# SAMPLING FROM A FLUID POPULATION

Benjamin F. King, Florida Atlantic University
220 S.E. 2nd Avenue, Ft. Lauderdale, FL 33301

KEY WORDS: Elusive population, location sampling, weights, design effect

## I. Introduction.

This paper deals with some of the problems that arise when sampling from a population that is impossible to frame. That is, there is no useable list from which last-stage sampling units (in this case, individuals) may be selected. The motivation is a survey (still in progress) called the Department of Labor Workplace Literacy Assessment. The target population is all persons who are either registered in JTPA (Job Training Partnership Act) programs or made eligible for UI/ES (unemployment insurance/ employment services) by appearing at local intake offices during a specified calendar time period.

Readers may decide that this problem falls most properly under one of the categories of elusive populations enumerated by Kish [1]. (See also [3].) There is a slight distinction, however, from other situations that suggests the use of the term fluid. Unlike many transient or nomadic groups the target individuals flow through well-defined catchment areas-- i.e., the local offices. These offices can be completely listed, and sampled at earlier stages. Within the local intake offices the individuals are listed, but if one waits to select the final sample from the list it is too late. There is little or no chance of contacting the individuals again for assessment in a way that is likely to be economically feasible. For the survey to be executable in a practical way the sample must be selected as it flows through and the testing done on the spot. (Monetary incentives are used to encourage cooperation.) Another way to describe this situation is as a form of location sampling (Sudman [2]), but it is distinct from the usual problem (e.g., shopping mall sampling) in that the population is clearly defined. (We cannot precisely describe the population found at shopping malls in reference to all consumers, except to say that they are the kind of people who go to shopping malls.)

To summarize, we are dealing here with a well-defined target population that passes through a well-defined set of locations during a well-defined period of time. The elusiveness comes from our having to grab the sample as it flows through because we will probably never have another chance at interviewing after the first encounter.

This paper will address the following problems:

1. Obtaining the necessary information for calculation of exact selection rates for the last stage of sampling.
2. The sensitivity of reported statistical measures (e.g., ratios, totals, standard errors) to errors in the calculation of selection rates and, accordingly, the sample weights used in design-based estimation. Is it worth spending a lot of money to achieve item 1, above?
3. Analysis of the design effect. How much of the deff is due to lack of control over selection at the last stage? Tying in with item 2, will biases prevent our knowing how well or how badly off we are?

## II. Basic Design of the Survey.

The Workplace Literacy Assessment consists of two sample surveys--one of the JTPA enrollees, and a separate and independent survey of UI/ES participants. The goals of this paper can be achieved by discussing either one of the surveys, hence we shall describe the details of the JTPA study. The sampling approach was a conventional stratified, multi-stage cluster design with selection probabilities proportional to size.

At the first stage the states were grouped into nine strata corresponding to Department of Labor regions or sub-regions and two states were selected by pps with replacement from each stratum. In the eighteen first-stage "hits" there were fourteen distinct states selected, hence four states were selected twice. The unit of selection at the second stage was the Service Delivery Area (SDA), usually corresponding to a county, city, or town within a state. For each first-stage hit the SDAs in the selected state were arrayed and four were systematically selected with pps. Most SDAs consisted of a number of JTPA centers where program intake occurs, thus it was necessary to perform a third-stage selection of one location within each selected SDA (again with pps). Finally, at the last stage the plan was to administer the literacy assessment to 56 "randomly" selected program enrollees.

For sampling at the final stage each local office was issued two randomly selected days out of the work week and also a randomly selected time of the day. The intake office was instructed to administer the assessment to the first new client to arrive after the chosen time. Firm refusals were replaced by the immediately following arrival. For certain offices where prior size information indicated that this procedure would not yield 56 completed cases in the calendar time available instructions were given for assessing two clients per randomly chosen day instead of only one. If one overlooks the bias of nonresponse and the effect of stopping the sampling as soon as the target 56 assessments are attained, the overall selection rate for this last stage can, in principle, be calculated. We know the rates of selection of days and hour, and the local offices were carefully instructed to keep counts of all eligible clients entering the premises on assessment days. Elaborate instructional materials and training sessions with SDA personnel in all states were employed (at non-trivial expense), but it was not possible to monitor the execution of the selection procedures and the recording of the intake counts at the local offices, with the exception of a handful of spot checks. Thus it is quite possible that, even if the effects of nonresponse and shut-down of sampling are assumed away, the calculated probabilities of last stage selection will be seriously in error.

Any errors in selection probabilities will be translated into errors in the weights applied to individuals in the analysis of the data. We propose to investigate the effects on estimation of these possible errors in weights.

## III. Analytical Approach

The technical note in the appendix shows the effect of a change in weighting on the estimated mean of a variable y. It can be seen that if the new weight is applied to cases in the sample with average y-value greatly different from that of the remaining cases, then the effect on the estimator is great-- an intuitively obvious result. It follows that the most serious errors in estimation will occur when the most extreme y-values have erroneous weights.

The approach taken in this paper is to start with a benchmark set of weights, assumed to be correct, and accordingly, the benchmark estimates of the population mean and the standard error of the estimator that are obtained by using those weights. Then the weights are perturbed according to the following model for random error generation:

$$\log_e(\epsilon_{hjk}) = \beta(\bar{y}_{hjk} - \bar{y}) + u_{hjk}$$

$$\text{New weight}_{hjk} = \text{weight}_{hjk} \cdot \epsilon$$

The subscripts identify the stratum, the psu, and second-stage cluster to which the weight is applied. The model assumes (as is true in this survey) that the weight is constant for all cases in a given last-stage cluster. The disturbances, u, are assumed to be distributed as Normal($\mu$,$\sigma$). The process of perturbation generates 500 random values of the error in weights for each cluster for given parameter values $\beta$, $\mu$, and $\sigma$. The average values of the estimated mean and standard error over the 500 replications are compared with the benchmark values to estimate the average biases resulting from erroneous weights under that particular model specification.

## IV. A Simulated Survey Outcome.

When this paper was conceived it was anticipated that the results of the actual survey would be available for the analysis of the effects of weighting error. Unfortunately, at this writing the survey is not complete, although eventually the analytical approach described above will be applied to the real data. For preliminary analysis, a set of simulated data was generated as follows:

1. A statistic of primary interest in the study of labor force literacy will be a test score, scaled to have a mean of 500 and a standard deviation of 50. Thus, for the 72 last-stage clusters (8 for each of 9 strata) mean scores for the 56 individuals contained therein were simulated in a manner consistent with the overall mean and variance. It was further assumed in the simulation that the between-strata component accounts for 20 percent of the total variance. No attempt was made to include in this simulation the effects of homogeneity within the first stage clusters-- our conjecture after the perturbation analysis is that it would not have much effect on the results.

2. The benchmark weights were based on the actual selection probabilities down through the drawing of the local offices. The final factor in the weight was based on the assumption of cooperation of 56 respondents in each cluster and the most recent update of estimated cluster size after the selection was made.

3. The simulated stratum means ranged from 469.35 to 541.18. When the weights were applied to the simulated test scores, the resulting benchmark values of the mean and standard error were 499.5 and 1.23. (The s.e. is estimated by the Keyfitz paired selection method.)

V. Results.

For this presentation the perturbation process was applied using 10 different sets of values for the error model parameters. The parameter $\mu$ determines the median value in the simulation of the lognormal error $\epsilon$. For the values of $\mu$ equal to -.69, 0, and .69, the corresponding median error factors are, respectively, .5, 1, and 2. When the second parameter, $\sigma$, is equal to .17 and $\mu = 0$ this implies that the probability is about .95 that the multiplicative error factor is between .7 and 1.4. For $\sigma = .35$, the .95 probability interval for $\epsilon$ is .5 to 2. The third parameter in the simulation is $\rho$, the correlation between the log error and the cluster mean y-value.

*COMPARISON OF DISTRIBUTIONS OF NEW AND OLD WEIGHTS*

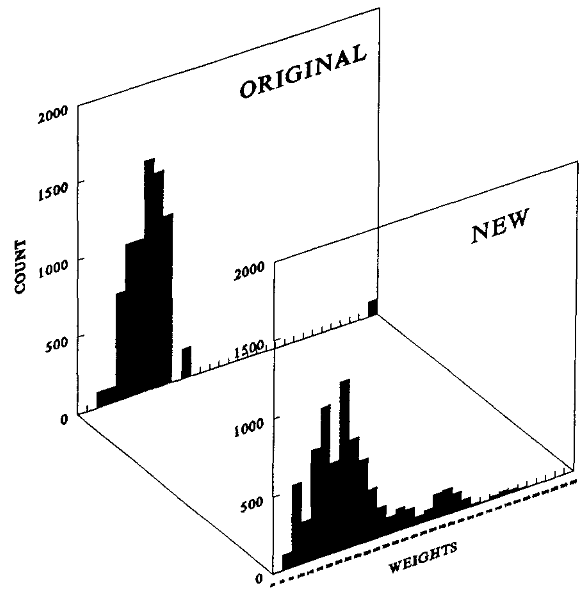


*Figure 1*

In Figure 1, above, we compare the perturbed weights with the original weights in one of the 500 replications for $\mu = 0, \sigma = .17$, and $\rho = 0$. It can be seen that the original weights contain one outlier due to an unforeseen large change in the estimated size of a certain last-stage unit. The effect of this particular error simulation is to increase the spread of the main body of the original distribution while eliminating the outlier.

The table below shows the average values of the estimated mean and standard error for the error simulation runs.

RESULTS

| $(\mu,\sigma,\rho)$ | Avg $\bar{y}$ | Avg s.e. |
|---|---|---|
| (0,.17,0) | 499.6 | 1.31 |
| (0,.35,0) | 499.6 | 1.57 |
| (-.69,.17,0) | 499.5 | 1.30 |
| (0,.17,.5) | 501.6 | 1.35 |
| (0,.17,.9) | 503.4 | 1.34 |
| (0,.17,.9) | 502.8 | 1.40[a] |
| (0,.35,.5) | 503.8 | 1.56 |
| (.69,.17,0) | 497.7 | 1.18[b] |
| (0,.35,.9) | 507.4 | 1.51 |
| (0,.35,.99) | 508.3 | 1.48 |
| Benchmark | 499.5 | 1.23 |

[a] If $|y_{hjk} - \bar{y}| < 20$, $\epsilon=0$.

[b] If not in Stratum 1, $\epsilon=0$.

Consider the first three rows of the table above where $\rho = 0$. We see that there is little effect on the mean value of the estimated mean, even in the case where the median error factor is .5. The estimated standard error, however, is increased as a function of increasing $\sigma$.

The row marked by footnote indicator b is a special case where the only multiplicative errors allowed are for cases in the first stratum (mean = 474.05). In that situation, with $\mu = .69$, the overall average estimated mean is driven downward to 497.7, and curiously the average standard error goes below the benchmark. This situation might correspond to a real world case where procedures were violated in one region of the country due to a breakdown in training or some such explanation.

It can be seen that the serious biases in the average estimated mean occur when there is positive correlation between the log error in the weights and the values of the variable y. Footnote a indicates a case where the error was only allowed to be nonzero for large deviations of y from its mean.

The question is, however, "Why should we expect errors in the weights due to careless bookkeeping, poor training, etc. to be correlated with the values of the variable under study-- in this case literacy of the clientele in a JTPA office?" The possibility of such a situation in the real world seems to me to be far-fetched.

## VI. Conclusions.

If one can accept the simulated survey test scores as realistic, then it would appear that the results are insensitive to the errors in the weights except under extremely pathological conditions that are unlikely to occur. Tentative recommendations are:

1. Do not spend too much money on getting accurate counts of the flow through the catchment areas in the sample.
2. Ratios appear only to be affected in bizarre situations that are unlikely to be encountered in practice.
3. For estimating totals one should use ratio estimation instead of blowup factors based on the selection rate.
4. The effects of the errors in weights on the estimated standard error and the estimated deff are in the conservative direction, i.e., overestimates.

This report is only a practice run for the analysis with the actual data from the survey. I shall be surprised, however, if the results of the extended study are very different for these preliminary results.

## REFERENCES

1. Kish, L. (1988), "A Taxonomy of Elusive Populations", Proceedings of the Survey Research Methods Section of the American Statistical Association.
2. Sudman, S. (1980), "Improving the Quality of Shopping Center Sampling", Journal of Marketing Research, XVII, 423-31.
3. Sudman, S., Sirken, M.G., and Cowan, C.D. (1988), "Sampling Rare and Elusive Populations", Science, 240, 991-6.

## TECHNICAL NOTE

Consider a weighted mean

$$\bar{y} = \frac{(\Sigma_1\bar{y}_1 + \Sigma_2\bar{y}_2)}{(\Sigma_1 + \Sigma_2)}$$

$$= W_1\bar{y}_1 + (1 - W_1)\bar{y}_2$$

where $\Sigma_1$ is the sum of weights for Subsample 1 and $\Sigma_2$ is the sum of weights for Subsample 2, with the subsample means defined accordingly.
$W_1$ is thus the proportion of the total sum of weights that is applied to the first subsample.

Suppose that an error was made in calculating the weights for Subsample 1 and that the sum of erroneous weights for that part of the sample is $a\Sigma_1$. Thus the new weighting factor for Subsample 1 is

$$bW_1 = \frac{a\Sigma_1}{(a\Sigma_1 + \Sigma_2)}$$

which implies that

$$b = \frac{a(\Sigma_1 + \Sigma_2)}{(a\Sigma_1 + \Sigma_2)}$$

The new (erroneous) weighted mean is

$$\bar{y}* = bW_1\bar{y}_1 + (1 - bW_1)\bar{y}_2.$$

It follows that the change in the weighted mean resulting from the erroneous weight is

$$\Delta_{\bar{y}} = \bar{y}* - \bar{y}$$

$$= (b - 1)W_1(\bar{y}_1 - \bar{y}_2)$$

$$= \Delta_{W_1}(\bar{y}_1 - \bar{y}_2).$$