underlie these approaches rely on the same set of formulas so using one approach over the other is dictated by study constraints or researcher preference.

The factors that determine statistical power in CRTs with continuous outcomes have been well-documented (e.g., Donner and Klar, 2000; Bloom, 1999; Murray, 1998; Raudenbush, 1997). Two general points are worth highlighting. First, the number of clusters is more important (and often, far more important) than the number of individuals within the clusters to achieve a set level of statistical power. In some cases, it may actually not be possible to achieve certain levels of statistical power by solely increasing the number of individuals within the clusters. Second, the proportion of total outcome variability lying between the clusters, known as the intraclass correlation (ICC), is inversely related to the amount of statistical power: a larger ICC means less statistical power. In a power analysis for a CRT, it is then crucial to have an estimate of the ICC in order to calculate either the required sample sizes at each level (e.g., number of clusters and number of individuals within the clusters) given an expected treatment effect, or the minimum detectable treatment effect given a set of sample sizes at each level.[1] Other factors influencing the statistical power include blocking and the use of covariates. Raudenbush et al. (2007) offer a detailed discussion of these factors in the context of a regular CRT.

The factors that determine statistical power in CRTs with non-continuous outcomes are not as extensively documented. In addition, seemingly divergent approaches have been

[1] Some analysts opt for using a design effect that is a function of the ICC instead of the ICC itself. We shy away from this approach, particularly for the more complex designs (e.g., with more than two levels of nesting), as we find talking about ICCs at each level to be more intuitive.

proposed. For example, for binary outcomes, Murray (1998) uses many of the same parameters as for the continuous outcomes case, including the ICC, while Moerbeek et al. (2001) propose an approach that does not use the ICC but the unstandardized within- and between-cluster variances instead. However, regardless of the approach, the two main general points about power for CRTs with continuous outcomes remain true for CRTs with binary outcomes: the number of clusters and the amount of between-cluster variation play the most crucial roles in determining the minimum detectable treatment effect. The following example builds on the approach proposed by Moerbeek et al. (2001).[2] We adopt this approach because of the type of information that needs to be elicited and because of the flexibility it allows. We also believe it relies on a more sensible set of assumptions.[3]

A CRT with a binary outcome

In a regular CRT, the main outcomes of interest are often defined at the individual level, with individuals nested within clusters and clusters being randomly assigned into treatment conditions. To be concrete, consider a CRT with 40 health facilities and 50 individuals per health facility in which the outcome of interest is whether an individuals' health improves (coded 1) or worsens (coded 0). Based on historical data, 67% of all individuals see their health improve under the comparison condition, with health facility-level percentages ranging from 55 to 90%. Researchers believe that the treatment being evaluated will boost success rates by 7 percentage points. A success in this context is simply the manifestation of the event under

[2] This is the approach adopted among other places in the Optimal Design software (Raudenbush et al., 2011). See the software documentation for details.
[3] See Spybrook et al. (2011) for details.

study. For example, a success can be a pregnant woman attending at least four antenatal care visits; conversely, a success does not necessarily mean a favorable outcome, and could also mean a newborn with low birth weight. So, for each individual we either observe or not the event (a positive or negative outcome) and for each cluster we can obtain the percentage of individuals exhibiting a success in that event (the manifestation of a positive or negative outcome). Furthermore, assume one treatment and one comparison group and a fully balanced design (equal allocation of health facilities into the treatment and comparison groups, and health facilities with the same number of patients).

Following the hierarchical linear modeling (HLM) framework (Raudenbush & Bryk, 2002), patients are at level 1 and health facilities at level 2. The underlying statistical model can be conceived as a generalized linear mixed model in which y_{ij} is the observed outcome for patient $i = \{1, \dots, n\}$ in health facility $j = \{1, \dots, J\}$, with $y_{ij} = 1$ if patient i in health facility j gets well (often called a "success") and $y_{ij} = 0$ if not. Let φ_{ij} be the probability of observing $y_{ij} = 1$ for patient i in health facility j. Using a logit link function, the model can then be written as

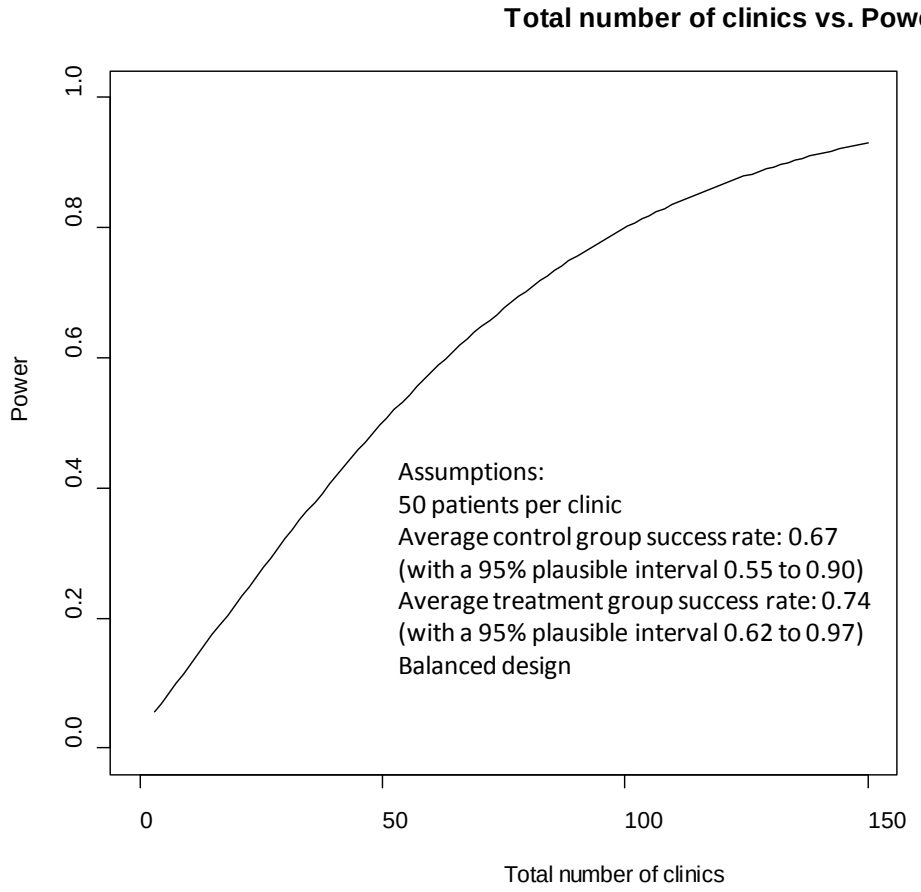
$$\log\left(\frac{\varphi_{ij}}{1-\varphi_{ij}}\right) = \gamma_{00} + \gamma_{01}W_j + u_{0j}$$

where γ_{00} is the average log-odds of success across all health facilities; γ_{01} is the treatment effect in log-odds metric; W_j is a contrast indicator taking a value of ½ for the treatment group and -½ for the comparison group; $u_{0j} \sim N(0, \tau)$ is the random effect associated with each health facility; and τ is the between health facility variance in log odds. Note the sampling model for the observed outcome is defined by φ_{ij} (the probability of observing $y_{ij} = 1$) so the

expected value of $y_{ij}|\varphi_{ij}$ is φ_{ij} with variance $\varphi_{ij}(1-\varphi_{ij})$. The derivation of the test statistic for $H_0: \gamma_{01} = 0$ vs. $H_1: \gamma_{01} \neq 0$ can be found in Moerbeek et al. (2001) and in the Optimal Design software documentation.

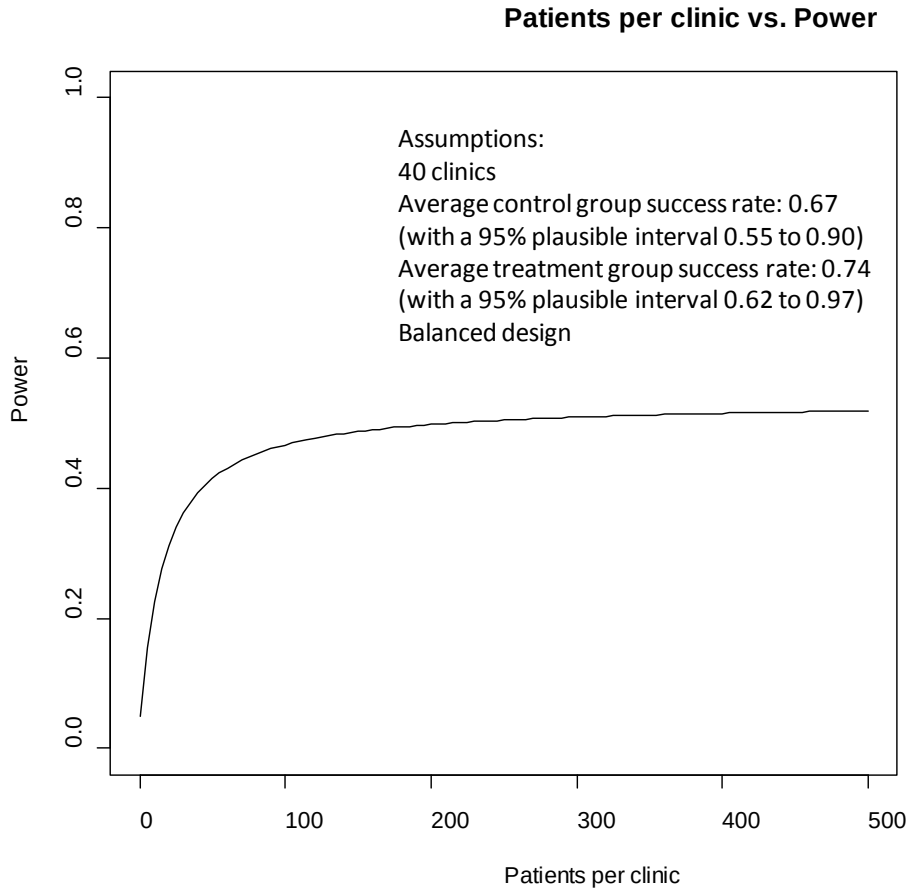
Sample size or Ex-ante approach

The figure below displays the total number of health facilities against the amount of statistical power. Note how power increases as the number of health facilities increases. With a total of 40 health facilities, half in the treatment group and half in the comparison group, and with 50 patients per health facility, the power to detect a difference in success rates from 0.67 to 0.67+0.07 i.e. 0.74 is only about 0.41. In order to have 0.80 power, about 100 health facilities would be needed, and to have 0.90 power, a total of about 134 health facilities would be needed.

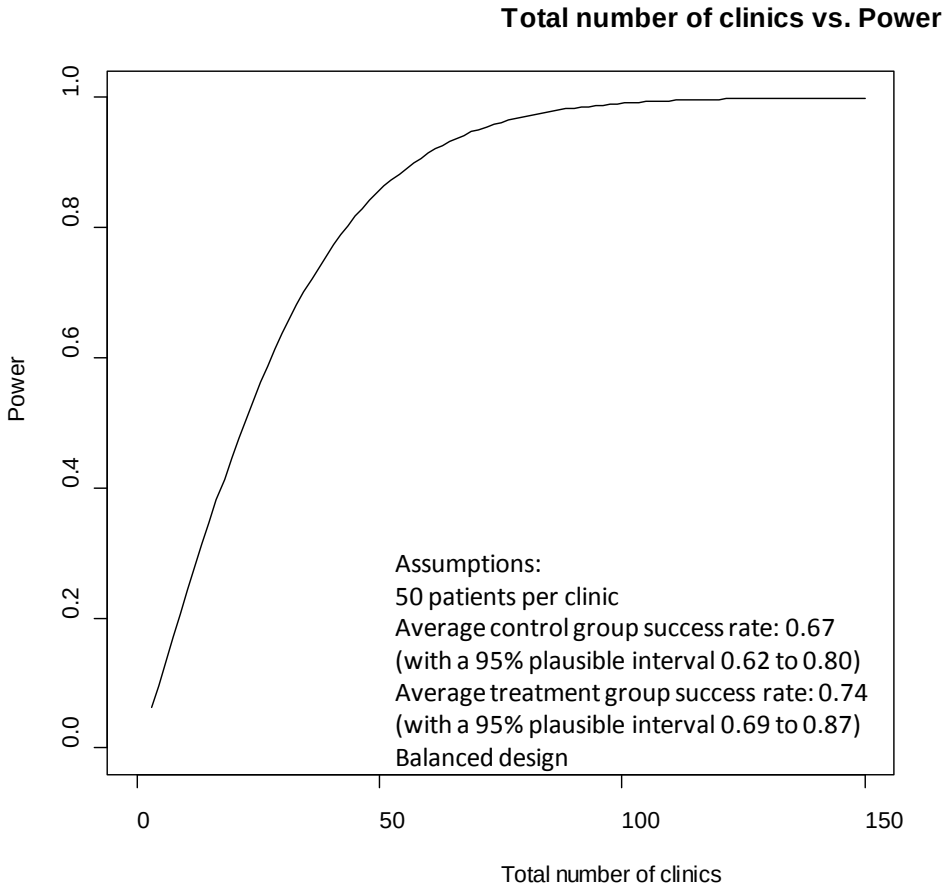


To see how the number of patients per health facility (the within-cluster sample size) has a different, more limited impact in the amount of statistical power, consider increasing the number of patients per health facility. As shown in the next figure, it would impossible to achieve acceptable levels of statistical power to detect a change in success rates of 7 percentage points from 0.67 to 0.74 with only 40 health facilities regardless of the number of patients per health facility. Actually, the maximum amount of power such a study would achieve would be about 0.53, which is unacceptable. The obvious conclusion here is that in order to achieve certain levels of statistical power, the number of clusters is far more important than the number of individuals per cluster. In many cases the amount of power will not reach

acceptable levels regardless of how many individuals per clusters are added if the number of clusters does not achieve a certain threshold.

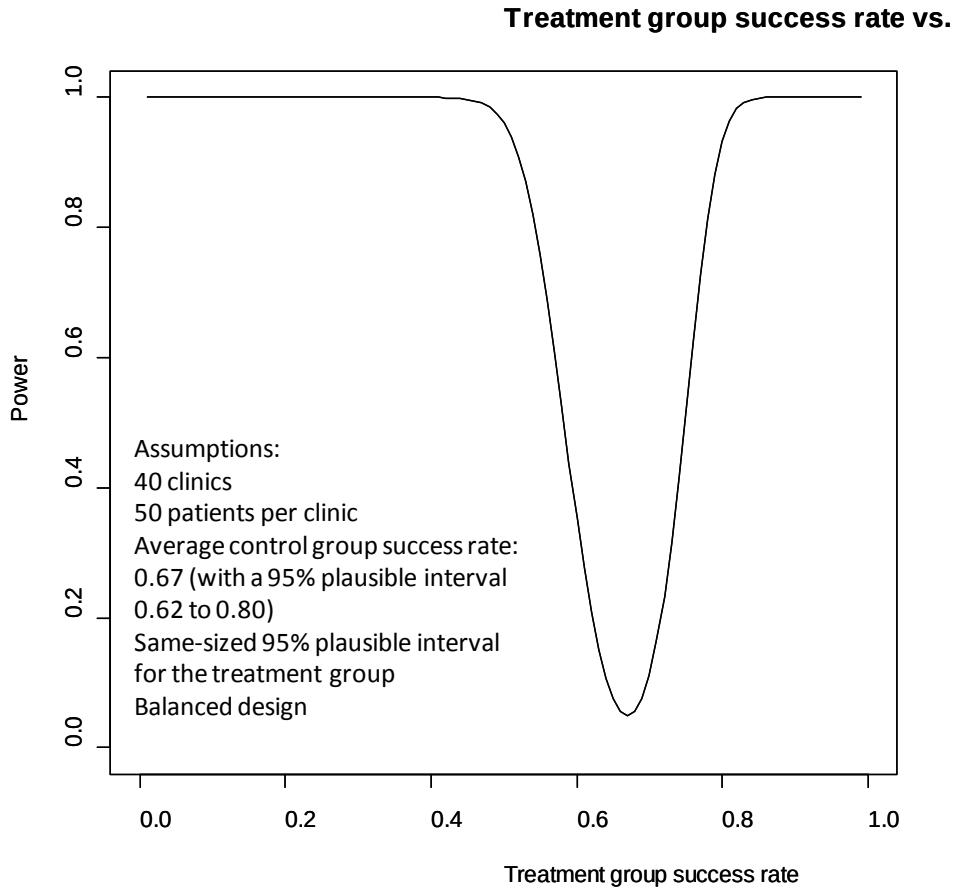


Consider now a narrower plausible interval for the between-cluster success rates (that is, smaller between-cluster variability). Suppose the plausible interval for the comparison group goes from 0.62 to 0.80, instead of the previous 0.55-0.90 interval. With the same 40 health facilities, power in this case would be about 0.77 and a 0.90 power would be achieved with 58 health facilities. The relationship is shown in the graph below.



Minimum Detectable Effect size or Ex-post approach

As was previously discussed, another approach to the power calculations is to start from a given sample size and determine the minimum detectable treatment effect that can be statistically detected to be different from zero. In this case, with 40 health facilities and 50 patients per health facility, the minimum detectable positive treatment effect at 0.80 power would be of about 16 percentage points. The relationship is shown in the graph below.



Discussion

Some general points are worth highlighting. First, for a given probability of success in the comparison group, power increases as the magnitude of the difference between that probability and the probability of success in the treatment increases. This makes intuitive sense given that as the difference between groups gets larger, the power increases. The relationship, however, is far from linear, and for a given comparison-treatment difference, power is maximized as the probability of success in the comparison group gets away from 0.5. Second, as was previously mentioned, the number of clusters is far more important than the number of individuals per cluster to achieve increased levels of statistical power. Third, power increases as

the plausible interval for the success rates decreases. Since the plausible interval for the success rates is a proxy for the between-cluster variance, a smaller interval means less between-cluster variance which translates into more statistical power. Issues not considered in this example are covariate adjustment and blocking. These strategies have the potential to increase the power of the design and should be explored during the design stage. In essence, both strategies seek to increase the power of the study by explaining some of the between-cluster variation (see Raudenbush et al., 2007 for a detailed discussion). Naturally, the quality of the power analysis depends heavily on the quality of the parameters used in those calculations. When possible, parameters estimates need to come from similar outcomes and contexts as the ones being planned for the evaluation study.

References

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs. Evaluation Review, 23(4), 445–469.

Bloom, H. S. (2005a). Randomizing Groups to Evaluate Place-Based Programs. In H. S. Bloom (Ed.), Learning More from Social Experiments: Evolving Analytic Approaches (pp. 115–172). New York: Russell Sage Foundation.

Bloom, H. S. (Ed.). (2005b). Learning More from Social Experiments: Evolving Analytical Approaches. New York, NY: Russell Sage Foundation.

Boruch, R. (Ed.). (2005). Place Randomized Trials: Experimental Tests of Public Policy. The Annals of the American Academy of Political and Social Sciences (Vol. 599). Thousand Oaks, CA: SAGE.

Boruch, R., Snyder, B., & DeMoya, D. (2000). The Importance of Randomized Field Trials. Crime and Delinquency, 46(2), 156–80.

Cook, T. D. (2005). Emergent Principles for the Design, Implementation and Analysis of Cluster-Based Experiments in Social Science. In R. Boruch (Ed.), Place Randomized Trials: Experimental Tests of Public Policy. The Annals of The American Academy of Political and Social Science (Vol. 599, pp. 176–198). Thousand Oaks: Sage Publications.

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

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal Experimental Designs for Multilevel Logistic Models. Journal of the Royal Statistical Society (Series D): The Statistican, 50(1), 17–30. doi:10.1111/1467-9884.00257

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

Raudenbush, S. W. (1997). Statistical Analysis and Optimal Design for Cluster Randomized Trials. Psychological Methods, 2(2), 173–185.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical Linear Models: Applications and Data Analysis Methods (2nd ed.). Thousand Oaks, California: Sage Publications.

Raudenbush, S. W., et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01) [Software]. Available from www.wtgrantfoundation.org or from sitemaker.umich.edu/group-based.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. Educational Evaluation and Policy Analysis, 29(1), 5–29.

Spybrook, J., et al. (2011). Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software Version 3.0. Available from www.wtgrantfoundation.org or from sitemaker.umich.edu/group-based.