BASIC CLUSTER SAMPLE DESIGN

C. H. Proctor, North Carolina State University
Department of Statistics, NCSU, Raleigh, NC 27695-8203

KEY WORDS: Coefficient of heterogeneity, Smith's b, intracluster correlation, optimum cluster size, required sample size

1. Introduction.

Sample surveys have been done in so many different fields of application that choosing a design usually means reviewing the experiences of others and settling on the most popular design. For present purposes such background experiences will be taken to be relatively absent. Suppose that one can visualize fairly well how the simple random sample design would work and wishes to consider increasing sampling unit size. That is, we have defined the element along with its measurement operation and now wish to construct clusters of elements to see if sampling such clusters would offer advantages over the sampling of elements.

First we review some formulas useful in setting sample size and finding the best size of cluster when doing basic cluster sampling. The word "basic" means that there are no natural units such as classrooms, farm fields, bales, or other chunks having boundaries that should be utilized as strata or as first stage units. The formulas are based on a measure of relative cluster heterogeneity called Smith's b and we will show how values for b can be obtained. The illustrative data come from dot sampling of aerial photos to estimate the proportion of land in forest. Although this type of sampling is more often done with spatially systematic (widely spaced) methods (Zeimetz, *et al*, 1976), forest areas do occur in clumps and clumping is also common for soils and vegetation characteristics for which enumeration on ground level is best done using compact clusters.

The following discussion starts by reviewing the steps one takes in designing a small scale cluster sample survey. We have in mind a household or farmstead survey as a typical case, although many surveys of plants, trees, soils, and other material should fit the description as well. We first furnish the formulas for required sample size, for best size of compact cluster and for design effect. Recall that design effect is the ratio of the number of elements sampled in clusters to elements sampled individually that give the same precision. We also furnish methods of finding the value of Smith's b which is used in the formulas for cluster size and for design effect, and we illustrate the methods with examples. Finally we illustrate methods for determining cost coefficients.

2. Preliminary Steps

It is helpful, in designing small or moderate sized sample surveys to begin by imagining the design as a multi-start systematic sample of clusters. Actual definition of the boundaries of the clusters would be done only in the vicinity of those selected so that frame construction costs are kept down. The frame is ordered, as best as can be, so the systematic design feature achieves hidden stratification. Use of several starts allows the sample to be drawn as replicated subsamples which can then be used, via the Tukey Jackknife, to estimate sampling variances and biases.

If the population to be sampled is novel then a challenging job will be to settle on the frame material. One will be told of ideal materials such as lists in some office somewhere, or sketch maps that one might be able to get permission to

use, or aerial photos that can be ordered, but it's usually best to start by considering the simplest materials such as homemade sketch maps or a tourist map and then try to find improved materials.

Small scale surveys often have time limitations as well as cost restrictions. Although the questionnaire must be pretested and the enumerators trained, fieldwork could conceivably commence within a few weeks so one cannot wait too long to assemble frame materials. The basic "skeleton" material must be capable of dividing up the population into fairly large, so called, count units each one having a well delineated boundary. Then there must be some information on sizes of the count units to be used in assigning numbers of clusters. Selection of a sampled cluster will then lead to its count unit, which itself must then be further subdivided in order to discover (by rerandomized selection) the sampled cluster. This final subdivision will commonly be combined with enumeration and done in the field.

The U. S. Census Bureau provides Block Statistics for moderate or larger sized cities and the city block then becomes the count unit, while numbers of households from the last census can be used as sizes. In open country or rural areas of the U. S. one can use BNA's (block numbering areas) as count units. The population census of any country provides an excellent list of count units as names of places such as villages with either population counts or map areas, as sizes. The nature of the size information will be chosen to represent a compromise between accuracy and cost. It may be simply the guesses of someone who knows the territory.

3. Sample and Cluster Size Formulas

In terms of elements, not clusters, one can calculate required sample size for estimating a mean or a total as:

$$n = \left[\frac{\text{Population CV}}{\text{Required Sample CV}}\right]^2 \qquad (3.1)$$

where Population CV is the ratio of guessed population standard deviation divided by guessed population mean and Required Sample CV is 1/10 for a "minimally adequate precision" or 1/20 for "adequate precision" or 1/100 for "good precision." Population CV's range from .1 to 3 or 4 but most are around .3 to 1 or 2 and for novel characteristics one simply inquires about the ranges where common values are found and then judges the magnitude of the Population CV. For the estimation of a proportion around 50% the Population CV is 1, and the three required sample sizes are thus 100, 500 and 10,000.

A bit more difficult design question is that of the size of a cluster. The basic formula gives optimum cluster size as:

$$M_{opt} = b\, C_1/(1-b)C_2\ , \qquad (3.2)$$

where C_1 is the cost of adding another cluster to the sample and C_2 is the cost of adding another element to a cluster, while b is Smith's (1937) b. Let's consider how to obtain a value for b first.

The quantity b equals 1 if there is no adjacency correlation and equals 0 if adjacency correlation is maximal. Thus b reflects relative independence or "heterogeneity" as H. F. Smith (1937) called it. If there were no adjacency correlation then the variance of cluster means would equal $\sigma^2/M = \sigma^2 M^{-1}$ where σ^2 is population variance. For a given shape of plot and a given characteristic, adjacency correlation is often found to cause the variance of cluster means to equal $\sigma^2 M^{-b}$ where b is in the range 0 to 1. This is Smith's "law."

In actual data one may notice gradual changes in b as M varies. For some variables the cluster variances may be poorly and only

erratically fit by Smith's law. Such cases can occur with economic data on households, for example, as one goes from city blocks to tracts to whole cities to regions. Fortunately, such cases are relatively rare and can be easily foreseen, while constancy, or a gradual change with near constancy in the range of interesting cluster sizes, is more usual.

There are at least four ways to get a numerical value for b. The first is by judgement. A default is .5. Smith found b = .75 for yield of wheat on plots ranging from 1/2 foot to 36 feet of row. We found b = .1 for disease incidences among tobacco plants in plots of all sizes within a field (Proctor, 1985). One can thus often judge the amount of clustering for the variable in question as intermediate between wheat yields and tobacco diseases -- especially if one has familiarity with biological phenomena.

The second way is to convert values of the intracluster correlation coefficient (written δ, ρ or "roh") to b values. In Hansen, Hurwitz and Madow's (1953) Chapter 6, Tables 3 and 4 show by way of values for δ that b is .4 for the agricultural items and ranges from .4 to .9 for some socioeconomic and demographic variables. The formula to use for conversion is

$$b = 1 - \ell og[(M-1)\delta + 1]/\ell og\ M. \quad (3.3)$$

An even more empirical (third) way is to find data on a variable similar to the survey variable and to estimate the population variance, as s^2, and that among cluster means, as s_y^2, for some given cluster size M. Then one solves Smith's law for b as:

$$b = -\ell og(s_y^2/s^2)/\ell og\ M\ . \quad (3.4)$$

A more elaborate variation on this method (the fourth way) when one has the data, is to do a nested analysis of variance for various sizes of nested clusters and use the estimation methods in (Proctor, 1985). Both of these latter two types of calculations will be illustrated for the data shown in Figures 1 and 2.

4. Illustrative Calculations for Smith's b

The ones in Figures 1 and 2 represent dots on an aerial photo that hit woodlands and the zeroes hit something else. The data were interpreted by Joop Faber (1971) from two 1:200,000 aerial photographs of the Lake Mickey watershed near Durham, NC. Each Figure has 2500 points in a 50-by-50 square lattice. On the ground the side of this square measured about 2.5 miles. If clusters are formed as 5-by-5 squares then there will be 100 square clusters. The variance among these 100 cluster means for Figure 1 can be found to be .0740 while the total variance is .2478 [since the proportion of forested points is p = .548 for the photo, the total variance is approximately p(1-p)]. Thus formula (3.4) may be used to find b = .38, for squares having M = 25 points. Applying the same calculation to Figure 2 yields b = .35.

For application of the more elaborate nested ANOVA we trimmed away the last two columns and last two rows of both figures to get a 48-by-48 lattice. Now we can create 2-by-2, 4-by-4, 8-by-8 and 16-by-16 nested squares as well as nested row and column transects with lengths 2, 4, 8, 16 and 48. The resulting analyses of variance and estimates of Smith's b are found in Tables 1 and 2.

The basic message for both squares and transects concerning the value of Smith's b is its gradual increase when the length of a side or transect exceeds 16. Below this size its value is stable at around .35 to .40. The global estimate with $\alpha = 0$ in tables 1 and 2 is based on supposing b to be constant and averaging interlevel b's roughly in accord with their degrees of freedom. The one with $\alpha = .01$ is based on supposing the departure from constancy is due to haphazard lack of fit and averages

interlevel b's more nearly equally. However, since the departures from constancy are not haphazard it is the actual pattern of interlevel b's that needs to be kept in mind for these data.

Table 1. Square Plots ANOVA and Smith's b.

1a. Analysis of Variance

Sizes	Degrees of Freedom	Mean Squares Photo 33	Mean Squares Photo 86
16 x 16 = 256	8	1.23318	4.64941
8 x 8 = 64	27	3.00434	2.28487
4 x 4 = 16	108	1.18432	.88759
2 x 2 = 8	432	.32465	.34505
1 x 1 = 1	1728	.12355	.11646

1b. Smith's b Estimation, Interlevel and Two Global Estimates

64 to 256	1.54	.64
16 to 64	.62	.47
8 to 16	.35	.41
1 to 8	.34	.33
α = 0 Estimate	.35	.31
α = .01 Estimate	.46	.39

Table 2. Transect ANOVA and Smith's B Estimates

2a. Analysis of Variance

Sizes	Degrees of Freedom	Mean Squares Photo 33 Row transects	Mean Squares Photo 33 Column transects	Photo 86 Rows	Photo 86 Columns
48	47	.45242317	.5694444	.6834275	.93076795
16	96	1.3346354	.57942708	.66666667	.97005208
8	144	.53993056	.76388881	.69487847	.50217014
4	288	.31597222	.49392361	.35980903	.35112847
2	576	.21788194	.18576389	.19748264	.21223958
1	1152	.11197917	.11371528	.11414931	.09765625

2b. Smith's b Estimation, Interlevel and Two Global Estimates

Interlevels

16 to 48	1.76	1.01	.98	1.03
8 to 16	.60	1.22	1.02	.61
4 to 8	.48	.80	.61	.57
2 to 4	.48	.40	.46	.48
1 to 2	.37	.37	.40	.33
α = 0 Estimate	.44	.43	.45	.36
α = .01 Estimate	.72	.77	.70	.61

This may be an opportune place to remark that the Smith's b values constitute a means of characterizing spatial autocorrelation that is in conceptual competition with the spatial correlogram, or a variogram or a spatial spectral density function. Obviously the Smith's b characterization, even when one recognizes the patterns of nonconstancy, is inferior to these more complex methods. On the other hand we believe that for the limited purposes of sample design the Smith's b is just right. Its only serious competitor here is the intracluster correlation coefficient, and again we believe the Smith's b formulation is both more flexible in fitting the data and leads to more convenient formulas.

5. Finding Cost Coefficients

As with the Population CV and the Smith's b, the values for C_1 and C_2, the cost coefficients, can be obtained by judgement, by experience with similar material or from data. Using judgement and an active imagination, one considers all the operations of the survey needed to collect and analyze the data, and expresses them in man hours or in dollars. Next we suppose an additional element is added to a cluster and ask how much time or money does this add to survey cost. This is C_2. Then we suppose an additional cluster must be drawn and calculate its expense. This is C_1. Since only the ratio C_1/C_2 enters the optimizing expression we are basically interested in it. Usually (translate "In my experience") C_1 to C_2 is around 10 to 1. Hansen, Hurwitz and Madow (1953) cite 2 to 1 and this can happen when household interviews take all day or when there will be repeated visits, but for physical measurements such as moisture readings, or weed counts or soil analyses the ratio can easily be higher than 10 to 1.

Although such judgements are entirely appropriate for small or moderate sized surveys, since departures from optimality will here not be too costly, as the scale of survey increases one should consider doing a field trial of the procedures to estimate C_1 and C_2. To illustrate the method, let's consider asking each student in a sampling class to conduct three surveys on his copy of an aerial photo. When we in fact carried out the exercise the three surveys were: A was of n = 20 single (M=1) points; B was of n = 10 clusters of M=5 points in a row; and C was of n = 5 clusters each of M=10 points in a row. Table 3 shows the data for 13 students and one can verify that regressing time = Y on X_1 = n and X_2 = nM (where the factor student at 13 levels was removed as a blocks variable) gives regression coefficients C_1 = .72 and C_2 = .06. The cost ratio is thus around 12.

6. Design Effect and Setting the Final Design Features

Having estimated the cost ratio and also having estimated Smith's b one uses (3.2) and finds

$$M_{opt} = .4 \times 12/.6 = 8.$$

This is a recommendation for using 8 points. Since cost coefficients are available only for the row shape we consider just that case. Since 8 is close to 10 and M = 10 is a more convenient cluster size we will actually use M = 10. We can now return to the design of the sample after, it will be recalled, having calculated the required sample size in elements.

Knowing M and b we can calculate the design effect as:

$$D^2 = M^{1-b} \qquad (6.1)$$

or, for our example, as $3.98 = 10^{.6} = 4$. This shows that one needs four times as many points in a cluster sample to get the same precision as a (scattered) simple random sample. Thus, with n = 400, for example, from (1) we would be led to a sample

design of 4 x n/M = 160 clusters, each of 10 points in a row. The 160 sampling units themselves would actually be drawn as a systematic sample (or as several such) and so might achieve more precision than the simple random selection underlying the formulas.

7. Postscript

Having seen in a fairly concrete form how one can design a cluster sample survey and thereby having gotten a notion of how it might be carried out, we hope the reader will be in a better position to evaluate cluster sampling as compared, say, to element sampling or to a census or, perhaps, to some other method such as a purposive sample. A close cousin to the cluster sample design is a sample in two or more stages. The cluster sample procedure as just described is two-stage in operation but, although the clusters in a count unit are randomly permuted just before selection, it is in effect a one-stage design.

It is this simplicity of design that makes for simplicity in tabulations. Each enumerated element will carry the same basic raising factor of N/n. These factors can then be adjusted for case nonresponse, for item nonresponse, for subsampling, if such had to be carried out because of surprises in the actual numbers of elements in some selected clusters, and so forth. Each enumerated element can also be provided with its replicate raising factor and the separate estimates from replicated subsamples can be simply calculated and can be used in variance estimation.

Although this review has moved fairly rapidly over a number of topics, they are the critical ones in cluster sample design. Sample size depends on the usual requirements for precision but also on the design effect which in turn depends on the pattern of adjacency correlation as reflected in Smith's b. This b value can be judged or based on data as we've shown.

Choice of cluster size requires that one know both the b value and a ratio of costs and again we illustrate a judgement method as well as an empirical approach for getting this cost ratio. The reason we emphasize judgement methods for Smith's b and for the C_1/C_2 ratio is that one can seldom justify the expense of pilot surveys. This is essentially the same reason one chooses the cluster design, that is so as to avoid the expense of more elaborate frame materials.

Table 3. Times in minutes required by 13 students to carry out three sample surveys of an aerial photo.

	Survey		
	A	B	C
Student	$n = 20$ $\mu = 1$	$n = 10$ $\mu = 5$	$n = 5$ $\mu = 10$
1	22	16	11
2	20	15	7
3	25	12	11
4	10	8	15
5	5	3	1
6	13	12	6
7	20	11	7
8	12	18	16
9	19	13	20
10	13	11	4
11	40	25	10
12	24	15	9
13	15	10	5

8. References

Faber, J. A. J. (1971). "Precision of sampling by dots for proportions of land use classes," Institute of Statistics Mimeo Series No. 773, Raleigh, NC.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory Vol. 1 Methods and Applications*. New York, Wiley.

Proctor, C. H. (1985). "Fitting H. F. Smith's empirical law to cluster variances for use in designing multi-stage sample surveys," *Journal of the American Statistical Association* vol. 80, pp. 294-300.

Smith, H. F. (1938). "An empirical law describing heterogeneity in the yields of agricultural crops," *Journal of Agricultural Science* vol. 28, pp. 1-23.

Zeimetz, K. A., Dillon, E., Hardy, E. E. and Otte, R. C. (1976). "Using area point samples and airphotos to estimate land use change," *Agricultural Economics Research* vol 28, pp. 65-74.