Sample Size Considerations for Multilevel Surveys

Michael P. Cohen
U.S. Bureau of Transportation Statistics
400 Seventh Street SW, Washington, DC 20590 U.S.A.
Michael.Cohen@BTS.GOV

1 Introduction

If researchers are planning to analyze data from surveys using multilevel models rather than concentrating on means, totals, and proportions, it is best to account for this in the survey design. One important aspect of the design is the sample size at each level (for example, the number of neighborhoods in the sample and the number of households sampled within each neighborhood). Cost models can be developed to determine the most efficient allocation of the sample.

To date, there has been only a limited amount of research on this topic. The most up-to-date account is given in Chapter 10 of Snijders and Bosker (1999). Except for the author's 1998 paper, the emphasis has been on small single-purpose surveys rather than on large federal surveys.

This talk will begin with an introduction and description of multilevel models. No prior knowledge of these models will be assumed.

2 Simple Two-Stage Design and 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 neighborhoods, and from each of the m neighborhoods, we select n households (a balanced sample design). It costs C_2 to include a neighborhood in the sample and an additional C_1 for each household sampled at the neighborhood. We wish to hold total sampling costs to our budgeted amount C where $C = C_2 m + C_1 mn$.

We refer to the *first stage units* as *neighborhoods* and the *second stage units* as *households* throughout this paper in order to avoid cumbersome terminology. Of course, the results apply more broadly.

In reality we would almost certainly select the neighborhoods by a stratified design. Additional levels (e.g., towns, persons) are possible. Unequal probability sampling might be used at any level. Our assumption of a balanced sample design (same number of households from each neighborhood) 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.

3 Traditional Sample Size Determination

Hansen, Hurwitz, and Madow (1953, pp. 172-73) have developed the formula for the optimal size n for the number of households to sample from each neighborhood. It applies to estimating means, totals, and ratios. A simple approximate version of the formula is $n_{\text{opt}} \doteq \sqrt{(C_2/C_1) \times (1-\rho)/\rho}$, where ρ is the measure of homogeneity, also called the intraclass correlation coefficient. The number of neighborhoods sampled is then $m_{\text{opt}} = C/(C_2 + C_1 n_{\text{opt}})$.

In the two-level setting, we have $\rho = \tau^2/(\sigma^2 + \tau^2)$, where σ^2 is the household level variance and τ^2 is the neighborhood level variance.

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. The determination of n and m is based on variance considerations but also ties in with the power of hypothesis tests and the widths of confidence intervals.

4 Regression Coefficients

For household i in neighborhood j, let us consider the simple the multilevel model $Y_{ij} = \beta_{0j} + r_{ij}$ where $\beta_{0j} = \gamma_{00} + \gamma_{01} z_{1j} + \cdots + \gamma_{0q} z_{qj} + u_{0j}$ and the $\{r_{ij}, u_{0j}\}$ are mutually independent random variables with $\mathrm{E}(r_{ij}) = \mathrm{E}(u_{0j}) = 0$, $\mathrm{var}(r_{ij}) = \sigma^2$, and $\mathrm{var}(u_{0j}) = \tau_0^2$. Notice that this simple model has no explanatory variables at the household level.

Suppose we want to estimate $\mathbf{a}'\boldsymbol{\gamma}$ where $\boldsymbol{\gamma} = (\gamma_{00}, \ldots, \gamma_{0q})'$ and $\mathbf{a}$ is a vector of constants $(a_0, \ldots, a_q)'$. This includes the case in which we are mainly interested in estimating a single coordinate of $\boldsymbol{\gamma}$. Let $\hat{\boldsymbol{\gamma}}$ be an

(asymptotically efficient) estimator of $\boldsymbol{\gamma}$. Let m denote the number of neighborhoods in the sample; let n denote the number of households in each neighborhood in the sample (assumed constant); and let $\mathrm{E}(\mathbf{z}_j) = \boldsymbol{\mu}$ and $\mathrm{var}(\mathbf{z}_j) = \boldsymbol{\Sigma}_z$. As in Snijders and Bosker (1993, pp. 248–249),

$$\mathrm{var}(\mathbf{a}'\hat{\boldsymbol{\gamma}}) \approx \frac{1}{m}\left(\tau_0^2 + \frac{\sigma^2}{n}\right)\ \mathbf{a}'(\boldsymbol{\mu}_z\boldsymbol{\mu}_z' + \boldsymbol{\Sigma}_z)^{-1}\mathbf{a}.$$

They show that for the cost model $C = C_2 m + C_1 mn$, $n_{\mathrm{opt}} = \sqrt{C_2\sigma^2/(C_1\tau_0^2)}$. For this choice of n,

$$\mathrm{var}(\mathbf{a}'\hat{\boldsymbol{\gamma}}) \approx \frac{1}{C}\left(\sqrt{C_1}\sigma + \sqrt{C_2}\tau_0\right)^2\ \mathbf{a}'(\boldsymbol{\mu}_z\boldsymbol{\mu}_z' + \boldsymbol{\Sigma}_z)^{-1}\mathbf{a}.$$

Clearly, if we want to know the total cost C needed to achieve a specified value of $\mathrm{var}(\mathbf{a}'\hat{\boldsymbol{\gamma}})$, this will be

$$C \approx \frac{1}{\mathrm{var}(\mathbf{a}'\hat{\boldsymbol{\gamma}})}\left(\sqrt{C_1}\sigma + \sqrt{C_2}\tau_0\right)^2\ \mathbf{a}'(\boldsymbol{\mu}_z\boldsymbol{\mu}_z' + \boldsymbol{\Sigma}_z)^{-1}\mathbf{a}.$$

5 The Intraclass Correlation Coefficient

The variance expression for the intraclass correlation coefficient ρ is given in, e.g., Snijders and Bosker (1999), p. 21. In order to get a useful form for n_{opt}, we optimize instead a slightly different but nearly equal expression. This gives $n_{opt} = [1+\sqrt{1+8\rho(1+C_2/C_1)}]/(2\rho)+1$. The variance expression at n_{opt} can be easily solved for C, giving the cost needed to achieve a given (approximate) variance for $\hat{\rho}$.

References

[1] Cohen, M.P. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *J. Official Statist.*, **14**, 3, 267-275.

[2] Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory, Volume II (Theory). Wiley.*

[3] Snijders, T.A.B., and Bosker, R.J. (1993). Standard errors and sample sizes for two-level research. *J. Educational Statist.*, **18**, 237-259.

[4] Snijders, T.A.B., and Bosker, R.J. (1999). *Multilevel Analysis. Sage.*

RESUME

Les techniques et les programmes machine sont devenus disponibles pour traiter des données emboîtées, permettant la formulation des modèles à multiniveaux explicites avec des hypothèses au sujet des effets se produisant à chaque niveau et à travers des niveaux. Si les utilisateurs de données projettent analyser des données d'étude en utilisant les modèles à multiniveaux, ceci doit être expliqué dans la conception d'étude. Les implications pour déterminer des dimensions de l'échantillon (par exemple, le nombre de voisinages dans l'échantillon et le nombre de ménages échantillonnés dans chaque voisinage) sont explorées.

Techniques and computer programs have become available for dealing with nested data, permitting the formulation of explicit multilevel models with hypotheses about effects occurring at each level and across levels. If data users are planning to analyze survey data using multilevel models, this needs to be accounted for in the survey design. The implications for determining sample sizes (for example, the number of neighborhoods in the sample and the number of households sampled within each neighborhood) are explored.