SOME ALTERNATIVES FOR THE TREATMENT OF FIRST PHASE TELEPHONE NUMBERS IN A WAKSBERG-MITOFSKY RDD SAMPLE

Charles D. Palit, University of Wisconsin-Madison
Johnny Blair, University of Illinois-Urbana

Statement of the Problem.

The Waksberg-Mitofsky RDD telephone design is a two-stage cluster sample. In the first stage, banks of telephone numbers are selected with probabilities proportionate to the number of working residential telephone numbers in the banks. In the second stage, calls are made into each selected bank until some predetermined cluster size of identified working residential telephone numbers is reached.

Carrying out the first stage requires that one make a determination of residential status for a simple random sample of telephone numbers. This is done by calling each sample number to determine its status. Most of the sample can be sorted into residential or non-residential categories quite easily. However, some telephone numbers, even after numerous callbacks, cannot be identified as residential or non-residential, because they are never answered. The sampler is, therefore, obliged to seek an alternative means of resolving the ambiguity. Sometimes this can be done by contacting the local telephone company, and inquiring after the status of the particular phone number. Often this is a successful strategy, but often the phone company is unwilling to cooperate and will not provide the information. When the telephone companies will not cooperate, then other methods of dealing with the problem must be found.

The alternative treatments of these numbers under these conditions and the statistical implications of the treatments are the subject of this paper.

Some alternatives for dealing with the telephone numbers whose residential status cannot be determined by calling them are:

a) keep those banks in the sample and generate secondary phone numbers for them

   1) Strictly keeping the banks associated with never answered phone numbers in the sample is equivalent to changing the definition of the target numbers to include both residential and never answered numbers. When the true target is residential numbers, the use of this option will contribute substantially to the variance of the cluster size for households, increasing the possible bias in the ratio estimators which would commonly be used with this data, and making it more difficult to control the survey's sample size.

b) discard those banks with never answered numbers

   1) The effect of discarding the banks associated with never answered numbers depends on the treatment of never answered numbers in the second phase of the sample design. If never answered numbers are retained, then the cancellations which give the Waksberg-Mitofsky design the same selection probability for each sample element will not occur. If the never answered numbers are replaced in the second phase, then a significant portion of the universe may be unreachable by the sample design.

c) determine the residential status of the number by contacting the phone company or using directories. While this is the cleanest alternative, it is not always possible, at reasonable cost.

d) An alternative to the keep or discard approach is to retest the block (cluster) for which no determination could be made. The retest procedure requires that a different element be sampled from each no answer cluster, and a new call made to determine the status of the replacement element. Should this retest fail in the same way, then a second retest may be made and so on. After some predetermined number of retests, if the status of the cluster was still unresolved, the cluster would be:

   1) discarded
   2) kept

We use the model described next to understand and evaluate the effect of the different strategies.

Model:

Let P be the true proportion of working residential numbers in the cluster, let p be the proportion of residential numbers which can be identified by the calling procedure, and q be the proportion of non-residential numbers in the cluster which can be identified by the calling procedure.

$$p+q =< 1$$

and the difference

$$t=1-(p+q)$$

is the proportion of the numbers in the cluster which cannot be identified as residential or non-residential by the calling (screening) procedure.

Now let
$$t1=P-p,$$
and
$$t2= (1-P) - q,$$
then
$$t1 + t2 = t.$$

The situation is illustrated in Figure 1.

FIGURE 1

Possible Result of Test			Real Status	
p	Identifiable Residential Non-Identifiable Resident	t1	Residential phone	P
q	Non-Identifiable Non-Res. Identifiable Non-Resid.	t2	Non-Resid. phone	1-P

The probability of identifying the first element selected from the cluster is p. If t=0 then p=P otherwise p<>P, but for many values of t, p can be considered as an approximation to P, and correspondingly 1-P is approximated by q.

In practice, many blocks (or clusters) have a non-zero value for t. For some this shows up as a no answer classification, while for others, particularly where t is small, the existence of a non-zero t passes unnoticed.

Effect of Retaining the No-Answer Blocks.

The probability of selecting any particular block of 100 suffixes for testing on a given trial is inversely proportional to the number of blocks of 100 in the universe, i.e., $1/N$, where N = number of blocks of 100 phone numbers in the sampling frame. If the no-answer blocks are retained, then the probability of selecting a cluster on any given trial is

$$\left(P_i + t_{2i}\right)/N$$

where i denotes the i^{th} cluster in the population. By Lahiri's (1951) agrument, the chance that the cluster will end up in the sample is

$$k\left(P_i + t_{2i}\right)/\sum\left(P_i + t_{2i}\right)$$

where k is the number of clusters selected into the sample and the sum is overall blocks in the population. If the ultimate cluster size is n and the no-answer numbers are retained in the second stage, then the selection probability for each residential or no-answer number is

$$nk/\left(\sum_i\left(P_i + t_{2i}\right)\right)$$

or

$$\frac{\text{(number of PSU's)(size of ultimate cluster)}}{\text{Total no. of resid no.'s and no-answer non-residential no.'s in population}}$$

In practice, the value of the denominator is usually unknown, but since it is a constant for all sample elements, its value is often not required for the production of population estimates.

If the no-answer phone numbers are replaced in the second stage of sampling, then the selection probability for each residential number is no longer a constant, and in practice may not be determinable. If it cannot be determined, then it would be unwise to use this design since it would be extremely difficult to find suitable estimators.

Effect of Discarding the No-Answer Blocks.

If the no-answer blocks are not retained, then the probability of selecting a block on any given trial is p_i/N

where p_i may itself be a random variable, since the probability of a residential phone being answered will vary. For convenience we will temporarily consider p to be fixed. Later we will relax this assumption. Under the assumption that p is fixed for each PSU, the chance of a PSU being included in the k PSU's selected for the first stage of the sample is

$$p_i k\Big/\left(\sum_i p_i\right)$$

and the selection probability of a residential answering number is

$$nk\Big/\left(100\left(\sum_i p_i\right)\right)$$

provided that no-answer phone numbers are replaced in the second phase of the sampling process.

If we consider p_i as a random variable in its own right, then the situation becomes more complex; the indications are that the expected value of the selection probability will remain constant.

Usually using the expected value of the selection probability in an estimator will still yield a good population estimate. What changes is the variance estimate. The expression for the variance estimate becomes more complex to reflect the fact that the world is more complex and to compensate for the fact that the actual probability is a random variable.

We are continuing work on this aspect of the problem and will report our results later.

Effect of the Retest Rule with No-Answer's Discarded After Retest.

When we use a one try retest rule, the retention probability is the probability of success on the first try plus the probability of a second try times the probability of success. This becomes,

$$\begin{aligned} p + tp &= p + (t1 + t2)p \\ &= P\text{-}t1 + (t1 + t2)(P\text{-}t1) \\ &= P\text{-}t1 + t1P - t1 ** 2 + t2P - t1t2 \\ &= P + (P(t1+t2) - (t1+t1**2+t1t2)) \\ &= P + (P(t1+t2) - t1 - t1(t1+t2)) \\ &= P + (t(P\text{-}t1) - t1) \end{aligned}$$

Clearly the one retry rule produces a retention probability somewhere between the other two alternatives, and one which may well be closer to P. Again the cancellation which makes the overall selection probability of sample elements constant is unlikely to occur.

Effect of the Retest Rule with No-Answer's Discarded After Retest.

In the case where the first phase cluster is retained if the second try produces another no-answer number, the retention probability becomes:

$$p + t(p+t).$$

Even if the second phase no-answer's are retained, the selection probability for each residential or never answered number is not likely to be a constant.

Discussion.

If the cost is not prohibitive, weeding out the non-household no-answer's using information from telephone companies, etc. is the best alternative. For those times when this cannot be done, Table 2 provides a summary of what we see as possible alternatives.

Of the other options listed in this table, the most appealing currently are alternatives 2a and 3.

In alternative 2a, if we can ignore the random native of p_i the effect of discarding the no-answer blocks in the first stage coupled with replacing no-answer phone numbers in the second stage seems at first glance to be exactly what we would like to happen. The sample is self-weighting and we have good control on the sample size.

The simplicity of this design is very appealing. Its great drawback is the ease with which a bad calling strategy can pass undetected. Since the population is defined in terms of answering phones, it is easy to overlook the effect of the non-coverage of non-answering residential phones on the quality and value of the data.

In terms of the population of answering phone numbers, the response rate is calculated with

$$\frac{\text{No. completed interviews in the sample}}{\text{No. completed + No. of non-response in sample}} \times 100$$

This response rate is equivalent to the response rate calculation for the upper bound for the response rate in the design 3, which is the response rate for the population of all residential numbers, both those answering and those non-answering.

In terms of Figure 1 both measures ignore the non-answering residential phone numbers in the population of all phone numbers. The number of non-answering residential phone numbers in the sample, as well as the proportion of residential non-answering phones in the population of all phone numbers is often unknown. Absence of information on the non-answering residential phone numbers presents a substantial problem in the measurement of quality for any RDD sample. Usually the problem becomes apparent in the computation of response rate.

For this design the effect of non-answering can be expressed as a coverage problem, or as a response rate problem. We believe that in the interest of consistency the effect of non-answering residential numbers should be included in the response rate computation. Either way, we stall out on the usual problem of too little information for an exact solution.

Common solutions to this problem provide for setting an upper and lower bound for the response rate or generating a single number using assumptions as in the CASRO solution. In a simple RDD design, an estimate for the lower bound of the response rate is

$$\frac{\text{No. of completed sample} \times 100}{\text{No. of completed sample + No. of non-response + No. of no answer}}$$

We cannot duplicate this formula exactly for the whole sample in design 2a, but we can duplicate it for part of the sample, and use the result to estimate the equivalent upper bound for the rest of the sample. The estimate is based on the sample results up to but before the no-answer numbers are replaced. The definitions of terms are, therefore, exactly the same as for a simple RDD sample.

TABLE 2

EFFECT OF ALTERNATIVES FOR THE TREATMENT OF UNANSWERED TELEPHONE NUMBERS IN A WAKSBERG-MITOFSKY RDD SAMPLE

			Evaluation of the Options	
Option	First Stage	Second Stage	Selection Probabilities	Coverage
1	Determine HU status from the telephone company		equal	unaffected
2	Discard unanswered numbers & replace with new numbers			
		a) Discard unanswered numbers & replace with new numbers	equal	may be reduced
		b) Do not discard unanswered numbers	unequal	may be reduced
3	Keep unanswered numbers in the sample			
		a) Discard unanswered numbers	unequal	unaffected
		b) Count unanswered numbers toward the cluster size	equal	unaffected
4	Test a second number from the cluster; if the second number is unanswered:			
	A) Keep it			
		a) Count unanswered numbers toward the cluster size	unequal	unaffected
		b) Do not count unanswered numbers toward the cluster size	unequal	unaffected
	B) Discard and stop testing in that cluster		unequal	unaffected
	C) Discard and continue testing in the cluster until ... ?		unequal	unaffected

Summary.

Never answered numbers can present a real problem for the sampler designing RDD samples—Waksberg-Mitofsky or otherwise. A seemingly clean and simple way of treating the problem is to allow never answered phone numbers in the first phase to select blocks or clusters for the second phase even though this may contribute to the inefficiency of the design.

Discarding all never answered numbers in the first phase and replacing never answered numbers in the second phase is an acceptable procedure, provided care is taken to compute an estimate of the response rate which takes the possible existence of never answered residential phones into account.

A third approach, that of retesting the first phase blocks whose test numbers are never answered numbers may result in variable selection probabilities, but under certain conditions may well be an appropriate procedure. We are continuing our investigation of the behavior of these procedures.

References.

Des Raj (1968) Sampling Theory, McGraw-Hill, Inc., New York.

Hansen, Hurwitz, and Madow (1953) Sample Survey Methods and Theory, Vol. 1, Wiley, New York.

Lahiri, D. B. (1951) A Method of Sample Selection Providing Unbiassed Ratio Estimates, Bull International Statistical Institute, 33.

Waksberg (1978) "Sampling Methods for Random Digit Dialling," Journal of the American Statistical Association, 73, pp. 40-46.