Journal of Official Statistics, Vol. 14, No. 3, 1998, pp. 267–275

Determining Sample Sizes for Surveys with Data Analyzed by Hierarchical Linear Models

Michael P. Cohen[1]

Behavioral and social data commonly have a nested structure (for example, students nested within schools). Recently techniques and computer programs have become available for dealing with such data, permitting the formulation of explicit hierarchical linear models with hypotheses about effects occurring at each level and across levels. An example is given where such models could be used. If data users are planning to analyze survey data using hierarchical linear models rather than concentrating on means, totals, and proportions, this needs to be accounted for in the survey design. The implications for determining sample sizes (for example, the number of schools in the sample and the number of students sampled within each school) are explored.

Key words: Multilevel; random effects; covariance components.

1. Introduction and Example

There has been an upsurge in interest among behavioral and social scientists and education researchers in analyzing data in a way that accounts for the naturally occurring nested structure, for instance, in analyzing students nested within schools. Linear models appropriate for such data are called *hierarchical* or *multilevel*. In part, the increased interest has been sparked by the availability of new software that properly handles the nested structure and facilitates the analyses. There has also been a realization that one can take advantage of the nested structure to explore relationships not amenable to other approaches.

Bryk and Raudenbush (1992), Goldstein (1987, 1995), and Longford (1993) are recommended for book-length discussions related to hierarchical linear models.

To illustrate these models, an example of Bryk and Raudenbush (1992, Chapter 4) will be summarized. This example is based on data from a subsample of the 1982 High School and Beyond Survey, a survey of high school students in the United States by the U.S. National Center for Education Statistics. The socioeconomic status (SES) of the student is a variable computed from the income, education, and occupation of the student's parents. The MEAN SES is the average over the students in the school of the SES values for the students. The following questions, quoted from Bryk and Raudenbush (1992, p. 61), were being explored:

1. How much do U.S. high schools vary in their mean mathematics (math) achievement?

[1] Mathematical Statistician, National Center for Education Statistics, 555 New Jersey Avenue NW, Washington, DC 20208-5654, U.S.A.
Acknowledgments: The author thanks the referees for helpful comments.

2. Do schools with high MEAN SES also have high math achievement?
3. Is the strength of association between student SES and math achievement similar across schools? Or is SES a more important predictor of achievement in some schools than others?
4. How do public and Catholic schools compare in terms of mean math achievement and in terms of the strength of the SES-math achievement relationship, after we control for MEAN SES?

These are the kinds of questions that hierarchical linear models (HLMs) can handle.

One student-level model for this example is

$$y_{ij} = \beta_{0j}^* + r_{ij} \tag{1.1}$$

with the school-level model

$$\beta_{0j}^* = \gamma_0 + u_{0j} \tag{1.2}$$

where the asterisk on β_{0j}^* indicates that the parameter is random, not fixed. The random parameter β_{0j}^* might be, for instance, the mean mathematics achievement of school j. The r_{ij} are mean zero, independent, normally distributed random variables, each with variance σ^2, for the $i = 1, \ldots, n_j$ students in school j. The u_{0j} are independent of each other and of the r_{ij}. They are normally distributed, each with mean zero and variance τ^2. The τ^2 are called the *student-level variances*, and the τ^2 are called the *school-level variances*.

We shall consider models a bit more general. The student-level model will be

$$y_{ij} = \beta_{0j}^* + \sum_{h=1}^{p} \gamma_h x_{hij} + r_{ij} \tag{1.3}$$

with the school-level model still given by (1.2) and with the same distributional assumptions as above on the r_{ij} and u_{0j}. These are called *random intercept* models. The x_{hij} are independent variables, and the γ_h are fixed unknown parameters. For instance, x_{1ij} might be student SES in the example with $p = 1$. We consider (1.1) to be a special case of (1.3) with $p = 0$.

There are HLMs not fitting the above setup. In (1.3), if we replace γ_1 by β_{1j}^*, where the β_{1j}^* satisfy an equation like (1.2), we would have a particular case of what is called a *random slope* model. One can consider, moreover, school-level models more complicated than (1.2). These, and other more general models, are of practical interest, but they are beyond the scope of this article. We shall also restrict attention to the balanced case $n_j = n$; that is, we shall assume the same number n of students are selected per school.

By substituting (1.2) into (1.3), we get the models considered in this article in their *combined form*:

$$y_{ij} = \gamma_0 + \sum_{h=1}^{p} \gamma_h x_{hij} + u_{0j} + r_{ij} \text{ for } i = 1, \ldots, n. \tag{1.4}$$

Having the full model in one equation is technically convenient, although the separate equations (1.3) and (1.2) are often easier to interpret.

In the next section we discuss some recent research related to ours. In Section 3, the sample design and cost function are described. We review traditional sample size

determination from the viewpoint of the survey sampler in Section 4. In Section 5, the analytical results on sample sizes are developed. A final comment is provided in the last section.

2. Some Recent Related Work

There is a large and growing literature on hierarchical linear models. The bulk of this literature emphasizes estimation and interpretation rather than sample design questions. There are three recent papers, though, that are particularly pertinent so they will be summarized in this section.

The work of Snijders and Bosker (1993) is the most similar to this article, especially to our Subsection 5.2. They used asymptotic approximations, supported by simulations, to get formulas for the covariance matrix of the estimators of the regression coefficients. They showed how to use these formulas to derive approximately optimal sample sizes by searching among possible within-school sample sizes, holding costs constant. Their cost function is equivalent to the one used in this article.

Afshartous (1995) performed an interesting empirical study based on subsampling schools (and hence, indirectly, students) from the base year data of the National Educational Longitudinal Study of 1988, a survey of U.S. eighth graders by the National Center for Education Statistics. He was interested in determining the minimum number of schools one can have in the sample and still get ''good'' (according to various criteria, e.g., unbiasedness, stability) estimates. He found that for estimates of variance components, 320 schools are needed whereas to estimate regression coefficients as few as 40 schools may suffice.

Mok (1995), in a very thorough study, investigated samples of students of a fixed size as the number of schools and number of students per school vary. Like Afshartous, Mok derived her samples from a real educational dataset, using data on a population of students at 50 New South Wales Catholic schools collected by M. Flynn. She considered a wide variety of estimators, including regression coefficients, variances, and covariances. She found that designs using more schools and fewer students per school are generally less biased and more efficient than ones with fewer schools and more students per school, holding the total sample size constant.

The constraint considered by Mok, a fixed number of students, is equivalent to the special case $C_s = 0$ and, say, $C_k = 1$ in the cost function that will be introduced in the next section and employed thereafter to evaluate sample designs.

The empirical evaluations of Afshartous and Mok are complementary to the analytical approach adopted in this article.

3. Simple Two-Stage Design with a Simple Cost Function

In order to gain insight into the problem, we restrict our attention to a simple two-stage sampling design with a simple cost function. We select m schools, and from each of the m schools, we select n students (a balanced sample design). It costs C_s to include a school in the sample and an additional C_k for each student (''kid'') sampled at the school. We wish to hold total sampling costs to our budgeted amount C where

$$C = C_s m + C_k mn$$

We refer to the *first stage units* as *schools* and the *second stage units* as *students* throughout this article in order to avoid cumbersome terminology. Of course, the results apply much more broadly (for example, to beds within hospitals or to books within libraries).

In reality we would almost certainly select the schools by a stratified design. Additional levels (e.g., school districts, classrooms) are possible. Unequal probability sampling might be used at any level. Our assumption of a balanced sample design (same number of students from each school) would almost certainly not hold exactly, but we do not expect that our results are very sensitive to this assumption, provided that the design is not too unbalanced.

4. Traditional Sample Size Determination

Hansen, Hurwitz, and Madow (1953, pp. 172–173) have developed the formula for the optimal size n for the number of students to sample from each school. It applies to estimating means, totals, and ratios and minimizes the sampling variance of the estimator for a fixed total cost. A simple approximate version of the formula is as follows:

$$n_{\text{opt}} \doteq \sqrt{\frac{C_s}{C_k} \times \frac{1-\rho}{\rho}} \tag{4.1}$$

where ρ is the measure of homogeneity, also called the intraclass (*intra-school* in our example) correlation coefficient. The number of schools sampled is then

$$m_{\text{opt}} = \frac{C}{C_s + C_k n_{\text{opt}}}$$

Under the HLM model, we have

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

where σ^2 is the student-level variance and τ^2 is the school-level variance. It will also be convenient to work with the *variance ratio* ω defined by $\omega = \tau^2/\sigma^2$. In terms of the variance ratio, (4.1) becomes

$$n_{\text{opt}} \doteq \sqrt{\frac{C_s}{C_k} \times \frac{1}{\omega}} \tag{4.2}$$

so that the optimal number of students to sample from each school in the traditional setting varies inversely with the square root of the variance ratio ω.

It is perhaps worth mentioning that we are interested in finding the optimal values of n and m, not with the notion that they should be adhered to exactly, but rather with the idea that they can serve as a guide in survey planning.

5. Sample Size Determination for Hierarchical Linear Modelling

In analyzing HLM models, it is important to be able to estimate not only the regression coefficients but also the school-level and student-level variances (τ^2 and σ^2) because these

quantities are of substantive interest. In this section, we first explore the sample size implications of needing to estimate τ^2 and σ^2. We then study, for a simple special case, the corresponding problem for the regression coefficients.

5.1. The student-level and school-level variances

Longford (1993, p. 58) shows that the maximum likelihood estimates of τ^2 and σ^2 have asymptotic variances

$$\text{var}(\hat{\sigma}^2) = \frac{2\sigma^4}{mn - m} \tag{5.1}$$

and

$$\text{var}(\hat{\tau}^2) = \frac{2\sigma^4}{mn}\left(\frac{1}{n-1} + 2\omega + n\omega^2\right) \tag{5.2}$$

as the number of schools m grows large. As before, $\omega = \tau^2/\sigma^2$ denotes the variance ratio. We aim to minimize these variances subject to the cost constraint of Section 3: $C = C_s m + C_k mn$ where C is the total allowable cost, C_s is the cost of sampling each school, and C_k is the additional cost of sampling each student. But then $m = C/(C_s + C_k n)$ so that m can be eliminated from the Equations (5.1) and (5.2).

For fixed values of C, C_s, C_k, σ^2, and ω (the latter two would have to be estimated from previous data), it is relatively easy to find the values of n and m that minimize $\text{var}(\hat{\sigma}^2)$ or $\text{var}(\hat{\tau}^2)$ with $m = C/(C_s + C_k n)$. We merely evaluate the variance equations for all reasonable values of n. This can be done very quickly on a computer. But the result does not convey an understanding of how the sample should be apportioned as the different parameters vary. We therefore seek analytical solutions.

Let us consider $\text{var}(\hat{\sigma}^2)$ first. Although (5.1) is minimized subject to the cost constraint by taking n (students per school) as large as possible, in fact, $\text{var}(\hat{\sigma}^2)$ is relatively flat even for moderate n. It is (5.2), again subject to the cost constraint, that is the critical one to minimize.

The expression for minimizing $\text{var}(\hat{\tau}^2)$ with $m = C/(C_s + C_k n)$ reduces to solving a fourth degree polynomial in n. We have obtained the solution, but the expression is too cumbersome to be of any practical use. We can, however, study the closely related expression

$$\text{var}(\hat{\tau}^2) \doteq \frac{2\sigma^4}{mn}\left(\frac{1}{n} + 2\omega + n\omega^2\right) \tag{5.3}$$

where we have replaced $n - 1$ by n in the denominator of the first term. We have made informal numerical comparisons of (5.2) and (5.3) and found, in our experience, that the best values of n are usually the same and, if not, almost always within one for the two equations. Figure 1 graphs both curves for the example in which $\sigma^2 = 1$, $\omega = .2$, $C = 10,000$, $C_s = 30$, and $C_k = 1$. Figure 2 treats the same example except that ω varies from .04 to 2 in increments of .04. This figure shows the percentage increase in (5.2) if the value of n that minimizes (5.3) is used instead of the one that minimizes (5.2). The percentage increases for this example are frequently 0 and always small. The only examples of large percentage increases that the author has observed occur when the n that minimizes

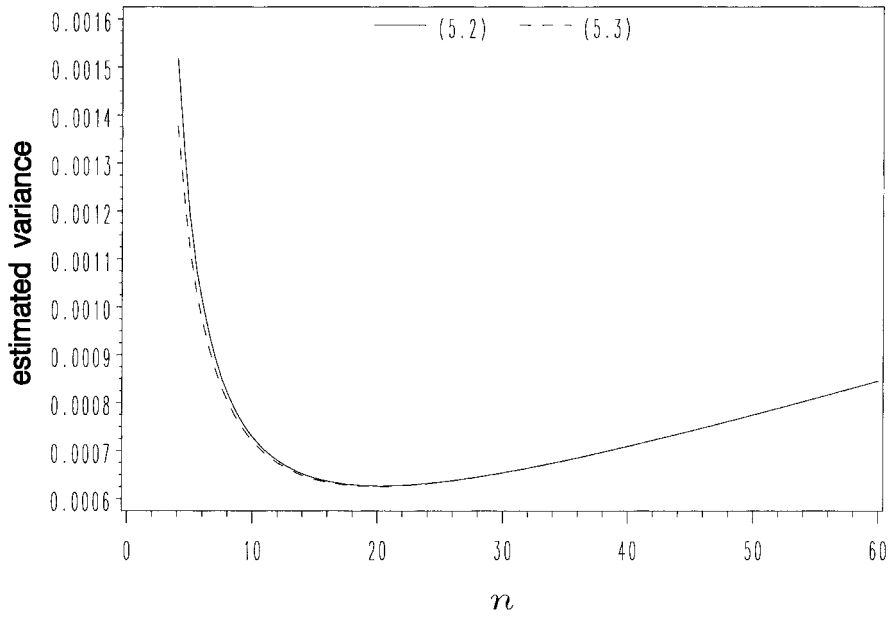

Fig. 1. Comparison of (5.2) and (5.3)

(5.3) is 4 or less. When n is large, (5.2) and (5.3) must be close. It turns out, by the way, that (5.3) is the correct asymptotic expression for $\text{var}(\hat{\tau}^2)$ when σ^2 is *known* (cf. Longford 1993, p. 59).

The solution to (5.3), subject to $C = C_s m + C_k mn$, is

$$n_{\text{opt}} = \frac{\sqrt{C_k(C_k + 8C_s\omega)} + C_k}{2C_k\omega}$$
$$= \frac{1}{2\omega} + \sqrt{2\frac{C_s}{C_k}\times\frac{1}{\omega} + \frac{1}{4\omega^2}} \tag{5.4}$$

In particular, comparing (5.4) with the traditional case of (4.2), we see from the first term under the square root sign in (5.4) that n_{opt} will be at least $\sqrt{2}$ times as large as it is for (4.2). If $1/\omega$ is large relative to C_s/C_k, the difference is even more marked. Suppose, say, $C_s/C_k = 30$. Then for $\omega = 1$, $n_{\text{opt}} = 5$ in (4.2) and $n_{\text{opt}} = 8$ in (5.4). For $\omega = 0.05$, $n_{\text{opt}} = 24$ in (4.2) and $n_{\text{opt}} = 46$ in (5.4). So estimation of τ^2 requires a larger sample of students within each school (and hence fewer schools) for a fixed cost than does estimation of traditional quantities (means, totals, ratios).

5.2. *The regression coefficients*

It is also, of course, important to be able to estimate the regression coefficients themselves. We denote the vector of regression coefficients by $\boldsymbol{\gamma}$, the design matrix by $\mathbf{X}$, and the vector of outcomes by $\mathbf{y}$. To illustrate the notation, consider the case where we have two students selected from each of a sample of two schools. If $p = 2$ in (1.4), that is, if there are two independent variables, then $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$ where $^\top$ denotes transpose. In this

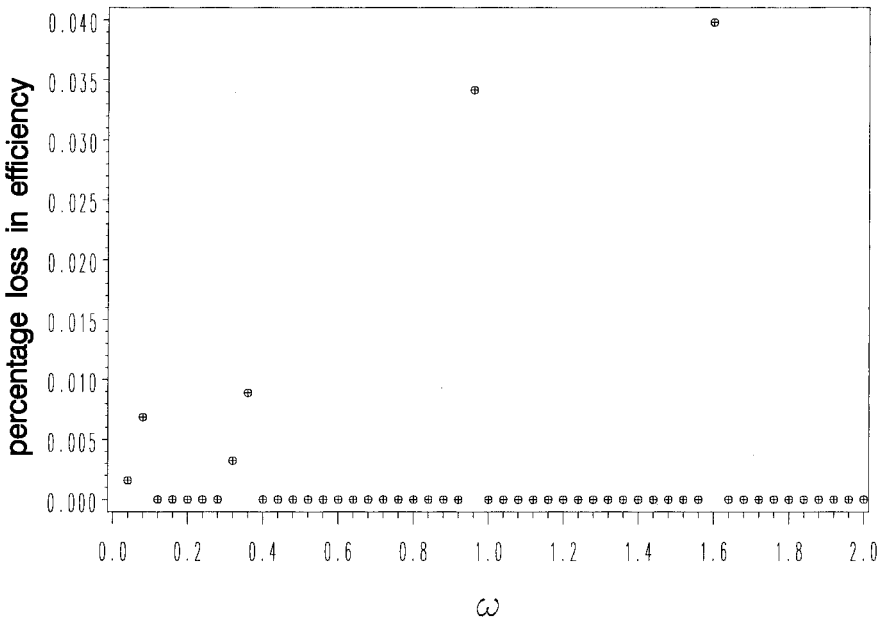

Fig. 2. Use of n_{opt} from (5.3) instead of from (5.2)

case, moreover,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{111} & x_{211} \\ 1 & x_{121} & x_{221} \\ 1 & x_{112} & x_{212} \\ 1 & x_{122} & x_{222} \end{pmatrix}$$

and $\mathbf{y} = (y_{11}, y_{21}, y_{12}, y_{22})^\top$.

The maximum likelihood estimator of $\boldsymbol{\gamma}$ as in (1.4) and its covariance matrix are given by

$$\hat{\boldsymbol{\gamma}} = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} \text{ and}$$

$$\operatorname{cov}(\hat{\boldsymbol{\gamma}}) = (\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X})^{-1}$$

(Longford 1993, p. 54), where $\mathbf{V}$ is a matrix of the form

$$\mathbf{V} = \begin{pmatrix} \tau^2 \mathbf{J}_n & & & \\ & \tau^2 \mathbf{J}_n & 0 & \\ & & \ddots & \\ & 0 & & \tau^2 \mathbf{J}_n \end{pmatrix} + \sigma^2 \mathbf{I}_{mn}$$

We are using $\mathbf{I}_d$ to denote the $d \times d$ identity matrix and $\mathbf{J}_d$ to denote the $d \times d$ matrix of all 1's. So $\mathbf{V}$ is a block diagonal matrix with entries of $\tau^2 + \sigma^2$ on the main diagonal, entries of τ^2 in the blocks but off the main diagonal, and 0's elsewhere. Note, in particular, that for $\tau^2 = 0$, $\mathbf{V}$ reduces to $\sigma^2 \mathbf{I}_{mn}$, and the maximum likelihood estimator $\hat{\boldsymbol{\gamma}}$ reduces to the familiar ordinary least squares estimator $\tilde{\boldsymbol{\gamma}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

Investigating the properties of the estimators of the regression coefficients is made

difficult by the dependence on the design matrix $\mathbf{X}$. We will only consider here a very simple design for a very balanced situation. We will let the first column of $\mathbf{X}$ be all 1's; this corresponds to estimating an intercept term in $\boldsymbol{\gamma}$. The second column of $\mathbf{X}$ will be a student-level indicator (''dummy'') variable, and the third column will be a school-level indicator variable. We assume the student-level indicator variable is balanced within a school and that the school-level indicator is balanced overall. This design is illustrated in (5.5) for the case of $n = 6$ students sampled per school and $m = 2$ schools sampled (but we are really interested in large m).

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{matrix}}\right\} \text{school 1} \\ \left.\vphantom{\begin{matrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{matrix}}\right\} \text{school 2} \end{matrix} \tag{5.5}$$

When m and n are both even, an explicit expression can be derived for the matrix $\text{cov}(\hat{\boldsymbol{\gamma}}) = (\mathbf{X}^\top \mathbf{V}\mathbf{X})^{-1}$ in terms of σ^2, ω, m, and n:

$$\begin{pmatrix} \frac{\sigma^2(2n\omega+3)}{mn} & -\frac{2\sigma^2}{mn} & -\frac{2\sigma^2(n\omega+1)}{mn} \\ -\frac{2\sigma^2}{mn} & \frac{4\sigma^2}{mn} & 0 \\ -\frac{2\sigma^2(n\omega+1)}{mn} & 0 & \frac{4\sigma^2(n\omega+1)}{mn} \end{pmatrix}$$

Let us minimize $\text{var}(\boldsymbol{\gamma}_0) = \frac{\sigma^2(2n\omega+3)}{mn}$, $\text{var}(\boldsymbol{\gamma}_1) = \frac{4\sigma^2}{mn}$, and $\text{var}(\boldsymbol{\gamma}_2) = \frac{4\sigma^2(n\omega+1)}{mn}$ subject to the simple cost constraint $C = C_s m + C_k mn$. The results are

$$n_{\text{opt},0} = \sqrt{\frac{3}{2}\frac{C_s}{C_k} \times \frac{1}{\omega}}$$

$$n_{\text{opt},1} = \frac{C - C_s}{C_k} \quad \text{and}$$

$$n_{\text{opt},2} = \sqrt{\frac{C_s}{C_k} \times \frac{1}{\omega}} \quad \text{respectively.}$$

The $n_{\text{opt},2}$ value is the same and the $n_{\text{opt},0}$ value is similar to that obtained in the traditional case (4.2). The $n_{\text{opt},1}$ value is equivalent to $m_{\text{opt},1} = 1$; we should only sample one school (were this practical) if we *only* want to estimate $\boldsymbol{\gamma}_1$. The variance of $\boldsymbol{\gamma}_1$, though, will be small in comparison to the variance of $\boldsymbol{\gamma}_0$ or $\boldsymbol{\gamma}_2$ for any reasonable design (no n in the numerator of the variance expression) so other considerations are more important. It is noteworthy that $n_{\text{opt},0} = \sqrt{3/2}n_{\text{opt},2}$ regardless of the costs. To settle on a single value for n_{opt}, one might consider minimizing $a\text{var}(\boldsymbol{\gamma}_0) + b\text{var}(\boldsymbol{\gamma}_2)$ for some $a \geq 0$, $b \geq 0$,

$a + b > 0$, subject to $C = C_s m + C_k mn$. The solution is

$$n_{\text{opt};a,b} = \sqrt{\frac{3a + 4b}{2a + 4b} \times \frac{C_s}{C_k} \times \frac{1}{\omega}}$$

In particular, if the two variances are weighted equally so that $a = b$, we have

$$n_{\text{opt};1,1} = \sqrt{\frac{7}{6} \frac{C_s}{C_k} \times \frac{1}{\omega}}$$

The author has informally explored some more complicated and less balanced cases, and the results were qualitatively like those given above. The variance of γ_1 may depend on ω (hence τ^2) but, in the cases looked at, does so in a bounded way.

It should be mentioned that the mathematical software Derive®, which does symbolic calculations, was employed heavily in doing these computations.

It seems that traditional sample designs may do very well in enabling us to estimate the regression coefficients. In analyses where it is important to estimate also the variance components, τ^2 in particular, one must instead sample more students per school (and fewer schools) as we saw in the previous subsection.

6. Final Comment

As hierarchical models become more widely used by researchers analyzing survey data, the need grows for survey design statisticians to understand the implications of such use for good survey design. This article, along with Snijders and Bosker (1993), Afshartous (1995), and Mok (1995), marks the beginning of an effort to develop such an understanding. But we have scarcely scratched the surface. Opportunities abound for further research on this topic.

7. References

Afshartous, D. (1995). Determination of Sample Size for Multilevel Model Design. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Bryk, A.S. and Raudenbush, S.W. (1992). Hierarchical Linear Models: Applications and Data Analysis Methods. Newbury Park, California: Sage.

Goldstein, H. (1987). Multilevel Models in Educational and Social Research. London: Griffin.

Goldstein, H. (1995). Multilevel Statistical Models. London: Edward Arnold.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Volume II (Theory). New York: Wiley.

Longford, N.T. (1993). Random Coefficient Models. Oxford: Clarendon.

Mok, M. (1995). Sample Size Requirements for 2-level Designs in Educational Research. Multilevel Modelling Newsletter, 7, 2, 11–15.

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.

Received July 1996
Revised October 1997

Sample size determination is the act of choosing the number of observations or replicates to include in a statistical sample. The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample. In practice, the sample size used in a study is usually determined based on the cost, time, or convenience of collecting the data, and the need for it to offer sufficient statistical power. In complicated studies there may be several different Determining sample size is a very important issue because samples that are too large may waste time, resources and money, while samples that are too small may lead to inaccurate results. In many cases, we can easily determine the minimum sample size needed to estimate a process parameter, such as the population mean . When sample data is collected and the sample mean is calculated, that sample mean is typically different from the population mean . This difference between the sample and population means can be thought of as an error.Â Assume that a previous survey of household usage has shown = 6.95 minutes. Solution We are solving for the sample size . A 95% degree confidence corresponds to = 0.05.