Sampling *

Tom A.B. Snijders ICS
Department of Statistics and Measurement Theory
University of Groningen
Grote Kruisstraat 2/1
9712 TS Groningen, The Netherlands
email t.a.b.snijders@ppsw.rug.nl

February 24, 2006

1 Multilevel analysis and multistage samples

There is an intimate relation between multilevel analysis and multistage samples, although this does not go as far as the two being inseparable. Multistage samples, as described in textbooks on sampling theory (e.g., Cochran, 1977) are useful when the population sampled is divided in subsets which may be considered exchangeable and which have a role of some administrative nature. Examples are the population of inhabitants of a country divided in municipalities, or a population of patients divided in hospitals. The subsets are conventionally called *primary sampling units* or *psu*'s. In a two-stage sample, first a sample is drawn from the primary sampling units (the first-stage sample), and within each *psu* included in the first-stage sample, a sample of population elements is drawn (the second-stage sample). This can be extended to situations with more than two levels, e.g., individuals within households within municipalities, and then is called a multistage sample. In the boundary case that each sampled *psu* is included entirely in the sample, i.e., the sampling fraction in the second stage is unity, the sample is called a *cluster sample*.

Clearly, multistage samples are used for precisely those nested populations where multilevel analysis also can be appropriate. The rationale, however, may be different. The usual motive for using a multistage sample is cost efficiency: if a sample is to be drawn of 100 appendicitis patients in some year in some country, it is much cheaper to draw a two-stage sample with a first stage of, say,

*Chapter 11 (pp. 159–174) in *Multilevel Modelling of Health Statistics*, A. Leyland and H. Goldstein (eds.), Wiley, 2001.

10 hospitals, than to draw a simple random sample of 100 patients – who might be dispersed over 99 hospitals. On the other hand, the usual rationale for multilevel analysis resides in the research question at hand: the phenomena under study themselves have a multilevel structure, as is evident, e.g., when studying contextual effects in a study of outcome measures for individuals nested in organisations (hospitals, schools, etc.), or in a longitudinal study where individual development as well as individual differences are relevant.

When a multistage sample is drawn, it usually is likely that population elements within *psu*'s will be more alike than elements of different *psu*'s. Some kind of multilevel analysis therefore seems called for. On the other hand, multilevel analysis can also be applied to data collected in different sampling designs. The dependence structures represented by the random intercepts and random slopes of multilevel modeling are brought about by the processes determining the phenomena under study, with or without a multistage sampling design. It can be concluded that a multistage sample will often lead to a multilevel analysis, but multilevel analysis also can be important for other data collection designs.

2 Model-based and design-based inference

Either of two types of mechanism is usually proposed as the basis for a probability model for statistical inference. When descriptive parameters of some finite population are to be estimated from a probability sample, it is usual to base inference on the sampling design. The investigator controls the sampling process which is the foundation for this design-based inference. An important advantage is that no extraneous assumptions are required for the unbiasedness of estimators of population parameters and the associated variance estimators.

Much statistical inference is, however, not aimed at the estimation of means or other parameters of well-defined finite populations, but rather at discovering or ascertaining mechanisms and processes in our world, reflected by the observation of measurable variables. The assumed generality of such mechanisms and processes implies that the population to which the results are supposed to apply is not only quite general but also somewhat vaguely circumscribed. Findings about the course of some disease, and the effects of relevant treatments, may be generalisable to the population of all *homines sapientes* afflicted with this disease in past, present, and future – a quite hypothetical population. Results found with respect to the consequences for the course of this disease of attitudes of the patient and his or her social environment, will be culture-dependent and therefore restricted to the vaguely defined population

of patients living in a given culture – hypothetical, circumscribed in an unsatisfactory way, but meaningful nevertheless. Procedures of statistical inference for such investigations can be based on plausible probability models including assertions about the distribution of random variables and their independence or conditional independence, etc. Such models do not come for free, their plausibility must be argued and their consequences checked, and if a model does not stand such tests it must be replaced by a more plausible one. 'Random terms' or 'error terms' in such models can be regarded as resulting from influences not included among the observed variables, or – less attractive – from deviations between model and reality.

In such investigations, all or part of the sampling design often consists of just a convenience sample. In the investigation of some rare disease, the researcher will obtain collaboration from a number of clinics and include in his data all patients in these clinics suffering from this disease. The results of the investigation may be thought to apply to anyone suffering from this disease. To argue that the patients included in the study can be considered a random sample from this population, the investigator has to consider carefully the selection processes that lead to a patient being included in the study, and whether there could be factors having to do with severity of the disease, comorbidity, general health status, etc., which are related simultaneously to the selection of the patient in the study and to the measured variables. Only if it is plausible that no such variables exist, it is reasonable to apply model-based procedures of statistical inference. Often such considerations lead to circumscribing the population to which the results can be generalised, e.g., patients who have been suffering from the disease for a protracted period or those who are well-motivated to comply with their therapy.

The multilevel statistical procedures treated in this book are examples of model-based statistical inference. E.g., the usual two-level hierarchical linear model implies assumptions of independence between level-two units; conditional independence between level-one units within each level-two unit, given the random effects associated to this level-two unit; and normal distributions for the error terms. The investigator must check critically whether these assumptions are plausible. If they are, the data can be analysed as if they are produced by a two-stage sample with random selection in both stages, although it is not necessary that the sampling procedure actually was carried out in this way.

If, on the other hand, one wishes to follow a design-based approach – e.g., because the study has a descriptive purpose – and the selection probabilities are not constant, the sampling design must be taken into account to obtain unbiased estimators. For the estimation of population means this is treated in the

standard textbooks about sampling theory that include multistage sampling, e.g., Cochran (1977). For more general statistical questions such as hypothesis testing and regression analysis this is treated in specialized literature, e.g., Skinner, Holt and Smith (1989). For the estimation of parameters in the hierarchical linear model, however, it is much more complicated to take unequal selection probabilities into account. Methods to do so are proposed in Pfefferman et al. (1998). The remainder of this chapter is about model-based[1] inference only.

3 Study design: Power and standard errors

The following sections of this chapter are mainly about the design of two-level studies, in particular the determination of optimal or adequate sample sizes. What complicates the choice of an adequate design for a multilevel study is the fact that there are sample sizes to be chosen at each level of the nesting hierarchy. E.g., when studying patients in hospitals, the researcher has to decide whether a given number of hospitals and a given number of patients within each hospital is adequate. In another example, when some outcome variable is measured repeatedly for a sample of patients, it has to be decided how many patients to include in the study and how often to measure the outcome variable for each of them. Another important choice in multilevel experimental design that does not occur in single level designs is the determination of the level of randomization. E.g., when studying a new medical treatment, the researcher may have to choose between randomizing within and randomizing between between hospitals.

Since considerations for the choice of a design always are of an approximate nature, only those designs are considered here where each level-two unit contains the same number of level-one units. Level-two units will sometimes be referred to as clusters. The number of level-two units is denoted N, the number of level-one units within each level-two unit is denoted n. These numbers are called the level-two sample size and the cluster size, respectively. The total sample size is Nn. If in reality the number of level-one units fluctuates between level-two units, it will almost always be a reasonable approximation to use for n the average number of sampled level-one units per level-two unit.

Optimality or adequacy of the design is primarily a function of the power of tests and the standard errors of estimators. This chapter concentrates on parameters in the fixed part of the model for which the estimator is approximately normally distributed. Denote this parameter by β and the standard error of estimation by $s.e.(\hat{\beta})$. Provided that sample sizes are not very small,

[1] The printed text mistakenly says "design-based".

the test for β can be approximated by the standard normal test applied to the t-ratio $\hat{\beta}/s.e.(\hat{\beta})$. If the significance level is denoted by α and the power by γ, the approximate relation between standard error and power is

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

where $z_{1-\alpha}\,, z_{\gamma}$ and $z_{1-\gamma}$ are the values for which the standard normal distribution has the indicated cumulative probability values. E.g., if $\alpha = .05$ and and a power is desired of $\gamma = .80$ if the effect size is $\beta = .20$, then the standard error should be no more than

$$\text{standard error} \leq \frac{.20}{1.64 + 0.84} = 0.081 \,.$$

It should be stressed that this approximation does not take into account the degrees of freedom for variance estimation, and it might be relevant to modify the conclusions of the following analysis for relatively low values of N and n. In the following sections the discussion will be mainly in terms of standard errors.

4 The design effect for estimation of a mean

To discuss the estimation of fixed effect parameters, first three important special cases are considered: the estimation of a grand mean, the estimation of the regression coefficient of a level-two variable, and the estimation of such a coefficient of a level-one variable without any level-two variance. This should give the reader an understanding of some issues which are important for standard errors of estimators for such parameters. Then the general case will be discussed.

The estimation of a population mean is a scientific question where model-based and design-based inference meet. We approach it in a model-based way, but a design-based approach for sampling from a finite population arrives at basically the same answers, if the sample is a two-stage sample using random sampling with replacement at either stage or if the sampling fractions are so low that the difference between sampling with and sampling without replacement is negligible.

Suppose that the mean is to be estimated of some variable Y in a population which has a two-level structure. As an example, Y could be the duration of hospital stay after a certain operation under the condition that there are no complications or additional health problems. Suppose also that it is reasonable to postulate the empty model of multilevel analysis,

$$y_{ij} = \mu + u_j + e_{ij} \,,$$

with the usual assumptions. The variance of the random intercept is $\mathrm{var}(u_j) = \tau^2$, the level-one variance is $\mathrm{var}(e_{ij}) = \sigma^2$. The parameter to be estimated is μ.

The overall sample mean,

$$\hat{\mu} = \frac{1}{Nn} \sum_{j=1}^{N} \sum_{i=1}^{n} y_{ij} \,,$$

is the obvious estimator and its variance is

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

The sample mean of a simple random sample of Nn elements from this population has variance

$$\frac{\tau^2 \,+\, \sigma^2}{Nn} \,.$$

The relative efficiency of the simple random sample with respect to the two-stage sample is the ratio of these variances,

$$\frac{n\tau^2 \,+\, \sigma^2}{\tau^2 \,+\, \sigma^2} = 1 \,+\, (n-1)\rho_{\mathrm{I}} \,, \tag{2}$$

where ρ_{I} is the intraclass correlation coefficient,

$$\rho_{\mathrm{I}} = \frac{\tau^2}{\tau^2 \,+\, \sigma^2} \,.$$

The quantity (2) is called the *design effect* of the two-stage sample (e.g., Cochran, 1977). It is the ratio of the variance obtained with the two-stage sample to the variance obtained for a simple random sample with the same total sample size. A large design effect means statistical inefficiency, but this disadvantage may be offset by the cost reductions of the two-stage design. A two-stage sample yields the same standard error as a simple random sample for which the total sample size is divided by the factor (2).

5 Effect of a level-two variable

When a two-level regression is carried out and the objective is to estimate the regression coefficient of a level-two variable X, then the estimated coefficient is practically equivalent to the estimated regression coefficient in the single-level regression analysis for data aggregated to the cluster means of all relevant variables. If the variance of the random intercept is denoted again by τ^2 and the residual level-one variance by σ^2, the residual variance for the aggregated regression analysis is $\tau^2 + (\sigma^2/n)$. Assume that Y is distributed according to a random intercept model,

$$y_{ij} = \beta_0 \,+\, \beta_1 x_j \,+\, u_j \,+\, e_{ij} \,.$$

When the variable X has variance s_X^2, the variance of the estimated regression coefficient is

$$\mathrm{var}(\hat{\beta}_1) = \frac{n\tau^2 + \sigma^2}{Nns_X^2} \,. \tag{3}$$

This implies that again the relative efficiency of the two-stage sampling design is given by (2).

Thus it appears that for estimating a population mean or, more generally, the effect of a level-two variable, and if the intraclass correlation is moderate or high, a large cluster size leads to a large statistical inefficiency in estimating the population mean.

6 Effect of a level-one variable

Now consider the opposite situation, where one wishes to estimate the regression coefficient of an independent variable X which is a pure level-one variable, i.e., its mean is the same in each level-two unit. Note that this implies that the intraclass correlation of X is negative, $-1/(n-1)$, which implies that X itself is not distributed according to the hierarchical linear model (which allows only nonnegative intraclass correlations). For simplicity, assume that the cluster mean of X is 0 and its variance is the same within each cluster, denoted by s_X^2. An example is the effect of a treatment that is randomly allocated to a fixed fraction of the level-one units within each level-two unit. Another example is the linear effect of time in a balanced longitudinal design.

The estimator for the regression coefficient now is the average of the within-cluster regression coefficients,

$$\hat{\beta} = \frac{1}{Nns_X^2} \sum_{j=1}^{N} \sum_{i=1}^{n} x_{ij}\, y_{ij}\,. \tag{4}$$

If Y is distributed according to a random intercept model,

$$y_{ij} = \beta_0 \,+\, \beta_1 x_{ij} \,+\, u_j \,+\, e_{ij}\,,$$

then the assumptions about the variable X imply that the estimator (4) is equal to

$$\hat{\beta}_1 = \beta_1 \,+\, \frac{1}{Nns_X^2} \sum_{j=1}^{N} \sum_{i=1}^{n} x_{ij}\, e_{ij} \tag{5}$$

and its variance is

$$\mathrm{var}(\hat{\beta}_1) = \frac{\sigma^2}{Nns_X^2}\,. \tag{6}$$

A simple random sample of Nn elements from the same distribution yields an OLS estimator for β with variance

$$\frac{\sigma^2 + \tau^2}{Nns_X^2}$$

(assuming that the variance of X in this sample also is precisely s_X^2). This shows that the two-stage sample here is more efficient than the simple random

sample, with the design effect

$$\frac{\sigma^2}{\tau^2 + \sigma^2} = 1 - \rho_{\text{I}} \,. \tag{7}$$

The greater efficiency in this case of the two-stage sample is well known in experimental design: the two-stage corresponds to blocking on the level-two units. In psychology, e.g., this is the often used within-subject design. Blocking is known to neutralize the main block effect as a variance component.

Since the design effect is less than 1 for level-one variables and larger than 1 for level-two variables, it may be concluded that if the study is a comparison of randomly assigned treatments in a random intercept model and the study costs are determined by the total sample size, Nn, then randomising within clusters is more efficient than randomising between clusters. The optimal level of randomization for two- and three-level designs is discussed extensively by Moerbeek, van Breukelen, and Berger (2000).

Example

Suppose that a new training program ('treatment') for nurses is to be compared with an existing training program ('control'), while hospitals are believed to be a major influence for the nurses' work. This question can be phrased in terms of the preceding sections as the estimation of the regression coefficient of the dummy variable that distinguishes treatment from control. Denote this variable by X, defined as 0 for the control and 1 for the treatment condition. When the treatment fraction is p, its variance is $\text{var}(X) = p(1-p)$; for $p = 0.5$ this yields $s_X^2 = 0.25$. Assume that the dependent variable is standardized to have a unit variance, and that it has an intraclass correlation of 0.10. Further assume that the treatment is equally effective for all hospitals, i.e., the random intercept model is adequate. Then $\sigma^2 = 0.9$ and $\tau_0^2 = 0.1$. Then the estimation variance for level-two randomization (3) is

$$\frac{0.4}{N} + \frac{3.6}{nN}$$

and for level-one randomization it is (6),

$$\frac{3.6}{nN} \,.$$

If group sizes n are predetermined, the advantage of randomization at level one is quite large.

6.1 A level-one variable with a random slope

The level-one variable X, however, may well have a random slope in addition to the random intercept. The model for Y then reads

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j}\, x_{ij} + e_{ij} \,. \tag{8}$$

Denote the intercept variance by τ_0^2 and the slope variance by τ_1^2. For the random slope model, the variance of the estimated regression coefficient (4) is

$$\mathrm{var}(\hat{\beta}_1) = \frac{n\tau_1^2 s_X^2 + \sigma^2}{Nns_X^2}\,. \tag{9}$$

The total residual variance of Y (where the level-two unit is random, i.e., marginalized) here is equal to

$$\sigma^2 \,+\, \tau_0^2 \,+\, \tau_1^2 s_X^2$$

so that the design effect now is

$$\frac{n\tau_1^2 s_X^2 + \sigma^2}{\tau_0^2 + \tau_1^2 s_X^2 + \sigma^2}\,. \tag{10}$$

This shows that the two-stage sample with level-one randomization only 'neutralizes' the random intercept and not the random slope of X as terms in the variance of the estimated regression coefficient.

In practice, the presence of a random slope for variable X means that the regression coefficient β_1 does not tell all of the story, and it is important to estimate the random slope variance τ_1^2 besides. This underscores the fact that design considerations should never focus narrowly on the estimation of just one statistical parameter.

7 Optimal sample size for estimating a regression coefficient

In studies leading to statistical models such as those treated in the preceding sections, the design is determined by the sample sizes N and n and the distribution of the X values. This distribution has a within-cluster and a between-cluster aspect. Like in OLS regression, if one has liberty to choose the X values, it is optimal to maximize their dispersion. With respect to optimal sample sizes, the multilevel, or two-stage design, requires the determination of the sample sizes at the two levels. This section is about the optimal choice of these sample sizes for the estimation of a regression coefficient under given budget constraints.

If the aim is to have a minimum variance for a given total sample size Nn and a given value for s_X^2, then it is clear from (6) that for a within-cluster deviation variable without level-two slope variation it does not matter how the total sample size is distributed over the level-two units, as long as one succeeds in constructing an X variable with constant within-cluster mean and within-cluster variance s_X^2. This implies, of course, that n is at least 2. For a level-two variable or a within-cluster deviation variable with positive random slope variation, (3) and (9) imply that it is optimal to let N be as large as

possible. This would imply $n = 1$, i.e., a simple random sample is optimal. If this is not feasible, then still it is best to have the clusters as small as possible.

Usually, however, study costs are not a function of total sample size but depend on total sample size as well as the number of level-two units. The costs often are well approximated by a function of the type $c_1N + c_2Nn$. Thus, an optimal design is obtained when the variance of the estimator is minimal, given the constraint

$$c_1N + c_2Nn \leq k\,, \tag{11}$$

where k denotes the total budget. In the preceding sections, it was shown that the variance to be minimised can be expressed as

$$\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{Nn}$$

for a suitable choice of σ_1^2 and σ_2^2, which can be found in equation (3), (6), or (9), respectively. The minimisation of this expression under the constraint $c_1N + c_2Nn = k$ is treated by Cochran (1977, Section 10.6). The optimal value for n is

$$n_{\text{opt}} = \sqrt{\frac{c_1\sigma_2^2}{c_2\sigma_1^2}}\,, \tag{12}$$

rounded upward or downward to an integer value. This is also the optimal n if the budget is to be minimised under the constraint that the estimation variance has a preassigned value. It may be noted that the optimal cluster size does not depend on the available budget or on the level-two sample size.

For explanatory variables defined at level one and having a constant mean across level-two units, and for which the dependent variables follows a random intercept model, it can be seen from (9) that $\sigma_1^2 = 0$ so that n_{opt} is infinite. This means in practice that n should be as large as possible: a single-level design, for which $N = 1$ and n is the total sample size, is preferable to a two-level design for the estimation of a regression coefficient of a level-one variable when the budget constraint is given by (11). If the explanatory variable X is defined at level two, on the other hand, we have $\sigma_1^2 = \tau^2/s_X^2$ and $\sigma_2^2 = \sigma^2/s_X^2$ so that the optimal cluster size is

$$n_{\text{opt}} = \sqrt{\frac{c_1\sigma^2}{c_2\tau^2}}\,.$$

This optimal sample size also is discussed by Raudenbush (1997, p. 177) and Moerbeek, van Breukelen, and Berger (2000). The latter paper also treats optimal allocation for three-level designs. Optimal allocation for two- and three-level designs for binary responses are discussed by Moerbeek, van Breukelen, and Berger (submitted).

8 Use of covariates

It is well-known in experimental design that controlling for relevant covariates can lead to important gains in efficiency. In a single-level design, a covariate that has a residual correlation with the dependent variable equal to ρ will yield a reduction of the unexplained variance by a factor $1 - \rho^2$. When the sample size is large enough for the loss of a degree of freedom for the variance estimate to be unimportant, this will allow the researcher to diminish the sample size by this factor while retaining the same standard error and power.

For a two-level design, the situation is – of course – more complicated. The reduction in standard error depends on the intraclass correlation of the dependent variable and on the within-group and the between-groups residual correlations between the dependent variable and the covariate. For more precise calculations in small sample situations, the degrees of freedom also play a part, but this is not considered in the following analysis.

Suppose that, as above, we wish to analyze the regression coefficient of some variable X, and we are interested to see how much gain in precision is obtained by controlling for some covariate denoted by Z. The model without control for Z is supposed to be the random intercept model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

while the model with control for Z is

$$y_{ij} = \tilde{\beta}_0 + \beta_1 x_{ij} + \beta_2 z_{ij} + \tilde{u}_j + \tilde{e}_{ij}\,.$$

It is assumed that Z follows a random intercept model,

$$z_{ij} = \gamma_0 + u_{Zj} + e_{Zij}\,.$$

Further assume that Z is, both within and between groups, uncorrelated with X. Then the regression coefficient β_1 remains the same when controlling for Z, which is reflected in the notation used in the two preceding equations.

Denote the population residual within-group correlation between Y and Z by ρ_W and their population residual between-group correlation by ρ_B. These are defined by (cf. Snijders and Bosker, 1999, Section 3.6)

$$\rho_W = \rho(e_{ij},\, e_{Zij}),\ \rho_B = \rho(u_j,\, u_{Zj})\,.$$

Some calculations show that the reduction in the variance parameters due to the control for Z is given by

$$\tilde{\sigma}^2 = (1 - \rho_W^2)\,\sigma^2,\ \tilde{\tau}^2 = (1 - \rho_B^2)\,\tau^2\,. \tag{13}$$

In a large-sample approximation (valid when n and N are large), these reductions are applied to the estimation variances given in (3) and (6). Thus, if X is a level-two variable, then formula (3) applies and the factor $\tau^2 + (\sigma^2/n)$ will be replaced by $(1 - \rho_B^2)\tau^2 + (1 - \rho_W^2)(\sigma^2/n)$. If X is a level-one variable, then

in (6) the factor σ^2 is replaced by $(1 - \rho_W^2)\,\sigma^2$. This illustrates that for pure level-one variables it is – naturally – only the within-group correlation which counts, whereas for level-two variables not only the between-group correlation but also the within-group correlation plays a role in the reduction of the estimation variance of β_1. If group sizes n are large, however, the influence of the within-group correlation for level-two variables will be of minor importance.

When one investigates the effect of a level-two variable X controlling for a level-one variable Z, one may be tempted to use the group means of Z rather than their individual values. The preceding analysis demonstrates that this leads to a loss in estimation efficiency. If n is large the loss will be negligible. This point is made also by Raudenbush (1997), who gives a more extensive discussion of the use of covariates, taking into account also the random nature of the observed residual covariances between Z and Y (but not the loss in degrees of freedom).

9 Standard errors for fixed effects in general

In practice, the assumptions made in Section 4 and 5 often are not an adequate simplification of reality; moreover, many researcher wish to estimate several regression coefficients from a single data set as precisely as possible. Exact formula for estimation variances are not available for arbitrary multilevel designs. Snijders and Bosker (1993) derived approximate formulae for estimation variances in two-level designs, valid under the restriction that variables with random slopes have a zero between-cluster variance and that n is not too small, say, at least 8. These formulae are calculated by the computer program *PinT* ('Power in Two-level designs') which can be dowloaded, with manual, from http://stat.gamma.rug.nl/snijders/multilevel.htm .

The main difficulty in applying this program is the requirement to specify plausible parameter values. This is, of course, a general difficulty in any power analysis (cf. Kraemer and Thiemann, 1987, or Cohen, 1992), but it is more pressing in the case of multilevel analysis because the random part parameters also must be specified. The use of *PinT* will be illustrated here by means of an example. The manual of *PinT* and Chapter 10 of Snijders and Bosker (1999) contain various other examples.

As an example, suppose that one is investigating the effect of the training of psychotherapists on therapy effectivity. Level-one units are patients, level-two units are therapists. The dependent variable is a patient-level outcome measure standardized to unit variance. The investigated training is a course represented by a 0-1 variable. In addition, the level of professional training of

the psychotherapists and a pretest measure of the seriousness of the patients' complaints are relevant. The question is, how many patients and how many therapists have to be investigated.

Assume that participation in the course will be randomized within groups of equal professional training in such a way that the fraction following the course is higher in the groups with lower professional training. Professional training and pretest are represented by numerical variables also standardized to mean 0 and unit variance. If a random intercept model applies with these three variables having fixed effects, the model can be expressed as

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3ij} + u_j + e_{ij}\,,$$

where X_1 indicates whether the therapist followed the course, X_2 is therapists' professional training, and X_3 is the pretest. The primary research variable is X_1.

The required information for running *PinT* consists of the means, variances, and covariances of the explanatory variables and all parameters of the random part.

First consider the means. Variables X_2 and X_3 have means 0. Suppose that the fraction of therapists who follow the course is thought to be 0.4. This is the mean of X_1.

Now consider the variances and covariances of the explanatory variables. Suppose that the therapists are somewhat different in the pretest values of their patients, this variable having an intraclass correlation of 0.19. The variance of X_1, being a 0-1 variable with mean 0.4, is 0.24. Suppose that the correlation between the therapist mean of the pretest and the professional training X_2 is known to be 0.5. Assume that the randomization of the course participation, which is conditional on X_2, will give a correlation between X_1 and X_2 of -0.4. The partial correlation between pretest mean and course participation, controlling for level of professional training, is expected to be nil, which leads to a total correlation between pretest mean and course participation of 0.2. The within-groups variance of X_3 then is $\sigma^2_{X(W)} = 1 - 0.19 = 0.81$ and the between-groups covariance matrix of (X_1, X_2, X_3) is

$$\Sigma_{X(B)} = \begin{pmatrix} 0.24 & -0.20 & 0.043 \\ -0.20 & 1.0 & 0.22 \\ 0.043 & 0.22 & 0.19 \end{pmatrix}\,.$$

Finally consider the parameters of the random part of the multilevel model. To get some insight in plausible values of the level-one and level-two variances it may be helpful to note that the variance of the dependent variable can be

decomposed as

$$\text{var}(Y_{ij}) = \beta_1^2 \sigma^2_{X(W)} + \beta' \Sigma_{X(B)} \beta + \tau_0^2 + \sigma^2 \ ,$$

where $\beta = (\beta_1, \beta_2, \beta_3)'$ (cf. Section 7.2 of Snijders and Bosker, 1999). This corresponds, for the variance decomposition of Y_{ij} in the empty model, to a total level-one variance of $\beta_1^2 \sigma^2_{X(W)} + \sigma^2$ and a total level-two variance of $\beta' \Sigma_{X(B)} \beta + \tau_0^2$.

Assume that the total level-one and level-two variances of the outcome measure are 0.8 and 0.2, respectively, and that the available explanatory variables together explain 0.25 of the level-one variance and 0.5 of the level-two variance. Then $\sigma^2 = 0.6$ and $\tau_0^2 = 0.10$. In terms of the decomposition of total variance this corresponds to a raw explained level-one variance of $\beta_1^2 \sigma^2_{X(W)} = 0.2$, and therefore a regression coefficient $\beta_1 = \sqrt{0.2/0.81} = 0.5$, and $\beta' \Sigma_{X(B)} \beta = 0.1$.

With respect to the cost structure, assume that the budget constraint can be expressed as (11) with $c_1 = 20$, $c_2 = 1$, and $k = 1000$. In other words, an extra therapist in the sample costs 20 times as much as an extra patient; and there would be enough funds to include, e.g., 40 therapists with 5 patients each, or 20 therapists with 30 patients each.

With this specification, *PinT* can be executed, and it produces the approximate standard errors of the estimated fixed effects for sample sizes satisfying the budget constraint $20N + Nn \leq 1{,}000$. The standard errors for β_1 are plotted as $*$ in Figure 1.

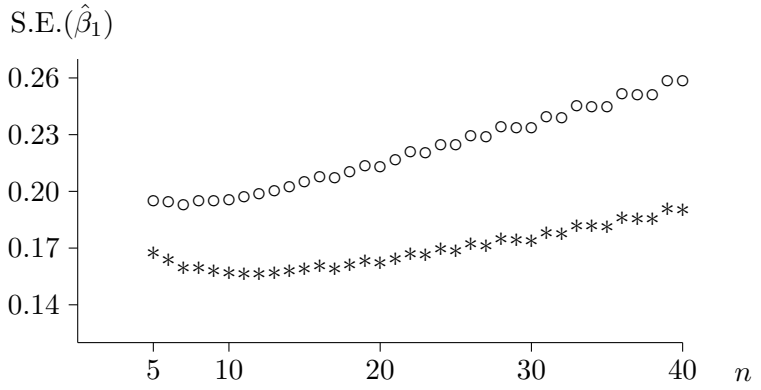

Figure 1 Standard error for estimating β_1, for $20N + Nn \leq 1{,}000$; $*$ for $\sigma^2 = 0.6$, $\tau_0^2 = 0.1$ and $\circ$ for $\sigma^2 = 0.5$, $\tau_0^2 = 0.2$.

The plot is a bit irregular due to the inequality constraint for the integer numbers N and n. The minimum is seen to be rather flat. The minimum standard error is 0.156, achieved for $n = 11$. For cluster sizes between 7 and 18 the standard error is less than 0.162. It can be concluded that, for these parameter values, average cluster sizes between 7 and 18 are fully acceptable.

To investigate the sensitivity of this result to the assumed parameter values, the calculations were carried out also for $\sigma^2 = 0.5$, $\tau_0^2 = 0.2$, all other parameters remaining the same. The results are shown in Figure 1 by the symbol $\circ$. The greater level-two variance leads to a bigger standard error for this level-two variable. The minimum is 0.192 for $n = 7$. The minimum is less flat than for the earlier parameter values; for $n \leq 13$, the standard error is less than 0.200. For these parameter values, the average cluster sizes would preferably be 13 or less. Note, however, that in the second situation the intraclass correlation is twice as big as in the first one, so the two situations are quite different.

The *PinT* program uses a rather rough large-sample approximation to obtain the standard errors. This is often adequate, because design questions usually are of a very approximate nature, but more precise approximations are desiable when they are available. For the special case of testing the effect of a level-two variable (representing the difference between a treatment and a control condition in a cluster randomized trial), controlling for one level-one covariate, a more precise approximation is given by Raudenbush (1997, p. 178–179). He obtains a result which for larger sample sizes boils down to (12). The greater precision can be important for small sample sizes.

10 Parameters of the random part

Usually the focus of the research questions is on the parameters of the fixed part. Sometimes, however, the design should be adequate also in view of the estimation of the random part parameters. The estimation of the intraclass correlation coefficient is treated here. For some remarks about the design of multilevel studies with respect to the estimation of other parameters of the random part, see Mok (1995), Cohen (1998), and Snijders and Bosker (1999, Section 10.5.2).

Donner (1986) proved that the standard error of the estimated intraclass correlation coefficient in an empty two-level model (i.e., a two-level model without any explanatory variables) with constant cluster size n is given by

$$\mathrm{S.E.}(\hat{\rho}_{\mathrm{I}}) = (1 - \rho_{\mathrm{I}})(1 + (n-1)\rho_{\mathrm{I}})\sqrt{\frac{2}{n(n-1)(N-1)}}\,. \tag{14}$$

This standard error depends on the parameter itself that is to be estimated. To obtain optimal sample sizes for estimating the intraclass correlation given the budget constraint (11), it is convenient to substitute $N = k/(c_1 + c_2 n)$ which transforms (14) into a function of n so that it can be plotted. From the graph, the optimum value for n can be deduced, as well as the sensitivity of this minimum to sub-optimal values of n.

As an example, suppose that it is desired to estimate the intraclass correlation with a budget constraint $20N + Nn \leq 1{,}000$ and the intraclass correlation is believed to be between 0.1 and 0.2. Figure 2 gives the graph of the standard errors of the intraclass correlation coefficients, using the substitution $N = 1{,}000/(20 + n)$ (neglecting the integer nature of the sample sizes), for $\rho_{\mathrm{I}} = 0.1$ and 0.2.

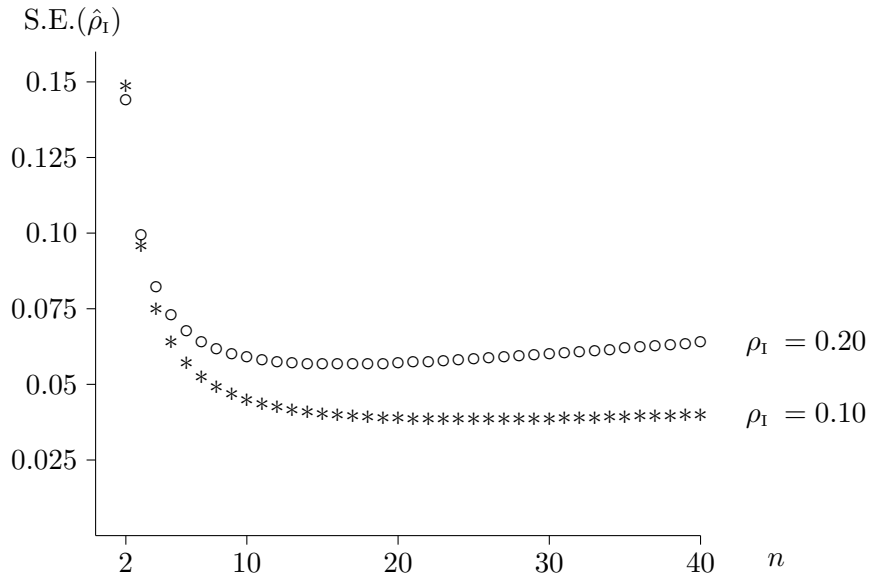

Figure 2 Standard error for estimating the intraclass correlation coefficient for a budget constraint $20N + Nn \leq 1{,}000$ with $\rho_{\mathrm{I}} = 0.1$ and 0.2.

For $\rho_{\mathrm{I}} = 0.1$ the minimum standard error is 0.03835, achieved for $n = 24$, 25, and the standard error is less than 0.040 for n between 16 and 40. For $\rho_{\mathrm{I}} = 0.2$ the minimum standard error is 0.05645, achieved for $n = 16$, the standard error being less than 0.59 for n between 10 and 27. In order to have a relatively small standard error for ρ_{I} in the range between 0.1 and 0.2, cluster sizes between 16 and 27 are fully acceptable.

References

Cochran, W.G. (1977) *Sampling Techniques*, 3d edn. New York: Wiley.

Cohen, J. (1992) 'A power primer'. *Psychological Bulletin*, 112, 155–159.

Cohen, M. (1998) 'Determining sample sizes for surveys with data analyzed by hierarchical linear models'. *Journal of Official Statistics*, 14, 267–275.

Kraemer, H.C., and Thiemann, S. (1987), *How many subjects? Statistical power analysis in research.* London, etc.: Sage.

Moerbeek, M., van Breukelen, G.J.P., and Berger, M.P.F. (2000) 'Design issues for multilevel experiments'. *Journal of Educational and Behavioral Statistics*, in press.

Moerbeek, M., van Breukelen, G.J.P., and Berger, M.P.F. (submitted). 'On the design of experiments with binary responses in multilevel populations'.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research, *Multilevel Modeling Newsletter*, 7 (2), 11–15. Available from http://www.ioe.ac.uk/users/hgoldstn/workpap.html .

Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998) 'Weighting for unequal selection probabilities in multilevel models'. *Journal of the Royal Statistical Society*, **B** 60, 23–40.

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

Snijders, T.A.B. and Bosker, R.J. (1993) 'Standard errors and sample sizes for two-level research'. *Journal of Educational Statistics*, 18, 237–259.

Skinner, C.J., Holt, D., and Smith, T.M.F. (eds.) (1989), *Analysis of Complex Surveys*. New York: Wiley.

Snijders, T.A.B., and Bosker, R.J. (1999) *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London, etc.: Sage.