Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

THE CONFIDENCE OF AN OPTIMAL SAMPLING SIZE BASED ON PREVIOUS DATA

Charles Fleming and Richard McGuinness[1]

ABSTRACT

For a multivariate survey which is based on simple random sampling, the problem of calculating an optimal sampling size becomes one of solving a stochastic programming problem in which each constraint corresponds to a bounded estimate of the variance for a commodity. The problem is stochastic because the set of data collected from a previous survey makes the components of each constraint random variables; consequently, the calculated size of a sample is itself a random variable and is dependent on the quality of that set of data. By means of a Monte Carlo technique, an empirical probability distribution of the optimal sampling size can be produced for finding the probability of the event that the prescribed precision will be achieved. Corresponding to each set of previously collected data, there is an optimal size and allocation across strata. While reviewing these over several consecutive periods of time, it may be possible to identify troublesome strata and to see a trend in the stability of the data. The review may reveal an oscillatory pattern in the sizes of the samples that might have evolved over time due to the dependency of one allocation on another.

KEY WORDS: Optimal Sampling Size; Convex Programming; Survey Design.

1. FORMULATION

When relying on voluntary cooperation for obtaining statistical information, the uncertainty that someone will respond to a question as well as the anticipated variability of the gathered data play important roles in determining the size of a sample. The problem of finding a suitable size of a sample can be completely described by a mathematical statement. But as rigorous as the mathematics may seem to appear, the optimal size of a sample is stochastic. It is not the solution of a deterministic problem. The optimal size of a sample is actually a random variable. It consolidates into one number the random characteristics of the estimates, the random characteristics of the estimated variances, and the random characteristics of estimating the cooperation of those who are selected to do the survey. It is a function of these important components, and its probability distribution can serve two purposes. It can be used in choosing a sampling size in terms of confidence intervals. It can be used in a quality control program.

In order to improve the efficiency of a survey, a statistician might obtain information about several subjects in a single interview rather than doing separate interviews for each subject. Thus the sampling size for a multivariate survey must account for the components of the survey for each variate simultaneously. A more efficient survey can be designed if a statistician has some foreknowledge of the results of a survey. To that end, stratifying the list of elements that exist in the population is a common practice. Another is to use information from a previous survey to describe the characteristics of the current population, provided that it is safe to assume that the current population is not too different from the older population. A principal goal of designing a survey leads one to make use of those practices in order to produce precise estimates

[1]Charles Fleming and Richard McGuinness, National Agricultural Statistics Service, U.S. Department of Agriculture, 1400 Independence Avenue, SW, Washington, DC 20250, USA.

with the least cost.
Although, techniques like stratifying the list of a population help in achieving that end, the precision of the estimates is extremely sensitive to the presence of misclassified elements of a list. Moreover, the problem of defining an optimal stratification in a multivariate survey is still an unsolved problem. But, given a stratification, the problem of determining an optimal sampling size can be formulated as a convex programming problem as follows:

$$\min \sum_{h \in strata} c_h n_h$$
$$such\ that$$
$$\sum N_h(N_h - \hat{D}_h n_h)\frac{s_h^2}{\hat{D}_h n_h} \le (CV\hat{\tau})^2$$
$$0 \le n_n \le N_h$$

where the summation runs over all the strata, c_h is the cost of conducting the survey in stratum h, n_h is the optimal sampling size for stratum h, $\hat{\rho}_h$ is the estimated rate of response from the people who are selected to do the survey, and $\hat{\tau}$ is the estimated quantity of a commodity. In the multivariate problem, there is a constraint for every commodity. By making n_h large enough, the estimated precision of $\hat{\tau}$ will improve until its CV will fall below a prescribed value but at the expense of increasing the cost of the survey. This convex programming problem is actually a stochastic programming problem because it contains random variables; therefore, its solution is a random variable which has associated with it a probability distribution. It is this probability distribution which will be used to assess the quality of the previously collected data.

The key idea that the natural variability of historical data affects the reliability of a calculated optimal sampling size should always be kept in mind. Knowing the extent to which the random variation of the data affects the optimal size of a sample provides two useful benefits. First, this knowledge permits a statistician to construct a probability distribution of the optimal sampling size so that he can regard the decision of specifying the sampling size in terms of reviewing confidence intervals. In other words, knowing the probability distribution of the optimal size, it is possible to gauge the probability that the survey will achieve the stipulated precision of an estimate for a given sampling size. The other benefit permits a statistician to assess the effectiveness of the chosen method of stratifying the list by inspecting the consistency of the optimal size at each stratum over the course of time.

For a multivariate survey, only numerical methods can solve the convex programming problem shown above. These same numerical techniques were used in a Monte Carlo scheme to produce many hundreds of optimal sizes per stratum, such that each calculation came from a set of randomly generated variates representing the standard error, nonresponse, and the expanded total. The total set of numbers produced by these calculations provided the basis for creating an empirical probability distribution of the optimal sampling size.

2. EMPIRICAL DISTRIBUTION

Before proceeding with the Monte Carlo simulation of the optimal sampling size, the possible existence of a correlation between the rate of response that someone will participate in a survey, the sample variances, and the estimates must first be ascertained in order to verify the validity of the theoretical assumption that they are independent. To look for the existence of a correlation, plots were made like the one shown in Figure 1. In that figure, the coefficient of variation and nonresponse from a combination of thirteen quarters and eleven strata related to the Quarterly Agricultural Labor Surveys which had been conducted in Minnesota show no

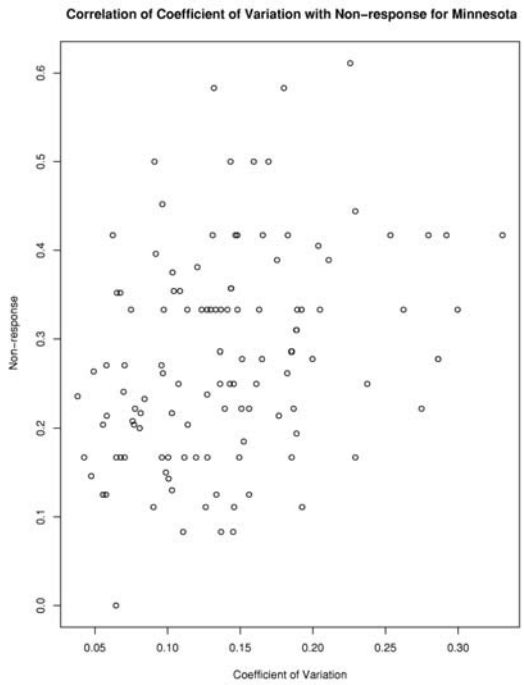

Figure 2. There is no apparent correlation between the coefficient of variation and the nonresponse.

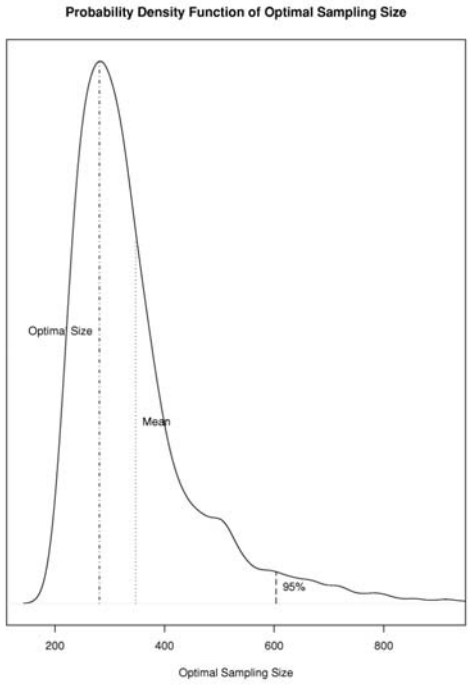

Figure 1. From the empirical probability distribution, the 95th percentile of achieving the prescribed CV occurs at 603.

correlation. As a result of inspecting the other different possible plots, the Monte Carlo technique was performed in which the sample variance, population total, and nonresponse were treated independently.

Although the exact distributions of the sample variance, nonresponse, and the expanded total are not known, there are good candidates with which to approximate them. For example, it was assumed that the square of the standard error, s^2, is distributed as a X^2 distribution. Also, because the range of the nonresponse is confined to the interval [0,1], it was assumed that the nonresponse follows a beta distribution. In order to design a stable Monte Carlo program in which the generated random variates are all nonnegative and behave realistically, a beta distribution was used for describing the population total rather than a normal distribution.

Due to the complexity of the problem, it is not possible to derive a theoretical probability distribution of the optimal sampling size. Instead we found, through the use of a Monte Carlo technique, an empirical probability distribution for one of the routinely conducted Quarterly Agricultural Labor Surveys. Produced from the set of 1,000 calculated optimal sizes using randomly generated values of the sample variance, nonresponse, and population total, the empirical probability density distribution appears in Figure 2. Our attention is drawn to the skewed shape of the distribution and, more importantly, to the conspicuously heavy right tail. Both features suggest that achieving a target CV will require a sampling size larger than what one would expect from the deterministic problem.

From the empirical distribution, we can see that the sampling size needed to achieve the prescribed CV for this particular survey at a level of 95 percent is 603. On the other hand, the probability which is represented by the area under the curve of achieving the prescribed CV given the deterministic sampling size of 281 is only 35

percent. Even for the simulated sampling size of 337, the probability of achieving the prescribed CV is 60 percent, a little more than flipping a fair coin.

To explain the origins of the heavy tail, we need to look at some detailed results. The ones displayed in Table 1 show the results of the Monte Carlo simulation for the Agricultural Labor Survey conducted in Minnesota. We notice that due to the heavy tail of the distribution, the simulated optimal size is larger than the deterministic optimal size in every stratum except one. Analogous to the goodness-of-fit X^2 statistic is our

Table 1.

Stratum	Population Size	Estimated Nonresponse	Optimal Size	Simulated Optimal Size	CHI2
50	9362	.278	5	6.2	.28
55	12423	.167	9	12.78	1.59
70	5020	.286	10	12.44	.59
75	612	.250	1	1.54	.30
79	28668	.146	42	61.65	9.38
85	787	.354	10	11.28	.16
92	16198	.167	81	99.05	4.02
93	2001	.167	31	34.63	.42
94	834	.250	40	45.26	.69
95	245	.333	29	32.41	.40
96	23	.250	23	19.45	.54
Total			281	336.95	18.42

CHI2 statistic, which appears under the column heading CHI2. This ad hoc statistic measures the influence of the heavy right tail on the sampling size by taking the sum of the quotients of the squared differences between the deterministic and simulated optimal sampling sizes by the deterministic size. It is a measure of the quality of the previously collected data which is reflected in the empirical probability distribution by the heavy right tail.

In Table 2, we see the tabulation of CHI2 for thirteen consecutive surveys spanning four years, from the most recent, denoted by 13, to the oldest, denoted by 1. As one scans over the thirteen surveys, the large values of CHI2 associated with surveys 6 and 2 stand out and suggest that the origins of the large values should be investigated. It was discovered upon studying the circumstances surrounding survey 6 that the actual sampling size was too small. Evidently, the precision of the estimates was satisfactory in the previous surveys, encouraging the Agency to try to use a smaller sampling size. Rather than using the traditional size of 350, a sampling size of 318 was used instead. Had a stochastic programming method been devised and available for use then, an application of the method would have revealed that the quality of the data from the previous survey was actually problematic and that a larger sampling size was needed instead of a smaller size. In response to that experience, we see a general decline in the values of CHI2 across all strata over time up to survey 13, the most recent survey, due to better stratification of the list and due to larger sampling sizes.

There is a pattern not only across time but across strata. Strata 55, 79, and 92 have consistently high CHI2, apparently due to misclassification of elements of the list concerning which strata they should belong to. We can see a steady decline in CHI2 in all strata after survey 8 to the present as new definitions of strata and larger sampling sizes are used. Likewise, for strata 50, 75, and 96, CHI2 is consistently low, indicating that the allocation of the sample for those strata is adequate and that the definitions of the strata are good. These strata have traditionally provided good estimates with few outliers.

Neither the skewness nor the kurtosis offered much help. Whereas the CHI2 statistic produced patterns which

Table 2.

Stratum	13	12	11	10	9	8	7
50	.28	10.35	1.41	.92	1.24	.01	2.45
55	1.59	3.62	.23	20.42	10.21	34.43	4.52
70	.59	2.61	.22	9.11	3.45	15.28	1.32
75	.30	0	0	.82	.85	.46	.03
79	9.38	6.34	.54	.46	2.58	.31	.44
85	.16	1.05	0	3.24	.61	2.67	.54
92	4.02	7.40	.30	25.81	5.92	105.43	32.03
93	.42	2.57	.06	10.61	3.75	20.23	1.27
94	.69	1.14	.16	8.83	4.65	16.16	4.57
95	.40	1.34	0	.96	.83	1.86	.51
96	.54	.02	.01	.54	.66	.91	.23

Stratum	6	5	4	3	2	1	
50	0	0	0	0	27.44	3.60	
55	28353.95	14.22	.14	.47	6569	4.28	
70	8508.47	25.57	1.10	.47	16.10	6.31	
75	870.25	.12	.34	.02	32.05	1.30	
79	49508.01	115.92	1.65	.92	192.57	17.18	
85	135.15	19.34	.92	.11	46.06	2.39	
92	23800.76	20.16	1.15	.89	.23	9.73	
93	1695.78	4.63	.42	.27	901.41	2.33	
94	157.3	3.92	.10	.90	2378.55	.37	
95	15.48	5.26	.16	.11	8.35	.94	
96	0	.74	.41	.24	1.08	.17	

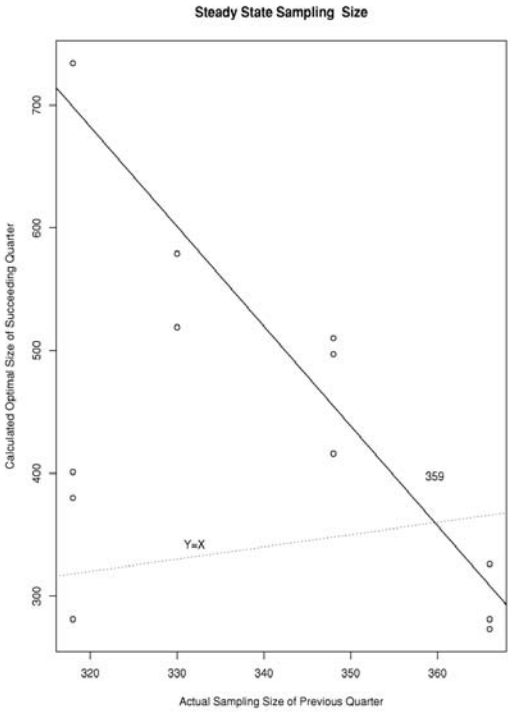

Figure 3. The relationship between the calculated optimal sampling size and the actual sampling size used by in the previous survey.

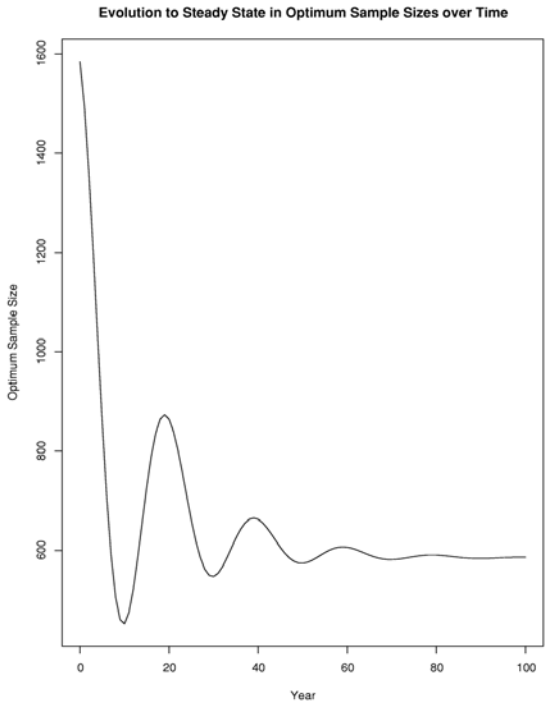

Figure 4. Qualitative depiction of the evolution of the optimal sampling size to a steady state sampling size over time.

could be explained when viewed over time and strata, no discernible patterns were observed in the tabulation of the skewness and kurtosis of the empirical probability distribution other than that they are almost always negative for stratum 96 and largely positive for the other strata.

3. STEADY STATE SAMPLING SIZE

The investigation of survey 6 brought to our attention the obvious correlation between sampling size of the current survey and the one for the previous survey. This correlation between the optimal sampling size and the actual size of the previous survey is shown in Figure 3. Superimposed on that graph is the least squares fitted line showing the linear relationship more clearly. The line shows that there is an inherent relation between the calculated optimal sampling size and the actual size of the sample used in the preceding survey. The existence of such a relationship should not be surprising because the size of the sample of the preceding survey directly affects the variability of the data gathered by the current survey, which will eventually become the historical data for determining the sampling size of the succeeding survey. In other words, the effects of a chosen sampling size propagate through subsequent surveys as in a Markov process. A large sampling size at one time may induce a smaller sampling size in the next as a result of producing good data with high precision, while conversely, a small sampling size may induce a large sampling size in the next survey as a result of producing noisy data with low precision. The optimal size of a sample depends on the size of the preceding survey sample. The sampling sizes from survey to survey are, therefore, not independent.

Consequently, if one is not careful, an oscillatory pattern could develop in the sampling sizes from survey to survey like the one depicted qualitatively in Figure 4.

Eventually, by a process of trial and error, a steady state sampling size will be reached, in which case the calculated optimal sampling size and the actual size of the sample will converge to the same number. By recognizing that there exists an inverse relationship between a previous sampling size and the optimal sampling size for the succeeding survey, we may assert that the steady state optimal sampling size will be the one in which the sampling sizes of the previous survey and the succeeding survey are equal. Rather than use the method of trial and error over many years to find a steady state sampling size, it can be determined by the intersection of the dotted line Y=X, with the curve relating the dependency of the optimal sampling size with the historical data. These intersecting lines appear in Figure 3 in the case of Minnesota for which the steady state optimal sampling size occurs at 359.

Another approach to determine a steady state sampling size is a method by which the variances from previous surveys are pooled. This pooled variance is then used in the calculation of an optimal sampling size. For a univariate survey in which the estimates, sample variances, rates of response, and stratification of the list are relatively stable over time, the approach of pooling variances might provide an easy alternative approach. But for a multivariate survey and especially one in which these components are not necessarily stable, then the pooling of information about a survey can be done transparently by solving the convex programming problem, so that a steady state sampling size can be found directly as we have already shown.

From the functional relationship of the sampling size on the estimates, sample variances, rates of response, and stratification given in the convex programming formulation, it is possible to arrive at a steady state optimal sampling size. Because the optimal sampling size is a random variable, its probability distribution can serve the purpose of not only finding the likelihood of achieving the prescribed precision of an estimate for a given sampling size, but by studying its skewed shape and heavy right tail by means of a statistic like the CHI2, it is possible to monitor the quality of the data across strata and over time.