Using Order Sampling to Achieve a Fixed Sample Size after Nonresponse

Pedro J. Saavedra[1], R. Lee Harding[1] and Francine Barrington[1]
[1]ICF Macro, 11785 Beltsville Drive, Suite 300, Calverton, MD, 20705

Abstract

There are situations when a study requires a fixed sample size, either for contractual reasons or because the cost of collecting data for too many cases is prohibitive. This makes the preferred practice of oversampling and then adjusting for nonresponse impractical. Under certain conditions a simple random sample can be obtained by randomly sorting the frame and selecting the first n in the random order. This yields a fixed initial sample size, but a variable respondent sample. In a case where potential respondents beyond the targeted number of completes can be approached in sequential order (exhausting contact attempts before going to the next unit), the sampling process can continue until the desired number of completes is obtained. Nonresponse adjustments can then be made as if the combined set of respondents and nonrespondents constituted an initial sample. A similar approach to the one described above could be used to achieve a fixed number of completes using Sequential Poisson Sampling or Pareto Sampling. Here the probability of selection is changed, but the difference may be minimal. Simulations using SRS, SPS and Pareto were conducted to examine this practice.

Key Words: replacements**,** Pareto sampling**,** adjustments**,** simulations**,** sequential Poisson sampling**,** propensity categories

1. Introduction

One difficulty in allocating for unit nonresponse is the inability to predict the response rate ahead of time. The problem is that if one oversamples by too small a number, the respondent sample can be too small, and if one underestimates the response rate, the respondent sample can be too large. The first problem will lead to estimates that do not meet the desired precision and the second will incur in unnecessary costs.

The problem is greater when one of the following conditions exists: 1) there is a contractual requirement as to the number of respondents, 2) the cost of a complete survey is high, 3) the survey is a longitudinal survey with subsequent attrition, or 4) the sampling a entails selecting a fixed number of SSUs (e.g. schools) per PSU.
There are various methods that have been used to fix this problem. They include the use of techniques that do not yield probability sampling, or the use of replicate groups, which are released as one refines the estimated sample size. The use of replicate groups is difficult to manage and requires close monitoring.

The method considered here is order sampling, and is effective only under certain circumstances. First, it calls for an inexpensive screening method (to determine willingness to response and characteristics useful in nonresponse adjustments) and a high cost for a complete survey. Second, it requires a list frame and third, it requires that potential respondents after the initial sample be contacted one at a time. An ideal example is record abstraction, where determining if the necessary information is in the record takes a short period of time, but the actual abstraction takes a much longer period.

1.1 Order Sampling

As mentioned above, the approach considered here is order sampling. In the case of Simple Random Sampling (SRS) this is the most common sampling approach if one has a complete list of the frame. One simply assigns a random number to each unit in the frame, orders the frame using the random numbers and selects the n cases with the lowest random numbers. The use of this approach with SRS or within stratum for stratified random samples has a number of advantages. The first advantage is that if one uses the same random numbers or a function of the two random numbers in more than one cycle or in two surveys with overlapping frames, one can control the overlap of the samples. This method is referred to as sampling with Permanent Random Numbers (PRN).

The Use of PRNs is straightforward for simple random sampling (SRS) or even for stratified random sampling. Each unit in the frame (or in the union of overlapping frames) receives a PRN which can be used across samples. For each sample, the units within a stratum are sorted by PRN and the first n are selected for the sample, where n is the allocation for the stratum. This creates a high overlap of units between surveys. If one wants to rotate the samples, one has merely to apply a linear transformation to the PRN. This is essentially the use of order sampling with PRNs and equal probabilities within strata.

When sampling with PPS is desired, one could use Poisson sampling, assigning a PRN between 0 and 1 to each unit in the sample. Letting $p(x)$ be the desired probability of selection of unit x and $r(x)$ the PRN for x, one can select the sample $S=\{x \mid r(x) < p(x)\}$ and one will have selected a sample where the probability of selection of unit x will be $p(x)$. Unfortunately, this approach yields a variable sample size, and while the expected size will be $\sum p(x)$, the actual size can vary considerably.

Ohlsson (1995) developed a method which used order sampling with PPS, thus allowing the use of PRNs to control overlap. The approach, which he called Sequential Poisson Sampling (SPS), merely sorted the frame (or the stratum) by $r(x)/p(x)$ and selected the first n units, where $n=\sum p(x)$. The probabilities of selection were not exact, but were extremely close to the designated probabilities, so that they could be used for most purposes.

Shortly thereafter, Rosen (1995) and Saavedra (1995), working independently, showed that this approach could be improved by sorting by $(r(x)-r(x)p(x))/(p(x)-r(x)p(x))$. Rosen proved that this formula was optimal among a family of PPS order sampling methods, and called it Pareto Sampling. While the probabilities were not exact, they were closer to the designated probabilities than SPS. Later, Aires (1999) developed an algorithm to calculate exact probability for Pareto Sampling, though Pareto comes sufficiently close to the designated probabilities that the Nieves algorithm is not strictly necessary for many practical purposes.

The concern for a fixed sample size is a very practical one, particularly when there is a contractual requirement for a particular number of completed surveys, when institutions are sampled as PSUs or in establishment surveys when the cost of data collection for each unit is particularly high. The problem with simply insuring that the sampling procedure yields a fixed number of initial units is that nonresponse will immediately reduce the sample size by a variable amount, unless some form of replacement is used. In addition, if nonrespondents are replaced, some form of non-response adjustment is necessary to control for the bias.

1.2 Sequential Replacement with Order Sampling

Ohlsson (personal communication, 1996) indicated that SPS could be used to order the frame, and then to continue to sample until the desired number of respondents is achieved. Consider how this might be done for a simple random sample. The traditional approach would be to oversample, identify weighting categories and then adjust the weights by weighting category. The alternative proposed here would be to order the frame and sample the cases near the beginning of the frame in order, until the desired number of completes is achieved. The sample is then treated as if there had been an oversample counting all nonrespondents (preceding the last sampled case) as if they had been part of the initial sample.

Thus, suppose N=120,000 and n=1,000. Suppose that it takes 1,200 initially sampled units to reach the 1,000th complete. Say there are two categories and we know from the frame which unit is in which category. Suppose we found 600 in each category, but of the 200 nonrespondents, 150 were in category A and 50 in category B. The initial weight would be 120,000/1200=100. But for members of category A we had 150 nonrespondents and 450 respondents, so the weight becomes 100 (600/450) =133.33. For category B the weight is 100 (600/550) or 109.09. These are exactly the same weights as if we had targeted 1,200 and done no sequential replacement. Note that if one uses order sample in this manner it is important not to skip any cases. In practice one can get ahead by attempting to contact the first 1,000 in any order, and then the number of cases left to get 1,000 in sequential order.

The above description works well for a simple random sample and the same process can be used if one uses Sequential Poisson Sampling or Pareto Sampling. However, with Pareto Sampling it is not clear what probability of selection to use in sorting the sample. Let m(x) be the measure of size for unit x. Then $p(x) = nm(x)/\sum m(x)$, however, it is not clear if one should use the intended number of respondents, or the expected number when respondents and nonrespondents are combined. For SPS it does not matter, because the one probability is a linear transformation of the other, so the sorting order is the same. However, SPS is not as efficient as Pareto Sampling, and for Pareto it does make a difference.

Thus one sees that when we sort an order sample with the intention of continuing until one reaches a certain number of respondents, one can calculate the variable one sorts by as:

1) The PRN divided by the measure of size (which is equivalent of the PRN divided by the probability of selection using either n).
2) Pareto Sampling using the intended number of respondents as the n in the formula.
3) Pareto Sampling using the expected combination of respondents and nonrespondents (which may not be easy to estimate beforehand) as the n in the formula.
4) Pareto Sampling using some other n reflecting some sort of conservative compromise.

This approached has been used in a number of surveys at different stages. The most direct implementation of this procedure was in the last several cycles of the EIA-782 petroleum price monthly survey. Replacements at the beginning of the survey were obtained using the second of the approaches listed. After the survey had started there was a concern for continuity and the replacement procedure required a replacement that was similar to the

unit that had dropped out. In addition, similar approaches with simple random samples were also implemented at the last stage of the HUD QC Recertification study.

2. Simulations

2.1 Simulation with Simple Random Sampling

In order to explore the various possibilities, a number of simulations were run using different approaches. The first set of simulations used a data set drawn from the American Community Survey, and used Simple Random Sampling. The expected result is obvious, but since the usual practice is to oversample, it is presented here.

First, all cases without total income were eliminated. Then a self-weighting sample was obtained from the remaining cases. This became the frame for this set of simulations, as well as for a later set. The simulated frame included 66,115 respondents to the ACS, sub-sampled so that each would have the same probability of inclusion in the simulation.
Sex was used as the weighting category, simulating a 70% response rate among males and a 90% response rate among females. Several sets of 10,000 samples were drawn several times, selecting samples with 1,260 initial records, for a mean of 1,006 simulated respondents. A second set of 10,000 samples was drawn with exactly 1,006 respondents in each sample, using sequential replacements. Weights were adjusted as described above. Estimates of total income were compared, using matched pairs, where the matched samples had identical PRN seeds. Three measures of accuracy were used: 1) the actual estimate of mean income (to measure bias), 2) the absolute value of the differences between estimated mean income and mean income from the frame, and 3) the square of the differences between estimated mean square income and mean square income from the frame. As expected, there were no significant differences found in bias, absolute deviations, or mean square deviations between the two approaches.

2.2 Pareto Sampling and the Baseball Data Base

The main simulations using PPS sampling and sequential replacement used baseball data. The frame include over 6,600 major league player/season dyads over seven years, where only players with at least one at least one at bat for the season were included. Records were sampled with probabilities proportional to times at bat. Estimates were made for number of hits and collective batting average (ratio of average number of hits to average at bats).

Four samples were selected using Pareto sampling, with a different n in the formula $p_u=n(s_u/\sum s_u)$. An additional set used Sequential Poisson Sampling and the last used Poisson sampling with an oversample. A total of 10,000 simulations were conducted for each of the six designs. Different response rates were simulated for pitchers and non-pitchers. Pitchers, of course, had a lower batting average than non-pitchers. Pitchers and non-pitchers formed nonresponse categories and different response rates were simulated for each, assigning a 70% response rate to non-pitchers and a 90% response rate to pitchers.

For the fixed sample size order sampling methods (all but Poisson), every unit was sampled in order until 400 simulated respondents had been sampled. The n used in the Pareto formula were the average total sample (576), the total number of respondents (always 400) and two other numbers (800 and 324). Non-response adjustments were conducted using pitcher vs. non-pitcher as a weighting class.

A General Linear Model design was used, absorbing random number seed (i.e. matching by seed) and comparing the six methods. The dependent variables were the actual estimates, the absolute value of the difference of each estimate from the population parameter and the mean square of the difference of each estimate from the population parameter. The results for average number of hits and collective average were, as expected, very similar.

The three methods:
1) Pareto with the expected total sample (576) as the n,
2) Sequential Poisson Sampling, and
3) Poisson Sampling were essentially tied.

All methods yielded on the average an underestimate. However, the Pareto method yielded the mean of means closest to the population mean, though not significantly. The number of hits had an extremely high correlation with the number of at bats, so the procedure was repeated with a second pair of variables – home runs and home run average. The results for the three methods were extremely similar, with Pareto using the expected initial size (including respondents and non-respondents), Sequential Poisson Sampling, and the variable sample size Poisson Sampling (with an oversample) essentially tied.

The estimates for both hits and home runs were fairly accurate, with an average absolute deviation of 0.3 hits (under 0.002 batting average) where the population average was 46.7 hits, and 0.14 home runs (under .0008 home runs per at bat) where the population average was 5.5 home runs.

2.3 Another ACS Simulation

The self-weighting file created from the ACS was used for a second set of simulations. Again, only cases with positive total income were used. As in the SRS simulation, different response rates were simulated by gender, and genders were used as weight adjustment classes. Estimates were obtained for wages and for self-employment income. Sampling probabilities were proportional to total income.

Six samples were drawn comparable to the six drawn from the baseball study. However, unlike the baseball study, no significant differences between the six methods were found for either variable.

Average absolute deviations were under 1% of total income for wages and under 0.6% of total income for self-employment. This suggests that all the estimates were rather accurate, and that for this reason all the methods provided estimates close to the population mean.

3. Conclusions

The results show that order sampling with sequential replacement is a viable method from the perspective of statistical accuracy. It seems particularly useful with record abstractions where there are missing items. The method used and the formula used to order the sample may or may not make a difference.

If one has a good estimate of the response rate, and thus of the total sample expected, including nonrespondents, Pareto may be the method of choice. However, when there is not prior information as to the expected initial sample, but some nonresponse is expected,

Sequential Poisson Sample may be preferable, since it does not require an estimate of the initial size.

References

Aires, N. (1999). Algorithms to find exact inclusion probabilities for Conditional Poisson Sampling and Pareto pps Sampling designs. Methodology and Computing in Applied Statistics, No. 4, pp. 463-475.

Ohlsson E. (1995). "Sequential Poisson Sampling". Institute of Actuarial Mathematics and Mathematical Statistics. Stockholm University. Report No. 182. June 1995.

Ohlsson E. (1996) Personal communication

Rosen, B. (1995) "On Sampling with Probability Proportional to Size", R&D Report 1995:1, Stockholm, Statistics Sweden

Saavedra, P. J. (1995). Fixed-sample-size approximations with a permanent random number. Proceedings of the Section on Survey Research Methods, American Statistical Association, Orlando.