Snijders, Tom A.B. 'Power and Sample Size in Multilevel Linear Models'.
In: B.S. Everitt and D.C. Howell (eds.),
Encyclopedia of Statistics in Behavioral Science. Volume 3, 1570–1573. Chicester (etc.): Wiley, 2005.

Power and sample size in multilevel modeling

Author: Tom A.B. Snijders

t.a.b.snijders@ppsw.rug.nl

Abstract:
Sample size determination in multilevel designs requires attention to the fact that statistical power depends on the total sample sizes for each level. It is usually desirable to have as many units as possible at the top level of the multilevel hierarchy. Some formulae are given to obtain insight in the design aspects that are most influential for standard errors and power.

Keywords:
power, statistical tests, design, multilevel analysis, sample size, multisite trial, cluster randomisation.

Power and sample size in multilevel modeling

Power of statistical tests generally depends on sample size and other design aspects; on effect size or, more generally, parameter values; and on the level of significance. In multilevel models, however, there is a sample size for each level, defined as the total number of units observed for this level. E.g., in a three-level study of pupils nested in classrooms nested in schools, there might be observations on 60 schools, a total of 150 classrooms, and a total of 3,300 pupils. On average in the data, each classroom then has 22 pupils, and each school contains 2.5 classrooms. What are the relevant sample sizes for power issues? If the researcher has the freedom to choose the sample sizes for a planned study, what are sensible guidelines?

Power depends on the parameter being tested, and power considerations are different depending on whether the researcher focuses on, e.g., testing a regression coefficient, a variance parameter, or is interested in the size of means of particular groups. In most studies, attention goes primarily to regression coefficients, and this article focuses on such coefficients. The cited literature gives methods to determine power and required sample sizes also for estimating parameters in the random part of the model.

A primary qualitative issue is that, for testing the effect of a level-one variable, the level-one sample size (in the example, 3,300) is of main importance; for testing the effect of a level-two variable it is the level-two sample size (150 in the example); etc. The average cluster sizes (in the example, 22 at level two and 2.5 at level three) are not very important for the power of such tests. This implies that the sample size at the highest level is the main limiting characteristic of the design. Almost always, it will be more informative to have a sample of 60 schools with 3,300 pupils than one of 30 schools also with 3,300 pupils. A sample of 600 schools with a total of 3,300 pupils would even be a lot better with respect to power, in spite of the low average number of students (5.5) sampled per school, but in practice such a study would of course be much more expensive. A second qualitative issue is that for testing fixed regression coefficients, small cluster sizes are not a problem. The low average number of 2.5 classrooms per school has in itself no negative consequences for the power of testing regression coefficients. What is limited by this low average cluster size, is the power for testing random slope variances at the school level, i.e., between-school variances of effects of classroom- or pupil-level variables; and the reliability of estimating those characteristics of individual schools, calculated from classroom variables, that differ strongly between classes. (It may be recalled that the latter characteristics of individual units will be estimated in the multilevel methodology by posterior means, also called empirical Bayes estimates; see the Encyclopedia entry on *Predicting the random effects*.)

When quantitative insight is required in power for testing regression coefficients, it often is convenient to consider power as a consequence of the standard error of estimation. Suppose we wish to test the null hypothesis that a regression coefficient γ is 0, and for this coefficient we have an estimate $\hat{\gamma}$ which is approximately normally distributed, with standard error s.e.($\hat{\gamma}$). The t-ratio $\hat{\gamma}$/s.e.($\hat{\gamma}$) can be tested using a t-distribution; if the sample size is large enough, using a standard normal distribution. The power will be high if the true value (effect size) of γ is large and if the standard error is small; and a higher level of significance (larger permitted probability of a type I error) will lead to a higher power. This is expressed by the following formula, which holds for a one-sided test, and where the significance level is indicated by α and the type-two error probability, which is equal to 1 minus power, by β:

$$\frac{\gamma}{\text{s.e.}(\hat{\gamma})} \approx z_{1-\alpha} + z_{1-\beta} ,$$

where $z_{1-\alpha}$ and $z_{1-\beta}$ are the critical points of the standard normal distribution. E.g., to obtain at significance level $\alpha = 0.05$ a fairly high power of at

least $1-\beta = 0.80$, we have $z_{1-\alpha} = 1.645$ and $z_{1-\beta} = 0.84$, so that the ratio of true parameter value to standard error should be at least $1.645 + 0.84 \approx 2.5$.

For two-sided tests, the approximate formula is

$$\frac{\gamma}{\text{s.e.}(\hat{\gamma})} \approx z_{1-(\alpha/2)} + z_{1-\beta} ,$$

but in the two-sided case this formula holds only if the power is not too small (or, equivalently, the effect size is not too small), e.g., $1 - \beta \geq 0.3$.

For some basic cases, explicit formulae for the estimation variances (i.e., squared standard errors) are given below. This gives a basic knowledge of, and feeling for, the efficiency of multilevel designs. These formulae can be used to compute required sample sizes. The formulae also underpin the qualitative issues mentioned above. A helpful concept for developing this feeling is the *design effect*, which indicates how the particular design chosen – in our case, the multilevel design – affects the standard error of the parameters. It is defined as

$$\textit{deff} = \frac{\text{squared standard error under this design}}{\text{squared standard error under standard design}} ,$$

where the 'standard design' is defined as a design using a simple random sample with the same total sample size at level one. (The determination of sample sizes under simple random sample designs is treated in the article in this Encyclopedia on *Sample size and power calculation*.) If *deff* is greater than 1, the multilevel design is less efficient than a simple random sample design (with the same sample size); if it is less than 1, the multilevel design is more efficient. Since squared standard errors are inversely proportional to sample sizes, the required sample size for a multilevel design will be given by the sample size that would be required for a simple random sample design, multiplied by the design effect.

The formulae for the basic cases are given here (also see [11]) for two-level designs, where the cluster size is assumed to be constant, and denoted by n. These are good approximations also when the cluster sizes are variable but not too widely different. The number of level-two units is denoted m, so the total sample size at level one is mn. In all models mentioned below, the level-one residual variance is denoted $\text{var}(R_{ij}) = \sigma^2$ and the level-two residual variance by $\text{var}(U_{0j}) = \tau^2$.

1. For estimating a population mean μ in the model

$$Y_{ij} = \mu + U_{0j} + R_{ij} ,$$

the estimation variance is

$$\mathrm{var}(\hat{\mu}) = \frac{n\tau^2 + \sigma^2}{m\,n} \ .$$

The design effect is $\mathit{deff} = 1 + (n-1)\rho_{\mathrm{I}} \geq 1$, where ρ_{I} is the intraclass correlation, defined by $\rho_{\mathrm{I}} = \tau^2/(\sigma^2 + \tau^2)$.

2. For estimating the regression coefficient γ_1 of a level-one variable X_1 in the model

$$Y_{ij} \ = \ \gamma_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + \ldots + \gamma_p X_{pij} + U_{0j} + R_{ij} \ ,$$

where it is assumed that X_1 does not have a random slope and has zero between-group variation, i.e., a constant group mean, and that it is uncorrelated with any other explanatory variables X_k $(k \geq 2)$, the estimation variance is

$$\mathrm{var}(\hat{\gamma}) = \frac{\sigma^2}{m\,n\,s^2_{X1}} \ ,$$

where the within-group variance s^2_{X1} of X_1 also is assumed to be constant. The design effect here is $\mathit{deff} = 1 - \rho_{\mathrm{I}} \leq 1$.

3. For estimating the effect of a level-one variable X_1 under the same assumptions except that X_1 now does have a random slope,

$$Y_{ij} \ = \ \gamma_0 + (\gamma_1 + U_{1j})X_{1ij} + \gamma_2 X_{2ij} + \ldots + \gamma_p X_{pij} + U_{0j} + R_{ij} \ ,$$

where the random slope variance is τ^2_1, the estimation variance of the fixed effect is

$$\mathrm{var}(\hat{\gamma}) = \frac{n\,\tau^2_1\,s^2_{X1} + \sigma^2}{m\,n\,s^2_{X1}} \ ,$$

with design effect

$$\mathit{deff} = \frac{n\,\tau^2_1\,s^2_{X1} + \sigma^2}{\tau^2_1\,s^2_{X1} + \tau^2 + \sigma^2}$$

which can be greater or less than 1.

4. For estimating the regression coefficient of a level-two variable X_1 in the model

$$Y_{ij} \;=\; \gamma_0 + \gamma_1 X_{1j} + \gamma_2 X_{2ij} + \ldots + \gamma_p X_{pij} + U_{0j} + R_{ij} \;,$$

where the variance of X_1 is s^2_{X1}, and X_1 is uncorrelated with any other variables X_k ($k \geq 2$), the estimation variance is

$$\mathrm{var}(\hat{\gamma}) = \frac{n\,\tau^2 + \sigma^2}{m\,n\,s^2_{X1}} \;,$$

and the design effect is $\mathit{deff} = 1 + (n-1)\rho_{\mathrm{I}} \;\geq\; 1$.

This illustrates that multilevel designs sometimes are more, and sometimes less efficient than simple random sample designs. In case 2, a level-one variable without between-group variation, the multilevel design is always more efficient. This efficiency of within-subject designs is a well-known phenomenon. For estimating a population mean (case 1) or the effect of a level-two variable (case 4), on the other hand, the multilevel design always is less efficient, and more seriously so as the cluster size and the intraclass correlation are larger.

In cases 2 and 3, the same type of regression coefficient is being estimated, but in case 3 variable X_1 has a random slope, unlike in case 2. The difference in *deff* between these cases shows that the details of the multilevel dependence structure matters for these standard errors, and hence for the required sample sizes. It must be noted that what matters here is not how the researcher specifies the multilevel model, but the true model. If in reality there is a positive random slope variance but the researcher specifies a random intercept model without a random slope, then the true estimation variance still will be given by the formula above of case 3, but the standard error will be misleadingly estimated by the formula of case 2, usually a lower value.

For more general cases, where there are several correlated explanatory variables, some of them having random slopes, such clear formulae are not available. Sometimes a very rough estimate of required sample sizes can still be made on the basis of these formulae. For more generality, however, the program `PinT` (see below) can be used to calculate standard errors in rather general two-level designs.

A general procedure to estimate power and standard errors for any parameters in arbitrary designs is by Monte Carlo simulation of the model and the estimates. This is described in [4] (Chapter 10).

For further reading, general treatments can be found in [1], [2], [4] (Chapter 10), [11], and [13] (Chapter 10). More specific sample size and design issues, often focusing on two-group designs, are treated in [3], [5], [6], [7], [8], [9], and [10].

Computer programs

`ACluster` calculates required sample sizes for various types of cluster randomized designs, not only for continuous but also for binary and time-to-event outcomes, as described in [1].
See website `www.update-software.com/Acluster` .

`OD` (*Optimal Design*) calculates power and optimal sample sizes for testing treatment effects and variance components in multisite and cluster randomized trials with balanced two-group designs, and in repeated measurement designs, according to [8], [9], [10].
See website `http://www.ssicentral.com/other/hlmod.htm` .

`PinT` (*Power in Two-level designs*) calculates standard errors of regression coefficients in two-level designs, according to [12]. With extensive manual.
See website `http://stat.gamma.rug.nl/snijders/multilevel.htm` .

`RMASS2` calculates the sample size for a two-group repeated measures design, allowing for attrition, according to [3].
See website `http://tigger.uic.edu/ hedeker/works.html` .

References and further reading

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

[2] Hayes, R.J., & Bennett, S. (1999). Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology*, 28: 319–326.

[3] Hedeker, D., Gibbons, R.D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, 24: 70–93.

[4] Hox, J.J. (2002). *Multilevel Analysis, Techniques and Applications.* Mahwah, NJ: Erlbaum.

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

[6] Moerbeek, M., van Breukelen, G.J.P., & Berger, M.P.F. (2001). Optimal experimental designs for multilevel logistic models. *The Statistician*, 50: 17–30.

[7] Moerbeek, M., van Breukelen, G.J.P., & Berger, M.P.F. (2001). Optimal experimental designs for multilevel models with covariates. *Communications in Statistics, Theory and Methods*, 30: 2683–2697.

[8] Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2: 173–185.

[9] Raudenbush, S.W., & Liu, X.-F. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5: 199–213.

[10] Raudenbush, S.W., & Liu, Xiao-Feng. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, 6: 387–401.

[11] Snijders, T.A.B. (2001). Sampling. Chapter 11 (pp. 159–174) in A. Leyland & H. Goldstein (eds.), *Multilevel Modelling of Health Statistics*. Chichester etc.: Wiley.

[12] Snijders, T.A.B. & Bosker, R.J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, 18: 237–259.

[13] Snijders, T.A.B., and Bosker, R.J. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling.* London etc.: Sage.