Reducing respondent burden for repeated samples

By Robert D. Tortora

Repeated sampling from a frame to make estimates in different subject matter areas may increase the burden on survey respondents. Two methods for determining selection probabilities are presented that reduce the burden. The first method takes into account the previous number of contacts of a sampling unit; the second, the length of interview on previous contacts.

Keywords: Respondent burden, sampling theory, selection probabilities.

INTRODUCTION

Reducing respondent burden must be a goal of any agency that repeatedly contacts the public to obtain data. This goal has been emphasized by the Office of Management and Budget (OMB) (2). OMB is particularly concerned with reducing the burden on private citizen respondents that are associated with data collection by the Federal Government. The Statistical Reporting Service (SRS) wants to reduce the burden on persons contacted repeatedly for voluntary responses to USDA surveys. More SRS survey work is being placed on a probability basis, and new methods should be presented to reduce the burden while retaining the probability characteristic of the surveys.

The principal factors that make up the burden placed on a particular respondent participating in surveys are: (1) the time it takes to complete an interview; and (2) the number of contacts during some time period, for example, a calendar year.

The burden on particular respondents is sometimes reduced by excluding from the next survey sample all persons surveyed previously. However, this practice violates the concept that each sampling unit must have a known positive probability of selection. Total respondent burden can be reduced by curtailing sample sizes, thus reducing the number of contacts made. However, this approach provides no relief for the respondents included in a smaller sample, and it may reduce precision (that is, increase the variance) to an unacceptable level.

I will address the two factors listed above for the case when the same frame L is used to select samples for K surveys covering different subject matter areas during some time period. Separate procedures for each factor will be suggested. Both procedures reduce respondent burden by increasing the probability of selection of sampling units not contacted previously, while decreasing the probability for those contacted previously.

In agricultural surveys, sampling frequently is done using lists of farm operators. For example, in each State, SRS is developing one list of farm operators for surveys. Each list contains the operator's name and address, and information about the types and numbers of livestock species, poultry, and crop acreage. Procedures should be instituted to reduce the burden placed on individual operators on this list.

Now let us assume that where the selection probabilities are altered to reduce respondent burden, the surveys are ad hoc; and no rotation of sampling units occurs. Further, the surveys are large enough so that estimators based on probability proportional to size (pps) without replacement sampling show only small gains over pps sampling with replacement. That is, we assume the proportion of the sample size to the population size (nN^-1) is small. In the method, we use the concept of replacement sampling. We can then account for the previous individual burden on a farm operator by adjusting selection probabilities.

SELECTION PROBABILITIES FOR NUMBER OF CONTACTS

When samples are drawn from a frame to estimate different subject matter areas, and the data collection does not involve lengthy interviews, the number of contacts represents the most important factor contributing to respondent burden.

Proposed Method

Suppose that the frame L is used to select K samples of size n_i, i = 1, . . . , K for K different subject matter surveys. Further, suppose that the first several surveys (P of them) have been conducted.

Let \gamma_j, j = 0, 1, . . . , M be the number of sampling units selected j times in the first P surveys. Note that M can be greater than P because we assume with replacement sampling. For the P+1st sample to be selected, we want to associate with each member of the frame a probability that reflects (inversely) the number of times that unit has been previously selected. So, if a sampling unit were not selected in any of the first P surveys from frame L, it should have a greater probability of being selected than a sampling unit which has been selected...
one or more times in any of the first \( P \) surveys. Therefore, we want to assign probabilities \( \pi_0, \pi_1, \ldots, \pi_M \) to each of the frame members that have been selected 0, 1, \ldots, \( M \) times such that \( \pi_j \) is the probability that each element that has been selected \( j \) times will be selected on the \( P+1 \)st survey. Thus, the probabilities \( \pi_j \) should be monotone decreasing and should satisfy the following equation:

\[
\pi_0 > \pi_1 > \ldots > \pi_M. \tag{1}
\]

Also, to have a probability mass function to sample from, the following equation holds:

\[
\sum_{j=0}^{M} \gamma_j \pi_j = 1. \tag{2}
\]

To satisfy equation (1) and to allow the selection procedure to determine how much more likely the sampling units that occurred \( j \) times are to be in the \( P+1 \)st survey than the sampling units that occurred \( i \) times, \( j < i \), the following set of equations must hold:

\[
\begin{align*}
\pi_0 &= a_0 \pi_1 \\
\pi_1 &= a_1 \pi_2 \\
\vdots \\
\pi_{M-1} &= a_{M-1} \pi_M 
\end{align*}
\tag{3}
\]

In equation (3) the \( a_i \)'s are constants greater than 1. Rewriting each probability in terms of \( \pi_M \) gives the following:

\[
\begin{align*}
\pi_0 &= a_0 a_1 \cdots a_{M-1} \pi_M \\
\pi_1 &= a_1 a_2 \cdots a_{M-1} \pi_M \\
\vdots \\
\pi_{M-1} &= a_{M-1} \pi_M 
\end{align*}
\tag{4}
\]

That is,

\[
\pi_i = \left( \frac{M-1}{a_i a_{i-1} \cdots a_1 \pi_M} \right) \pi_M
\]

Substituting these values into equation (2) gives the appropriate probability associated with all sampling units in the frame that have been selected \( M \) times in the first \( P \) surveys:

\[
\begin{align*}
\pi_M &= (\gamma_0 a_0 a_1 \cdots a_{M-1} + \gamma_1 a_1 a_2 \cdots a_{M-1} + \cdots \\
&\quad + \gamma_{M-1} a_{M-1} + \gamma_M) \pi_M 
\end{align*}
\tag{5}
\]

Using equations (5) and (4) gives the desired probabilities associated with the other units in \( L \).

Up to now, the choice of the \( a_i \)'s has been rather arbitrary because only equations (1) and (2) must hold. However, if we assume that the first \( P \) surveys were conducted so that each element of \( L \) had equal probability of selection \( \pi \), then equation (1) should be rewritten as:

\[
\pi_0 > \pi_1 > \ldots > \pi_M. \tag{6}
\]

We would choose the \( a_i \)'s so that equation (6) holds.

**Example One**

The following example illustrates the method of determining the selection probabilities \( \pi_j \). Suppose the frame \( L \) contains \( N = 200 \) sampling units. Four surveys have been conducted and, for definitiveness, simple random samples of size 20, 20, 25, and 30 have been selected. Of the 76 (distinct) sampling units that have been selected, 1 was selected on 3 surveys, 17 on 2 surveys, and 58 on 1 survey. Thus \( \gamma_0 = 124, \gamma_1 = 58, \gamma_2 = 17, \) and \( \gamma_3 = 1 \). Let \( \pi_3 \) be the probability of selecting any element that was selected three times in the four surveys.

For equation (6) to hold, where \( \pi = 0.005 \), and if three other conditions occur, equation (4) implies \( a_0 = 10/7, a_1 = 7/5, \) and \( a_2 = 50 \). The three conditions are these: the probability associated with any unit that was not selected \( (\pi_0) \) is 100 times \( \pi_3 \); the probability associated with any unit selected once \( (\pi_1) \) is 70 times \( \pi_1 \); and the probability associated with any unit that was selected twice is 50 times \( \pi_3 \).

Equation (5) becomes:

\[
\begin{align*}
\pi_3 &= (17311)^{-1} = 5.77 \times 10^{-5} \\
\pi_0 &= 100/17311 = 5.77 \times 10^{-3} \\
\pi_1 &= 70/17311 = 4.04 \times 10^{-3} \\
\pi_2 &= 50/17311 = 2.89 \times 10^{-3}
\end{align*}
\]

If we examine the special case of \( a_i = a \) for all \( i \)'s, equation 5 becomes:

\[
\pi_M = \left[ a^M (\gamma_0 + a^{-1} \gamma_1 + \ldots \\
+ a^{-(M-1)} \gamma_{M-1} + a^{-M} \gamma_M) \right]^{-1}
\tag{7}
\]

When \( a_i = 1 \) for all \( i \)'s and each frame unit has probability of selection \( 1/N \), equations (4) and (5) imply that \( \pi_i = 1/N \) for all \( i \)'s. Thus, we can attain a "minimum" spread on our new probabilities of selection for the remaining surveys by taking \( a_i = a \) just larger than 1.

Also, note that for large \( N \) and \( a > 1 \), an approximation to \( \pi_M \) is given by:

\[
\pi_M \approx \left[ a^M (\gamma_0 + a^{-1} \gamma_1) \right]^{-1}
\tag{8}
\]

For example, taking \( a = 1.1 \) in equation (8) gives a "minimum" spread as \( \pi_3 = 3.86 \times 10^{-3}, \pi_0 = 5.66 \times 10^{-3}, \pi_1 = 5.14 \times 10^{-3}, \) and \( \pi_2 = 4.68 \times 10^{-3} \).
SELECTION PROBABILITIES
FOR GROSS INTERVIEW TIME

In many instances, the number of contacts made with a
person is less important than the total time of the
interview. This characteristic is often the case when
detailed expenditure-type data are collected. The follow-
ing approach resembles those shown in the preceding
sections.

Proposed Method
Again, suppose K surveys are conducted from the
frame L. Let \(m_i, i = 1, \ldots, K\) be the average gross time
of interview for each survey. Define for each frame unit
the total time of interview for the first \(P\) surveys by:

\[
t_i = \sum_{j=1}^{P} m_{ij}, \quad i = 1, \ldots, N,
\]

where \(X_{ij}\) is the number of times the \(i\)th unit is selected
in the \(j\)th survey. Now group all frame units with the
same \(t_i\)'s, and, of course, include a group for those ele-
ments with \(t_i = 0\). Let \(\gamma_j\) be the number of frame units
in the same group wherein the index \(j\) runs over the
groups. Assume \(\gamma_j\) takes the possible values 0, 1, \ldots, \(M\).
Thus, the \(\gamma_j\)'s are defined similarly to the approach in
“Selection Probabilities for Number of Contacts.” Similar
calculations can be performed to obtain selection
probabilities based on gross interview time. The example
below illustrates this procedure.

Example Two
Suppose the situation of example one, but \(N = 600\).
Let the average time of interview for the four surveys be
\(m_1 = 1, m_2 = 0.5, m_3 = 3,\) and \(m_4 = 0.5\), respectively.
Group 0 contains the number of elements that were not
selected in any of the 4 surveys and, hence, each has
\(t_i = 0\) associated with it. Let Group 1 contain those
frame units that have \(t_i = 0.5\) associated with it. Groups
2, 3, 4, 5, and 6 are those frame units with a \(t_i\) of 1, 1.5,
3, 3.5, and 4, respectively. The number of frame units in
each group is \(\gamma_0 = 524, \gamma_1 = 27, \gamma_2 = 19, \gamma_3 = 5,\)
\(\gamma_4 = 19, \gamma_5 = 5,\) and \(\gamma_6 = 1\). Using equation (8) and tak-
ing \(a_i = 1.1\) for all \(i\)'s gives:

\[
\pi_7 = \left[ (1.1)^7 (524 + (1.1)^{-1} 27) \right]^{-1} = 9.35 \times 10^{-5}
\]

It follows that the remaining new selection probabilities
are these:

\[
\pi_0 = 1.82 \times 10^{-3}, \quad \pi_1 = 1.66 \times 10^{-3}, \quad \pi_2 = 1.51 \times 10^{-3},
\]
\[
\pi_3 = 1.37 \times 10^{-3}, \quad \pi_4 = 1.25 \times 10^{-3}, \quad \pi_5 = 1.13 \times 10^{-3},
\]
\[
\pi_6 = 1.03 \times 10^{-3}.
\]

ESTIMATION

In the assumptions in the “Introduction,” we
restricted ourselves to ad hoc surveys. Since we have also
assumed that \(nN^{-1}\) is small, the estimator for the popula-
tion total that is used with the above selection probabili-
ties is as follows:

\[
\hat{Y} = n^{-1} \sum_{i=1}^{n} \pi_i^{-1} y_i
\]

with variance

\[
V(\hat{Y}) = n^{-1} \sum_{i=1}^{N} \left( \pi_i^{-1} y_i - \bar{Y} \right)^2
\]

This equation becomes zero if \(\pi_i \propto y_i\); that is, \(\pi_i = y_i/\bar{Y}\).
Thus, the success of the sampling schemes (in terms of
more efficient estimation) rests on whether or not the
“sizes” suggested are proportional to the item totals \(y_i\).
At least for the list of farm operators discussed above
and for the situations described in the examples, this
may not be unreasonable. However, empirical studies
should be conducted to measure the proportionality.

If \(nN^{-1}\) is not small, sampling without replacement
is used with Horvitz-Thompson estimators (3) of the
form:

\[
\hat{Y}_{HT} = \sum_{i=1}^{n} \pi_i^{-1} y_i
\]

to estimate the population total.

Here, recall that \(\hat{Y}_{HT}\) may have an estimated nega-
tive variance. However, as outlined in (3), various tech-
niques have been proposed to minimize this problem.

CONCLUSIONS

Two methods of determining selection probabilities
that reduce respondent burden have been presented.
The first method gives probabilities based on the prior
number of previous contacts. The second method is
based on the gross interview time involved in previous
contacts. It should be clear that the schemes presented
can be generalized in two ways. First of all, the exten-
sion to stratified designs is easily made by computing
selection probabilities for each stratum. Secondly, if
one develops a more complex index of respondent
burden, it can be applied as suggested in “Selection
Probabilities for Gross Interview Time.”

These methods of determining selection probabili-
ties were developed for two reasons:

- To reduce respondent burden.
- To retain the probability concept behind the
  survey.
Thus, a fundamental concept behind probability sampling must be retained; each sampling unit in L must have a known positive probability of selection for each of the K surveys. The methods presented ensure that this concept is retained.

The procedures presented can be used in the total planning of surveys that are to be conducted by an agency. Suppose the "important" surveys are identified. By selecting the samples for these surveys at the beginning of the reference time period, analysts can apply procedures to the potential number of contacts or the potential total interview time per respondent.

Finally, further research is called for because empirical studies must be conducted to evaluate the efficiency of the pps estimator that would be used and that could be applied to stratified samples.

REFERENCES


IN EARLIER ISSUES

When designing a survey, the technical sampler usually tries to achieve the minimum sampling error consistent with a given cost or he tries to obtain a specified sampling error at a minimum cost. This is the principle on which much of the research work on sampling methods has been based. . . . To apply the principle, costs and sampling errors of alternative methods are needed. But the principle of minimum sampling error per dollar is not entirely satisfactory, for it deals with only one of two major components of error in survey results and furnishes no answer to the question of how much different degrees of accuracy are worth.

Earl E. Houseman
Volume 1, Number 1, p. 3
January 1949