SAMPLE WEIGHTS FOR COVERAGE: A STOCHASTIC INTERPRETATION

Charles D. Cowan, Opinion Research Corporation
Susan Ahmed, National Center for Education Statistics

Keywords: Inference, Estimation, Maximum Likelihood

Abstract

The derivation and development of weights for sample surveys has come almost exclusively from the sample survey tradition. Yet the forms of sampling that are most commonly used have direct interpretations as realizations of stochastic events arising from well defined probability distributions. This paper attempts to develop weights for estimators from sample surveys using maximum likelihood methods applied to the likelihood functions for different sampling designs. Of particular interest is a derivation of stages of weights analogous to those used to correct for coverage and nonresponse problems in sample surveys.

1.0 Introduction

This paper is philosophical in nature. It considers two approaches to the same problem: the derivation and application of weights as a method for estimation and as an adjustment for the various factors that impact the estimation process in statistics. The two approaches are the classical sampling approach to estimation and maximum likelihood estimation.

This discussion is intended to be a review of current practice in sampling, placed in the context of probability theory. We hope to show the correspondence between the two approaches to estimation and how for certain classes of distributions the two approaches lead to the same results, but via different paths. The paths are what make this investigation interesting, because they force the statistician to consider what assumptions are being made regarding how the data has arisen and whether the procedures used adequately reflect real life.

In particular, we attempt to deal with sample weighting as an adjustment for coverage problems in the sampling frame, where coverage problems are dealt with as a stochastic event. This in turn leads us to consider models for the process of sampling and simultaneously for the development of the frame used for sampling. The models lead us to consider a number of issues related to bias, variance, and the assumptions that go into our estimates.

This paper will look at some of the literature related to model based estimation, coverage, and relationships between the two approaches, and then consider some specific models for sampling from a finite population and models related to how frames might be constructed.

2.0 Literature Review

The approach taken in this paper crosses several broad areas related to survey sampling: weighting for coverage, the theory of inference in survey sampling, post-stratification, the EM algorithm. We will not attempt to give a comprehensive review of all of these topics, but rather will point out a few of the references in the literature which are most relevant to our discussion.

Rao and Bellhouse (1989) give an overview of the theoretical foundations of inference from survey data, including a discussion of the design-based vs model-based approach to survey sampling. Sarndal (1978) gives a comprehensive comparison of the design vs model-based approach. Brewer and Sarndal (1983) discuss six approaches to enumerative survey sampling covering both classical and model based approaches to inference. Smith (1976) reviews the historical joint development of survey design and finite population inference. All of the above papers refer to the works of Godambe, Royall, and Hartley and Rao, all of whom attempted to link the survey sampling approach to estimation to the classical approach. Godambe (1966) considered the likelihood function arising from drawing a random sample from a labelled finite population and found that the likelihood was flat and thus provided no information for estimating a total Y. Hartley and Rao (1968, 1969) and Royall (1968), by ignoring the labels, arrive at a likelihood function based on the hypergeometric and multinomial distributions which is informative and show that the customary estimator for the population total:

$$Y = N\ (\Sigma\, y_i\ /n)$$

is essentially the maximum likelihood estimator. They extend their approach to regression estimators, some special cases of unequal probability sampling and two stage sampling, and Bayesian estimators. Our discussion of models for sampling and the resulting estimators, in many respects, parallels the approach of Hartley and Rao. Royall takes a similar approach and, in addition, mentions briefly the case of poststratification where the sizes of the poststrata are known.

Not a great deal has been written on sample weights for coverage. Groves in his book on Survey Errors and Survey Costs (1989) has a chapter on "Costs and Errors of Covering the Population". He refers to making poststratification weighting adjustments to adjust for undercoverage. Often coverage adjustments are considered as a special case of nonresponse adjustments. The most common approach to adjusting for coverage problems is through poststratification or raking ratio estimation using iterative proportional fitting (Deming & Stephan (1940), Bishop, Fienberg, and Holland (1975)) when only marginal population totals are known. Kalton (1983) discusses weighting adjustments for coverage using raking ratio estimation and discusses briefly its relation to log linear models. Cox and Cohen

(1985) discuss poststratification adjustment where sizes of the poststrata are known as an adjustment for both nonresponse and undercoverage.

A comprehensive discussion of the EM algorithm and its wide range of applications is given by Dempster, Laird, and Rubin (1977). In our discussion of weighting for coverage, the EM algorithm is used as an iterative procedure for obtaining maximum likelihood estimates of population cell sizes in a manner similar to the use of iterative proportional fitting.

3.0 Sampling and Estimation under Different Situations, and Corresponding Models

This section will consider different common sampling methods and develop the corresponding models to account for both the sample design and issues in coverage. While we consider both the situation where we use simple random sampling and stratified random sampling, we avoid any discussion of clustered sampling so as to keep the models tractable.

3.1 Simple Random Sampling Without Replacement

In the simplest situation, where we have perfect coverage of the population, the frame perfectly lists the entire population, but we may not have available on the frame the information we require. So we draw a sample from the frame, typically using simple random sampling without replacement (srswor). To make our discussion more concrete, let's consider a simple example: our frame is the population of teachers, and we want to estimate the total number of math teachers in the population of all teachers. And no discussion is complete without the introduction of confusing notation, and so we have:

N = Number of teachers of all types in the population
T = Number of math teachers in the population
n = the sample size
t = the number of math teachers observed in the sample

N and n are known and fixed, t is observed as a result of sampling, and T is to be estimated.

The sampling statistician develops an estimator by considering the probability that a member of the population/frame is selected, n/N, and uses this probability to weight each member of the population to make him "represent" others not sampled but in the population. The criterion used to develop an estimator is that it be minimum variance and unbiased, or minimum mean square error (and so possibly biased, but the extent of the bias is small and the gain in variance reduction is larger than the bias squared). Derivation of such an estimator is a standard exercise early in any complete sampling text.

The probabilist considers the probability distribution for the set of observations collected, in this case the hypergeometric distribution. Such a distribution for this case might look like:

$$p(t) = \frac{\binom{T}{t}\binom{N-T}{n-t}}{\binom{N}{n}}$$

We know N and n, and we observe t, the number of math teachers in the sample; we want to know T, the number of math teachers in the population. The function above can be taken to be a likelihood function with T as the unknown parameter for which we are attempting to discover the most likely value. We can take the derivative of the above function with respect to T and calculate the maximum of the function to find the maximum likelihood estimator of T. Unfortunately, the function above is very difficult to manipulate, and we wind up resorting to indirect approaches or approximations. One approach is to calculate the ratio L(T)/L(T-1) and set it equal to unity. The point at which the ratio is equal to unity is approximately where the maximum of L(T) occurs. The ratio L(T+1)/L(T) works equally well, though it gives a slightly different answer since both are approximations because of the integer nature of T. The first ratio gives an MLE of:

$$\hat{T} = t \frac{(N+1)}{n}$$

and the second gives a similar but different value. Stirling's approximation applied to all the factorials in the combinatorics gives us:

$$\hat{T} = t \frac{N}{n}$$

which is the traditional estimator found in sampling theory.

Now we make the problem slightly more complicated. Suppose the frame does not have perfect coverage, but that we have some external information that fixes the population size. We have to consider two new totals as part of the estimation process. These are:

M = the number of teachers of all types on the frame (while N remains the number in the population), and

F = the total number of math teachers who are on the frame

and we have two events with which we have to be concerned: the event of being included on the frame, which we take here to be a stochastic event, and the event of being sampled. We now have to consider conditional likelihoods: a teacher cannot be sampled unless he has been included on the frame. We can now write the likelihood function as:

$$L(T,F) = \frac{\binom{T}{F}\binom{N-T}{M-F}}{\binom{N}{M}} \times \frac{\binom{F}{t}\binom{M-F}{n-t}}{\binom{M}{n}}$$

$$= L(T|F)\ L(F)$$

The left half of the likelihood assumes that F is now an observed value, and this half of the likelihood deals with the probability of the event of getting onto the frame. In fact we do not know the value of F, how many math teachers are on the frame, and this becomes a nuisance parameter since it is necessary to find the MLE for T. To rid ourselves of this nuisance, we solve first for the MLE of F, and then substitute this MLE into the likelihood L(T|F) to solve for T. This sequential procedure is commonly used in problems where parts of the likelihood are unobserved and other parts are "sufficient" in the formal sense for estimation. This procedure is also reminiscent of the EM algorithm in that we solve for an expected value first, and then use this value to solve for the primary parameter of interest.

The solution to the problem above, deriving an estimator for T, is:

$$\hat{T} = \hat{F}\frac{N}{M}$$

and F is estimated from the second half of the equation by:

$$\hat{F} = t\frac{M}{n}$$

Substituting the estimator for F into the estimator for T, we get:

$$\hat{T} = t\frac{N}{n}$$

which is the traditional estimator and does not rely on the size of the frame. What it does rely on is an assumption that all teachers have an equal probability of getting on the frame, and an equal probability of being sampled, whether they are math teachers or some other type. This point is crucial in understanding why a probabilistic/maximum likelihood approach is useful. The development of the likelihood functions forces us to explicitly state the assumptions underlying the selection process. Consideration of the coverage of the frame as a stochastic process also forces us to think about whether the frame can be adequate as a representation of the population. If the assumptions underlying the likelihood function are not believable, then estimates from the frame will not be unbiased for the population.

Another point to consider is that even in a very simple problem like this one, we have brought in outside information to the estimation process. We sample from the frame, but the weighting information uses the population size, which is a number not available from the frame nor the sampling process. It has to come from another source and must be incorporated as a parameter in the stochastic process to be used. These two points will be considered again in a later example.

3.2 Stratified Sampling

Stratified sampling leads us to consideration of two special situations. The first is the situation where there is information available on the frame that can be used to both partition the sample and also for use in establishing parameters in estimation. In this case we can write down a likelihood function, such as the product multinomial or the multivariate hypergeometric, which has parameters corresponding to the sizes of the strata, always for the frame and usually for the population. When the sizes are available for both the frame and the population, the problem of estimation becomes just the conditional likelihood problem described at the end of the previous section. In this case the assumption is made that the probability of being part of the frame is different in each of the strata.

A situation that is only slightly more complicated is found when the stratum sizes for the frame are known, but the stratum sizes for the population are not known, and the overall population size is known. Using the conditional likelihood, weights are derived for the part pertaining to the frame, and then a proportional adjustment is made to the totals which result to force them to add up to the population total. Again, the assumption is made that membership on the frame is a stochastic event, but as an event that happens with equal probabilities across all the strata. Since we have less information available for estimation, we have to restrict the number of parameters to be estimated.

3.3 Control Totals for Post-Stratification

The final part of the discussion has to do with the situation where information is not available at the time the sample is drawn (i.e. the frame does not have the information necessary to partition the sample on some key variable), but control totals are available for weighting. The simplest situation is the univariate case, where control totals are available for use in estimation with a simple random sample. The likelihood function in this case is very interesting and informative.

$$L = \prod \frac{\binom{T_i}{F_i}\binom{N_i - T_i}{M_i - F_i}}{\binom{N_i}{M_i}} \times \frac{\prod \binom{F_i}{t_i}\binom{M_i - F_i}{n_i - t_i}}{\binom{M}{n}}$$

In this likelihood, we see that there are no subscripts in the denominator of the right side of the likelihood, indicating that the sample was drawn as a simple random sample without replacement, and so all cases fall into the sample with equal probability. The sample can be arrayed according to the categories of the variable used as a control total, and these are our sufficient statistics. However, the frame cannot be arrayed in such a way since the information is not available on the frame. Even though we do not have the sizes, M_i, to complete the likelihood, we can still complete the estimation process by acting as if we did know these totals. Using the same derivation process described above in the first section, we find that the M_i are included in the estimators for T_i (the number of math teachers) and F_i (the size of the frame for coverage group i), and these values cancel when the estimator for F_i is substituted in the equation for T_i. In our notation we have:

$$\hat{T}_i = \hat{F}_i \frac{N_i}{M_i}$$

$$\hat{F}_i = t_i \frac{M_i}{n_i}$$

and by substitution:

$$\hat{T}_i = t_i \frac{N_i}{n_i}$$

In this case, we have a great deal of information about the coverage variable in the sample, and we know that all cases in the sample were drawn with equal probability, so the additional degrees of freedom in the sample allow us to assume different coverage rates - different probabilities of being included on the frame for the coverage groups - which is why we use the post-stratification adjustment. Even though we do not know the totals for the coverage groups on the frame, we are still able to incorporate the population information into the estimation process.

The final situation to be described is one in which we cannot explicitly write down a likelihood that makes use of all the information we have available. This is the situation where we have two or more control variables for which we know control totals, but do not have these variables available on the frame. We can summarize this situation in tabular form, as seen in Table 1.

In this table, the t_{ij} and the n_{ij} are known, the M_{ij} are all unknown, and the N_{ij} are unknown but sums of these values are available as the control totals. This means that the N_{i+} and the N_{+j} are known, and we would like to use these values in the estimation process.

One common way to think about this problem is to consider the table above as a large four way table with incomplete data, and to use log likelihood techniques. To be consistent in our presentation, however, we will outline a different approach to the problem. Our approach is to write out the complete likelihood as if the N_{ij} were known and the M_{ij} were known, and to use the EM algorithm to estimate the parameters in the likelihood, which in turn will give us our weights.

The EM algorithm is a two stage procedure that allows one to estimate the parameters in a likelihood function in the presence of missing data. The first stage assumes the parameters are known, and uses these parameters to calculate expected values for all the unknown totals in the likelihood function, conditional on what is known about these totals. In this case, we would calculate the expected values of the N_{ij} conditional on knowing the N_{i+} and the N_{+j}, and the expected values of the M_{ij} and N_{ij} - M_{ij} conditional on the expected values of the N_{ij}. The second stage assumes all the data are observed and uses this data and the density function to estimate the parameters of the likelihood function. Since the data are not all observed, the expected values calculated in the first stage are used in place of observed values. This procedure is repeated, the first stage followed by the second stage, until it converges. Estimates of the parameters of the distribution obtained in this way are maximum likelihood estimates.

Using the EM algorithm in this case, we can fit the likelihood and derive estimates and so weights. In this case also, since we had a single sampling rate, but a great deal of information regarding the distribution of the sample by the control variables, we can make estimates of the coverage rates for each of the control variables separately. However, when we make estimates of the coverage rates that are ultimately used for weighting the data, some assumptions have to be made regarding the relationships between the coverage variables. The assumption ultimately made is that the higher order interactions between the control variables are preserved for the population as were observed on the frame, though the univariate sample and population distributions may differ.

4.0 Conclusions

The assumptions discussed in section 3.0 are essentially the same as those which apply when deriving sample weights, and in particular when raking is used to fit a sample to a set of control totals. The purpose of this paper was to show that these assumptions are made explicit when the classical sampling and estimation process is contrasted with maximum likelihood procedures which require definitive statements about how the data observed are generated.

5.0 References

1. Bishop, Y.M., S.E. Fienberg, and P.W. Holland (1975), Discrete Multivariate Analysis: Theory and Practice, The MIT Press: Cambridge, Mass.

2. Brewer, K.R.W., and Sarndal C.E. (1983), "Six Approaches to Enumerative Survey Sampling". In Incomplete Data in Sample Surveys, vol. 3, Proceedings of the Symposium. W.G. Madow, and I. Olkin (eds.), Academic Press: New York.

3. Cox, B.G. and Cohen, S.B. (1985), Methodological Issues for Health Care Surveys, New York: Marcel Dekker: New York.

4. Deming W.E. and Stephan, F.F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known", Ann. Math. Stat., 11, 427-444.

5. Dempster, A.P., Laird, N., and Rubin, D.B. (1977), "Maximum Likelihood From Incomplete Data Via the EM algorithm", JRSS-B, 39, 1-38.

6. Godambe, V.P. (1966), "A New Approach to Sampling from Finite Populations", JRSS-B, 28, 310-328.

7. Groves, R.M. (1989), Survey Errors and Survey Costs, Wiley & Sons: New York.

8. Hartley, H.O., and Rao, J.N.K. (1968), "A New Estimation Theory for Sample Surveys", Biometrika, 55, 547-557.

9. Hartley, H.O., and Rao, J.N.K. (1969), "A New Estimation Theory for Sample Surveys II" In New Developments in Survey Sampling. N.L. Johnson and H. Smith (eds.), Wiley-Interscience, 147-169.

10. Kalton, G. (1983) Compensating for Missing Survey Data, Institute for Social Research, Univ. of Michigan: Ann Arbor.

11. Rao, J.N.K. and Bellhouse, D.R. (1989), "The History and Development of the Theoretical Foundations of Survey Based Estimation and Statistical Analysis", 1989 Proceedings of the American Statistical Association - Sesquicentennial Invited Paper Sessions.

12. Royall, R.M. (1968), "An Old Approach to Finite Population Sampling Theory", Journal of the Am. Stat Assoc., 63, 1269-1279.

13. Smith, T.M.F. (1976), "The Foundations of Survey Sampling: A Review", JRSS-A, 139, Part 2, 183-204.

This article was prepared by Susan Ahmed in her private capacity. The views are those of the author, and no official support by the U.S. Department of Education is intended or should be inferred.

Table 1: Crosstabulation of Sample, Frame, and Population Data using Control Information

Math Teachers

		Control 2					
		1	2	...	j	...	c
Control 1	1						
	2						
	...						
	i				t_{ij}		
	...						
	r						

Total Sample

		Control 2					
		1	2	...	j	...	c
Control 1	1						
	2						
	...						
	i				n_{ij}		
	...						
	r						

Frame Totals

		Control 2					
		1	2	...	j	...	c
Control 1	1						
	2						
	...						
	i				M_{ij}		
	...						
	r						

Total Population

		Control 2					
		1	2	...	j	...	c
Control 1	1						
	2						
	...						
	i				N_{ij}		
	...						
	r						