Session VIIB: Sample Design and Estimation

Discussant Comments by
Stephen J. Haslett
Bureau of Labor Statistics and Massey University, New Zealand.
haslett_s@bls.gov or s.j.haslett@massey.ac.nz

1. Introduction

In discussing the two papers in this session, "Sample size considerations for multilevel surveys" by Michael P. Cohen and "Two sided coverage intervals for small proportions based on survey data" by Philip S. Kott, Per Gosta Andersson, and Olle Nerman, I will follow a format.

I will first make some general comments. Then for each of these papers (which I will consider in order of their presentation), I will:

(A) Discuss the contents of the paper
(B) Place this discussion in a wider context, including some comment on related problems and possible solutions

Firstly, general comments. Both sets of authors are to be commended, both for their choice of topic and for their presentations. The two problems have strong implications for survey design and analysis. They are also important issues when minimizing or reducing survey costs. Survey designs should be optimized for the intended type(s) of analysis, or else resources are not properly focused. Good confidence and coverage intervals should neither be under-estimated nor overly conservative, especially when sampling for rare or relatively rare population characteristics.

2. Optimal design for multilevel surveys

The paper "Sample size considerations for multilevel surveys" by Michael Cohen considers how best to design a sample survey when the intended analysis is not the usual estimation of (sub-) population means or construction of cross-classified tables. Instead, he has looked at the question of how best to design a survey when the primary analysis will be multilevel modeling of the survey data. ie the fitting of linear or non-linear models with hierarchical structure and fixed or random coefficients.

The distinction made between descriptive and analytic surveys is certainly an important one. Because of improved computing resources and ease of use of statistical techniques for survey data (through packages like Pccarp, Sudaan and WesVar), the emphasis is slowly moving toward analytic surveys, where more than means and simple contingency tables are required. For this

reason, as Michael Cohen notes, considering optimal design for such surveys, rather than simply relying on known optimality results for descriptive surveys is an important area, not just from a statistical but also a survey cost point of view

I see this as an important area for another reason. I cannot comment on surveys undertaken in the USA, but in New Zealand large and consequently expensive national surveys are too often undertaken and either a very large percentage of the budget is spent on design and field work and little beyond summary statistics is completed (leaving a rich resource largely untapped), or else analytic results are wanted after field work is complete but the design and/or budget place severe constraints on what can then be done.

I have a number of particular comments on the paper.

It is important to distinguish between structure in the population and structure in the sample, eg there can be clustering in the sample design without clustering in the population, but often there is both.

There is as Cohen noted a rather confusing difference in notation used in sample surveys and in the multilevel modeling literature: eg

- sample surveys: neighborhood = 1^{st} stage, household=2^{nd} stage
- multilevel modeling: neighborhood = level 2, household=level 1

Consistency of notation would be welcome, but these opposing conventions seem well established.

The analysis in Cohen's paper looks at the case where the number of units sampled at the finer level is equal for all groups, or first stage units. In practice, eg for people within households, there can be unequal numbers sampled. It would be useful to have an extension of the optimal design results to such unbalanced models.

The focus of the paper is on randomization inference. It is also useful to consider what in multilevel models for surveys is random with respect to the survey design (leading to randomization or design-based inference), and what is random with respect to the multilevel model (leading to superpopulation inference). The two inferential frameworks are often linked, and it is useful to make them explicit. For example, in the simple model used in Section 5 of the paper, the superpopulation model assumes exchangeability within neighborhood (ie households within each neighborhood have the same statistical properties (eg mean, variance), with the same variance but different means for all neighborhoods. This is worth checking both against the sample data, and against what is known about the population at design stage. There is no point having an optimal design for a superpopulation model that does not adequately describe the actual situation. The point is that superpopulation models are not to be 'plucked from the air', but must reflect the relevant aspects of the underlying population.

The approximation that $n\sim=n-1$ made in equation (3), does not hold when sample size at the second stage is small, and the extension of the results to this case (eg for people within households) would be useful, even if as Cohen states, the results are not mathematically elegant. Similarly the equation for n_{opt} does not guarantee that n_{opt} is an integer, and the rounding effect

may be important when n_{opt} is small, as would be the case if n_{opt} were number of people to be sampled within household. Rounding n_{opt} by say 20% also has the same percentage effect on total sample size, since the same percentage rounding occurs at all higher levels in the model (ie earlier stages of the design). The importance of finite population corrections in this context may also warrant investigation. An analysis of how n_{opt} is affected by the fact that rho is estimated rather than known would also be useful.

Further possible extensions may include:

- multivariate optimality: looking at principal components has proved useful for descriptive surveys. Is it applicable or useful in analytic ones?
- using a combined randomization / superpopulation approach, given the analytic models have a superpopulation setting (eg time series, regression, logistic regression, relative risks, generalized linear models)

What follows below is a summary of some of currently unpublished results (which I presented last year in a seminar to the Washington Statistical Society). These begin to clarify the links between optimal sample design and the superpopulation structure for generalized linear models.

Linear models can be written in the form:

$$\mathbf{Y} = X\boldsymbol{ß} + \mathbf{e}$$

where $\mathbf{Y}$ is a vector of observations, X variables are assumed known predictor variables, $\boldsymbol{ß}$ is a vector of regression coefficients, e is the error vector with mean zero and variance covariance matrix V which would be $\sigma^2 I$ for a simple linear regression. In general V has to be a positive definite symmetric matrix.

Generalized linear models extend this framework by allowing $\mathbf{Y}$ to be some function g of the observations $\mathbf{Y}_0$, namely $g(\mathbf{Y}_0)$. Linear models have this link function g equal to the identity. The solution using generalized least squares makes it clear that, at each iteration for a generalized linear model, a linear model is actually being fitted, but that the underlying metric specified via V is changing at each iteration. See for example, del Pino (1989). Extensions to the case where some of the components of $\boldsymbol{ß}$ are random are possible.

To extend the generalized linear model to a superpopulation and then a randomisation context, let $\mathbf{Y}=(Y_1, Y_2, Y_3,\ldots Y_N)^T$ where N is the population size, X is N*p, e is N*1, and V is N*N. Then define the superpopulation by ξ and the expectation with respect to ξ by E. Note that E is defined for each i and leaves these labels i intact, in general. If the ξ distribution is stochastically degenerate, as is the case for purely design based inference, then the elements of e are fixed quantities because Y_i has only one value for a given i.

Define an N*N matrix c with consisting of zeros and ones, with all off-diagonal elements equal to zero, and n of the N diagonal elements of c equal to one. (Note that here n is the total sample size.) Then c is a sample selection matrix, and if p is defined as the N*N matrix with all off

diagonal elements equal to zero and ith diagonal element equal to π_i the selection probability for the ith population element, and E is expectation with respect to the survey design, p, then

$$E(c\ p^{-1}\ \mathbf{Y}) = \mathbf{Y}$$

Note that for a particular sample s only n of the N elements of $c\ p^{-1}\ \mathbf{Y}$ are non-zero and each of the non-zero elements is just Y_i / π_i. The elements of $c\ p^{-1}\ \mathbf{Y}$ can be permuted so that the first n are non-zero without loss of generality.

What we know have is a structure that allows us to see the relationship between the superpopulation structure and the design, and hence to optimize that design for a given model, in this case the generalized linear model which determines the ξ superpopulation distribution.

The discussion given below can be extended to random and mixed parameter models for $\mathbf{ß}$, but the fixed parameter case only is considered below for simplicity.

Given a sample of size n, construct

$$c\ p^{-1}\ \mathbf{Y} = c\ p^{-1}\ X\ \mathbf{ß} + c\ p^{-1} e$$

Distinguish two cases:

(a) e is fixed with respect to the superpopulation, which is the usual design based assumption
(b) e has stochastic properties with respect to the superpopulation.

Under (a)

$$Var_p\ (c\ p^{-1} e) = p^{-1} (p_{\mathbf{d}} \# e\ e^{\mathbf{T}})\ p^{-1}$$

where Var_p denotes expectation with respect to the design, # indicates a Hadamard product (ie elementwise multiplication), and $p_{\mathbf{d}}$ has i,j th element $[\pi_{ij} - \pi_i\ \pi_j]$.

ie $Var_p\ (c\ p^{-1} e)$ has i,j th element $[(\pi_{ij} - \pi_i\ \pi_j) / (\pi_i\ \pi_j)]\ \varepsilon_i\ \varepsilon_j$

In general there are approximations necessary because $c\ p^{-1}\ X$ is stochastic, but this is true of any regression technique applied to unit level survey data. In practice if X is known for all population elements, using X rather than $c\ p^{-1}\ X$ is better. This is the case, for example, if the matrix X designates membership of various subpopulations (the size of which is known, and membership of which is known for all sampled i), or if X is null so that $c\ p^{-1}\ \mathbf{Y} = c\ p^{-1} e$.

In this latter case, the estimator of a sample mean is the Horwitz-Thompson estimator $\Sigma_s\ Y_i / \pi_i$ and the variance estimator is the usual estimator ie $\Sigma_{i \varepsilon s}\ \Sigma_{j \varepsilon s}\ [(\pi_{ij} - \pi_i\ \pi_j) / \pi_{ij}]\ Y_i Y_j / ((\pi_i\ \pi_j)$.

This result shows that the Horvitz-Thompson estimator is the optimal estimator for the simplest possible generalized linear model, where the error structure rather than being stochastic with respect to some superpopulation, is fixed for every i=1,2,….N.

Under (b), e is stochastic with respect to the superpopulation.

More specifically, this implies, $E(\mathbf{e}) = \mathbf{0}$, and $E(\mathbf{e}\mathbf{e}^{\mathbf{T}}) = V$, so that

$$\begin{aligned} \mathrm{Var}_{p\xi}(c\ p^{-1}\mathbf{e}) &= \mathrm{Var}_{p\xi}(p^{-1}c\ \mathbf{e}) \qquad \text{since } c \text{ and } p \text{ are both diagonal matrices} \\ &= p^{-1}\,\mathrm{Var}_{p\xi}(c\ \mathbf{e})\ p^{-1} \\ &= p^{-1}\,(\mathbf{p_{nd}} \# V)\ p^{-1} \\ &= V_{p\xi} \text{ say} \end{aligned}$$

where $\mathrm{Var}_{p\xi}$ denotes variance with respect to the joint design superpopulation distribution, $\mathbf{p_{nd}}$ has i,jth element π_{ij} , $\pi_{ii} = \pi_i$, and as before # denotes the Hadamard product.

Note that if V is diagonal then $V_{p\xi} = p^{-1} V = V\ p^{-1}$, and if V is diagonal with equal diagonal elements σ^2, then $V_{p\xi} = p^{-1}\sigma^2$.

In general, inversion of $V_{p\xi}$ is necessary to get estimates of the parameters $\boldsymbol{\beta}$ in the generalized linear model. The n*n submatrix of $V_{p\xi}$, $V_{11p\xi}$ say, that is defined by the sample is not however all that is necessary to specify the corresponding n*n submatrix of $V_{p\xi}^{-1}$, namely $V_{p\xi}^{(11)}$ (see Rao, 1973, p33, for example) so that in general the design and superpopulation properties of all N units are required to find GLS or IGLS solutions, even though $c\ p^{-1}\mathbf{Y}$ and $c\ p^{-1} X$ contain only n non-zero elements or rows.

When $V_{p\xi}$ is diagonal however, $V_{p\xi}^{(11)} = V_{11p\xi}^{-1}$, and when $V = E(\mathbf{e}\mathbf{e}^{\mathbf{T}})$ is diagonal with equal elements, then , $V_{p\xi}^{(11)} = V_{11p\xi}^{-1} = p_{\mathbf{s}}^{-1}\sigma^{-2}$, where $p_{\mathbf{s}}$ is the diagonal n*n submatrix of p which contains selection probabilities for the sampled elements only. Defining $X_{\mathbf{s}}$, $\mathbf{Y_s}$ in the parallel manner, then

$$\hat{\boldsymbol{\beta}} = (X_s^T p^{-1} X_s)^{-} X_s^T p^{-1} \mathbf{Y}_s$$

which is the 'π-weighted' estimator used for example in Skinner, Holt and Smith (1989), and by Sudaan and WesVar. This estimator does not use any of the joint selection probabilities $\{\pi_{ij}: i=1,2,\ldots N;\ j=1,2,\ldots N\}$ however, and also requires very strong distributional assumptions on V. Being able to ignore the joint selection probabilities, and assuming no clustering in the population are both assumptions that will only be met occasionally with real populations and complex survey designs. eg in the linear regression case, the variance of $\hat{\boldsymbol{\beta}}$ is estimated to be the same for a simple random sample as for an epsem cluster design, which is clearly not always true. The methods outlined above are more general, because they allow for more detailed specification of superpopulation structure, and for incorporation of joint selection probabilities.

They also lead to some interesting results about optimal survey design.

Detailing these in full would take more time and space than is available here. What is given here is a very brief summary, without proof, of some useful results:

- If V were diagonal, we could choose $V_{p\xi} = \sigma^{-2} I$ where I is the identity matrix for some σ^2 ie $\pi_i = 1 / \sigma_i^2$ where $V = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2, \ldots\ldots \sigma_N^2)$.
- However V is not diagonal in general, and it is not possible to make the off-diagonal elements of $V_{p\xi}$, ie $(\pi_{ij} \sigma_{ij} / \pi_i \pi_j)$, equal to zero since setting π_{ij} equal to zero is not a possibility if unbiased estimates of variance are required, even for a simple mean statistic.
- McElroy (1967) has shown that ordinary least squares and general least squares give the same parameter estimates, when both models include $\boldsymbol{i}$ (a vector of ones) as one of the columns of X, if and only if the variance – covariance matrix of the errors (which here is $V_{p\xi}$) can be written as $V_{p\xi} = a\boldsymbol{I} + b\boldsymbol{i}\boldsymbol{i}^T$, where a and b are scalars. Thus if V has equal diagonal elements, the optimality requirement is that $(\pi_{ij} \sigma_{ij} / \pi_i \pi_j) = k$ where k is a constant, ie π_{ij} is proportional to $\sigma_i\sigma_j / \sigma_{ij} = 1 / \text{corr}_{\xi} (\varepsilon_i, \varepsilon_j)$, so that the π_{ij} should be chosen as inversely proportional to the corresponding superpopulation covariances.
- There is also a more general extension of McElroy, due to Mathew (1983) which considers when two general least squares solutions are equal. Using this, and making approximations of $O(N^{-1})$, yields π_i proportional to $1 / \sigma_i^2$, and π_{ij} proportional to $1 / \sigma_{ij}$ ie choose selection probabilities inversely proportional to the superpopulation variance for each i, and joint selection probabilities should be smallest when two population elements are highly correlated. What is interesting is that is what experienced survey designers actually do (even in large government surveys when superpopulations are seldom mentioned).

3. Two sided coverage intervals for small proportions

The paper by Philip S. Kott, Per Gosta Andersson, and Olle Nerman,. "Two sided coverage intervals for small proportions based on survey data" is another important area. Small proportions occur often, for example in health surveys, when sampling for relatively rare characteristics. However the central issue of getting only a small amount of information on people having the characteristic, even from a relatively large survey, is not confined to health surveys, as the comments below will indicate. Surveys of this type are very resource intensive and more accurate information on confidence intervals or coverage for small proportions would be useful not just for analysis but also at design stage, especially where information is sought for sub-populations.

Again, I have a number of particular comments on the paper.

The focus in the paper is on two sided confidence intervals; a solution to the 'one sided' problem remains illusive.

It would be useful to have empirical results on the utility of the methods in the paper in 'real life' situations, as well as those presented here from simulation.

A comparison of this paper with that of Korn and Graubard (1998) would be useful . Korn and Graubard look at using the Clopper-Pearson confidence intervals for small proportions, and replacing the sample size n in these formulae with the effective sample size n* = n/deff, where deff is the design effect. They showed, based on a household survey where the ratio of largest to smallest sample survey weighting was less than ten, that the coverage properties of such 'exact' binomial intervals were better than logit transform confidence intervals, normal based confidence intervals, and intervals based on the Poisson approximation to the binomial.

It is interesting to compare Korn and Graubard's conclusions with the comment in the present paper that this technique of using effective sample sizes instead of actual sample sizes does not work particularly well with the Wilson method of estimating coverage intervals. It would be useful to know why. Some consideration of the number of moments that must be estimated, even given a small sample, may be useful when considering stability properties of the Kott - Andersson - Nerman estimator. Its optimality for particular proportions rather than over the whole possible range of p is however clearly an advantage if the stability issue can be resolved.

Further possible research areas include:

Sensitivity issues, given the discrete nature of the problem. For a small proportion there can be a small number of successes even in a large sample. ie the sample size on its own is not a particularly good measure of the information content, eg the New Zealand Gaming Survey 1999, where n~=6000, but number of problem and pathological gamblers in the sample was less than 200. This is particularly important for sub-population estimates.

It would be useful to have a technique for confidence interval estimation (or for estimating coverage) that will allow for balanced repeated replicates since this technique is frequently used in practice. One possible issue with the Wilson method and its amendments is that required moments beyond the second cannot be estimated from the half samples inherent in BRR. This issue is connected to the open question of how well these higher moments can be estimated in situations where even estimating the first two (ie mean and variance) is not always entirely straightforward.

Extensions to the situation where BRR is used, or one-sided confidence intervals are required may be possible, eg by use of the bootstrap.

At this stage, I would like to describe a novel use of the bootstrap as it can be applied to sample survey problems (We have called the technique the 'exchangeable bootstrap'). Although it also can be used for more general problems, it provides an alternative solution to the problem of confidence intervals with good coverage for small proportions. This is illustrated with reference to a nationwide survey of problem and pathological gamblers in New Zealand in Gray, Haslett, and Kuzmicich (2001).

Essentially, the exchangeable bootstrap partitions the population into sub-populations that can be considered exchangeable, eg subgroups having the same (superpopulation) mean and variance. For example, when estimating a mean, a simple random sample is an optimal design technique

when all the population units can be considered as a single exchangeable group. The concept of exchangeable groups can be considered as the explicit basis for determining strata, when minimizing the design variance of a mean, total, or proportion during the design stage of a survey. Exchangeable groups need not however be strata. They may be subdivisions of strata, but they may also cut across strata and may use the same or different allocation variables.

Once the exchangeable groups have been determined, the realized sample can be allocated to these groups. The bootstrap sample is now drawn and, for a linear estimator, the bootstrap estimate is formed as follows:

- For each unit, i, in the sample, take the sample survey weight for unit i, and multiply it by the realized value of an observation drawn from the particular exchangeability group to which unit i belongs.
- Sum the products over the whole sample.

This bootstrap estimator can be shown to have some useful properties, given the correct exchangeability groupings:
e.g. For estimators of means, totals and proportions,

- It is design-superpopulation unbiased if the sample survey estimator is design unbiased.
- It is consistent for the population mean if and only if the design based sample survey estimator is consistent.

One very nice feature of the exchangeable bootstrap is that if the exchangeable groups are poorly chosen, then the average of the bootstrap estimates is very different to the sample survey average. This provides a very useful diagnostic for proper choice of exchangeable groups. For further details, and a detailed example based on the New Zealand Gaming Survey data, see Gray, Haslett and Kuzmicich (2001).

The exchangeable bootstrap has potentially wider application than to estimation of small proportions or even of linear or linearizable statistics. Some of its properties for linearizable statistics are already known (again see Gray, Haslett and Kuzmicich, 2001). In particular, it holds promise for sound estimation of small proportions and their confidence intervals from sample survey data, even for sub-populations. As Gray, Haslett and Kuzmicich show it has useful stability properties even when the ratio of smallest to largest sample survey weight exceeds fifty.

4. Conclusion:

Both of the presented papers focus on important topics, not only because the problems are statistically interesting, but also because they have major cost implications for large scale sample surveys.

In addition to the possibility of cost reductions, both papers also offer opportunities for better survey design, better data, and sounder statistical analysis. They are both topics deserving considerable further research.

The authors are to be thanked both for their papers and for their choice of topics.

References

del Pino G (1989) The Unifying Role of Iterative Generalized Least Squares in Statistical Algorithms; Statistical Science. 4, p394–408.

Gray, A., Haslett, S. and Kuzmicich, G. (2001) Confidence intervals for proportions estimated from complex sample designs, *Statistics New Zealand Research Report*, No. 21, Statistics New Zealand, Wellington, New Zealand.

Korn, E.L. and Graubard, B.I. (1998) Confidence intervals for proportions with very small expected number of positive counts estimated from survey data, *Survey Methodology*, 24, p193-201.

McElroy, F.W.(1967) A necessary and sufficient condition that ordinary least squares estimators are best linear unbiased, *Journal of the American Statistical Association*, Vol. 62, p1302-1304.

Mathew, T. (1983) Linear estimation with an incorrect dispersion matrix in linear models with a common linear part, *Journal of the American Statistical Association*, Vol. 78, p468-471.

Rao, C.R. (1973) *Linear Statistical Inference and Its Applications*, 2nd Edition, John Wiley and Sons.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989) *Analysis of Complex Surveys*, John Wiley and Sons.