OPTIMAL BATCHING IN SAMPLE SURVEYS

Benjamin King, University of Washington

1. Introduction

In many sample surveys, members of the target population cannot be identified in advance of the selection of sampling units. An example is the population of "persons 65 years of age or over." In sampling from a geographic area, one can only select households and then enquire at each sample address whether persons 65 or over are present. This process is called screening.

In the case of screening for the elderly there are Census counts available that make it possible to fairly accurately determine the number of households that must be selected in order to obtain the desired number of persons 65 or over. In many other screening situations, however, prior information is not available and one must rely heavily on subjective judgment. How many households must one screen, for example, in order to obtain 100 interviews with widows (female) under 35 years of age? Or how many screeners are required in order to find a sample of 100 persons who call themselves "born-again Christians"?

The nature of field operations in a survey research organization is such that when the required number of members of the target population has been found one cannot simply cut off the screening process at that point. It is well known that field personnel tend to cover the easiest cases first, saving refusals and not-at-homes until the clean-up stage of the field period. For all of the usual reasons of selection bias, a sample that has been cut off cannot be defended as a probability sample from any population that can be easily described. The total number of selected screeners must be completed (up to the point of budgetary restrictions on call-backs). Thus, in a state of great uncertainty about the eligibility rate, i.e., the probability that a screened sampling unit will produce one or more members of the target group, one runs the risk of having to perform unnecessary screening if the number of original selections is too large. If the field procedures require interviewing on the spot when a screener is successful, then one must also pay for unnecessary interviews. At the other extreme, a screening sample that is too small to produce the required number of eligibles means that an additional sample will have to be launched with all of the concomitant start-up costs. It is probably fair to say that survey organizations tend to be conservative, in the sense of sending out a number of screeners that is on the high side, rather than risking undershooting of the target group.

An approach that is designed to remedy some of these problems and provide greater control over field operations is called batching. The incremental cost of selection of sampling units is usually very small. Thus, a large number, B, of batches of size k is selected in such a way that each of the batches is a probability sample from the frame of sampling units. The B batches constitute a set of interpenetrating, or replicated, subsamples so that any subset of b batches (if the response rate is adequately high) can be pooled and treated as a single probability sample for the purpose of estimation and inference.

A sequence of batches of size k is metered out to the field during the screening process; i.e. if the first batch does not produce sufficient eligibles, a second batch is launched, and so on, until the desired number of eligibles has been obtained. When the report that the required eligibles have been found reaches the main office, field personnel are instructed to complete all cases in the batches presently in the field, but no new batches are launched.

If the batch size k is small, one has accordingly a great deal of control over the field work, and the number of excess screeners and interviews is kept at a minimum. If, however, the start up cost of additional batches is high, there is reason to set k equal to larger values. In short, one must balance the cost of launching new batches against the cost of overrun, or performing unnecessary screening and interviewing. The determination of the optimal value of k, the batch size, is the subject of this paper.

2. The Objective Function

Define

- r = the number of eligible units that is desired;
- n = a random variable, the number of sampling units that must be screened in order to find r eligibles, $(n \geq r)$;
- k = the predetermined batch size;
- b = a random variable, the number of batches of size k required to find r eligibles. $(bk \geq n)$;
- c = the start-up cost of launching each additional batch after the first;
- c_s = the cost of screening each additional unit beyond the n units required to find r eligibles.[1]

The cost associated with batch size k is considered to be a linear function of the number of batches after the first and the number of excess screeners (and possibly excess interviews.):

$$\text{Cost}(k) = (b-1)c + (bk - n)c_s. \qquad (2.1)$$

The expression above can be simplified by defining:

$c^* = c/c_s$, the start-up cost relative to the unit cost of excess screening,

and writing

$$\text{Cost}^*(k) = (b - 1)c^* + (bk - n). \qquad (2.2)$$

With n and b random variables, the objective in choosing k is to minimize the expected cost.

$$\gamma(k) = (c^* + k)E_k(b) - E(n) - c^*, \qquad (2.3)$$

or, ignoring the constants E(n) and c*, to minimize

$$\gamma^*(k) = (c^* + k)E_k(b). \qquad (2.4)$$

3. The Expected Value of b

The subscript k in the expression $E_k(b)$ denotes that the expected value of the number of batches required to find r eligibles depends on k, the batch size. As discussed previously, proper field procedures require that the batch that contains the rth identified eligible (and, correspondingly, the nth screened unit) must be screened completely so that the pooled set of b batches can be treated as a probability sample. It follows that:

$$P_k(b = i) = P([i - 1]k + 1 \leq n \leq ik). \qquad (3.1)$$

Thus,

$$E_k(b) = \sum_{i=1}^{\infty} iP([i- 1]k + 1 \leq n \leq ik)$$

[1] In the case where an interview would be performed on the spot if an eligible were found, c_s contains a factor equal to the cost of an excess interview weighted by the prior probability of finding an eligible respondent.

$$= \sum_{i=1}^{\infty} i\,[P(n \geq [i - 1]k + 1) - P(n \geq ik + 1)$$

$$= \sum_{i=0}^{\infty} P(n \geq ik + 1). \qquad (3.2)$$

4. Evaluation of the Expected Number of Batches

To examine the characteristics of the optimization process, we assume that successive screenings are independent Bernoulli trials, with unknown p, the probability that a screened unit produces an eligible. We call p the eligibility rate.

Uncertainty about p is expressed by means of a beta prior density with parameters r' and n':

$$f_\beta(p|r',n') \propto p^{r'-1}(1-p)^{n'-r'-1}, \quad 0 < p < 1. \quad (4.1$$

It follows that the probability mass function for n, the number of trials to obtain r eligibles, is a beta mixture of Pascal mass functions, and we can write

$$E_k(b) = \sum_{i=0}^{\infty} G_{\beta Pa}(ik + 1|r,r',n'), \qquad (4.2)$$

where $G_{\infty Pa}(\cdot|r,r',n')$ is the right tail area of the beta-Pascal distribution with parameters r,r',n'.[2]

After considerable struggle, a closed-form expression for the evaluation of (4.2) does not appear to be obtainable, hence a computer program has been written to calculate the required tail areas and sum the terms until the incremental contribution to the sum is very small. A check on the adequacy of this approximation is obtained from the fact that $E_k(b)$ for k = 1 is equivalent to

$$E(n) = r(n' - 1)/(r' - 1), \qquad (4.3)$$

the mean of the beta-Pascal distribution for r, r', n'.

Furthermore, there is no neat analytical expression that enables one to differentiate the expected cost (2.4) and solve for the minimizing k, but with a vector of values of $E_k(b)$ it is easy to use a computer routine to search over successive values of k for the minimum cost. As will be shown in the next section, the expected cost function is not perfectly U-shaped, but, rather, may have several local minima.

[2] See, e.g., Raiffa and Schlaifer, Applied Statistical Decision Theory, MIT Press, (1968) pp. 237-241.

5. An Example

A recently encountered real world example is as follows: A large, national organization sells a certain general service to business firms. Interviews are desired with 100 business firms who have purchased a particular type of special service under the heading of the general product. The only way to identify users of the special service is by going to regional offices (approximately 100 of these nationally) and examining records for the selected firms. For the national organization to use the regional facilities, considerable intra-company negotiation and planning is required. The screening of the records at the regional offices requires the use of staff who must be relieved of their usual duties and specially trained. If the initial batch of k screeners is insufficient to provide the necessary firms with the particular service, then subsequent batches would require additional negotiation, planning, and set-up at the regional offices. The national organization wants to avoid having to approach the regional offices any more than absolutely necessary. On the other hand, the screening of records is an expensive and time-consuming operation, and the total amount of screening must be kept within reasonable limits--thus k should not be too large.

There is uncertainty about the exact value of p, but it is generally agreed that, whatever the prior distribution, E(p) is about 0.2, i.e., about one in five screened firms will have the special service of interest. Fig. 1 shows three possible beta prior densities for p, all with the same expectation, 0.2. As the parameters r' and n' increase, the dispersion of the distributions decreases. Thus the distributions represent a range of possible subjective assessments of the eligibility rate before screening, from the

rather vague and informationless case (r'=2, n'=10) to the sharply spiked density about p = 0.2 (r' = 200, n' = 1000).

If the ratio r'/n' is held constant at p = 0.2 and r' and n' allowed to increase to infinity, the prior density approaches a single spike and the corresponding beta-Pascal distribution for n, the number of trials required to produce r eligibles approaches a Pascal distribution with parameter p = 0.2.

With the aim of finding r = 100 eligible units, values of $E_k(b)$ were calculated for each of the parameter pairs (r', n') for successive k from 1 to 1000 using (4.2). Fig. 2 shows a rough sketch of the objective function involving k, $E_k(b)$, and an assumed value of $c^* = 20$; i.e., the launching of a new batch is twenty times as expensive as performing an unnecessary screening. It can be seen that for the relatively diffuse case (r' = 2, n' = 10) the unique minimum is attained with k = 185. For the tighter prior (r' = 20, n' = 100) the minimizing value of k is 146, and for the prior with the smallest variance (r' = 200, n' = 1000), the minimum minimorum is attained at k = 281. Although the exact minimizing values of k differ for the three priors, one can conclude from an examination of Fig. 2 that by choosing k between, say 150 and 250, one would be satisfying all three of the sets of subjective beliefs.

The objective function for the tightest prior (r' = 200, n' = 1000) has much more marked highs and lows than the curves for the more diffuse priors. It can be seen that k = 594 yields almost the same value as does the optimal k = 281; yet there is a local maximum between the two values of k at 438, with a value of the objective function that is about 34 percent higher than the minimum. An intuitive explanation of this phenomenon is as follows: The expected number of trials required to obtain 100 eligibles is E(n) = 502.01. At the optimal value of k, 281, the expected number of batches that will be launched is 2.13--thus the expected total number of screeners released to the field is about 600, and the expected overrun in screening is about 20 percent of the expected number of trials needed. Similarly at k = 594, the expected number of batches is 1.05 and the same argument applies. For k = 438, however, the expected number of batches required is closer to 2, and in the event that the second batch is needed in order to find 100 eligibles the expected overrun is about 75 percent of E(n), thus the higher exp-cted cost for that batch size.

As a further example, the objective function for the three priors is sketched in Fig. 3 for the case c* = 100. As one would expect, with the cost of a new batch relative to an unnecessary screening increased, the optimal value of k moves to the right. Also, in the case of the extreme prior (r' + 200, n' = 1000) the values of the expected cost for the various locally minimizing k decrease as k increases. (The minimizing value k = 599 is the global minimizer although the rest of the function is not shown.) The intuitive reason for the multiple minima and maxima is as discussed above, except that with the higher cost of a new batch there is a greater penalty imposed on the smaller near-optimal batch sizes, accounting for the downward sloping appearance of the wavy curve.

6. Conclusion

This paper demonstrates the feasibility of brute force calculation of expected numbers of batches and minimizing batch sizes, k, for various cost and prior parameter assumptions. Although it is difficult to draw general conclusions from particular examples, we have shown that in a large scale sample survey where the stakes are high and great expense is involved in all stages of the operation it may be worthwhile to make a few computer runs to see the implications of various assumptions that might be made when screening is to be performed without precise knowledge of the eligibility rate.

A curious result is that even when one is quite sure about the value of p (i.e., with a very tight prior) it is not necessarily optimal to send out a single batch with k equal to E(n). If the cost of a new batch relative to the cost of excess screening is high, the best batch size is higher than E(n) in order to allow for the uncertainty about the exact n required. If c* is small, it may be optimal to send out smaller batches.

Note: Copies of the figures mentioned in the text can be obtained from the author, Professor of Quantitative Methods, School of Business Administration, University of Washington, Seattle, WA 98195