A COMPARISON OF METHODS FOR DRAWING LINKED SCHOOL SAMPLES USING PERMANENT RANDOM NUMBERS

William Robb, Pedro Saavedra, Michael Errecart
William H. Robb, Macro International Inc., 126 College Street, Burlington, Vermont 05401

KEY WORDS: simulation, poisson sampling, YRBS, overlap, Keyfitz

Introduction

In many survey situations it is desirable to coordinate samples across surveys to maximize or minimize overlap of units. For example, in school surveys it is typically desirable to coordinate the samples of students at each grade level to minimize the numbers of schools in the overall . Ideally, sampling would be performed independently at each grade level, however this is not practical in terms of field costs, as large numbers of schools would be drawn. Several methods have been used to draw samples that minimize the number of individual schools drawn while achieving sampling weights as close as possible to those generated under independent grade level sampling. Methods include standardizing school/grade structures prior to sampling or using specialized sampling techniques to coordinating school samples at each grade. In this paper, the recently developed Odds Ratio Sequential Poisson Sampling procedure is applied to this problem and compared to the Keyfitz and Sequential Poisson Sampling in terms of precision and overlap.

Techniques for Controlling Sample Overlap

In the past one of the most important methods for controlling overlap of PPS samples was due to Keyfitz (1951), but more recently methods based on permanent random number techniques have been proposed that better control sample sizes, such as Poisson sampling.

Under Poisson sampling each unit is assigned a desired probability of selection that may be based on size or arbitrary. A random number, uniformly distributed between 0 and 1, is assigned to each unit. The unit is sampled only if the random number is less than the probability of selection.

The major disadvantage of Poisson sampling is that the sample size is not fixed. To remedy this, two variants of Poisson sampling have been proposed that do produce fixed sample sizes:

- Sequential Poisson Sampling (SPS)(Ohlssen, 1995) assigns to each unit the ratio of the random number (r) to the desired probability of selection (p), sorts the list based on this ratio (r/p), and selects the first n units. The actual probabilities of selection are not exactly PPS, and there is no easy way of calculating them, but they are close to the desired probabilities of selection and asymptotically converge to them. Furthermore, Ohlsson showed that by using the desired probabilities rather than the actual ones in developing weight estimates, one could obtain estimates that were just as efficient as those of Poisson sampling.

- Odds Ratio Sequential Poisson Sampling (ORSPS)(Saavedra, 1995) is similar to SPS, except the sort is based on the quantity (r-rp)/(p-rp). ORSPS was found to reproduce desired probabilities better than the Ohlsson method, but this does not seem to translate into more efficient estimates (Saavedra et. al.,1996).

Poisson sampling and the SPS and ORSPS techniques tend to produce samples with high degrees of overlap when the random numbers are retained across sample draws (hence, become permanent random numbers) and if the probabilities of selection do not vary too much.

The Keyfitz procedure takes a different tact towards controlling overlap that does not involve permanent random numbers. Beginning with a given sample and the assumption that the probabilities of selection for the population have been updated, the procedure applies a decision rule to each case in the

sample. If the desired probability of selection has increased the case is retained, but if it has decreased a binomial trial is conducted that might possibly reject the case. For each case rejected another case is selected according to an adjusted probability of selection from among cases not in the sample that have had increases in their probabilities of selection. Like Poisson sampling, the Keyfitz procedure is a PPS procedure that does not produce a fixed sample size.

The next section compares the Keyfitz, SPS, and ORSPS in a simulation study involving four samples of schools for which it is desired to maximize sample overlap. Performance is evaluated based on the accuracy of estimates produced and the extent of sample overlap.

Methodology

Sampling Frame: The sampling frame for these simulations consist of a subset of the sampling frame used in the 1995 Youth Risk Behavior Survey sponsored by the Centers for Disease Control and Prevention. More specifically, the subset consists of 34 PSUs containing 662 schools that had at least one grade in the range 9-12. For each school the total enrollment per grade was taken as the measure of the size. Within each school, a selection probability was computed for each of the eligible by dividing the school's enrollment at a given grade by the total enrollment in the PSU and then multiplying by 3, the number of grades to be selected for that PSU. Selection probabilities were set to 1 for certainty grades, and the probabilities renormalized so that they summed to n over the PSU.

Sampling: In each iteration of the sampling, four samples of schools were selected that had grades 9, 10, 11 or 12 respectively. Each sample consisted of 3 schools per PSU selected within the 35 fixed PSUs. 1,000 iterations of each method were performed. In the implementation of the Keyfitz procedure the initial sample was taken for the 11th grade, the 11th grade sample was then used to produce the 9th, 10th and 12th grade school samples. The number of eligible grades per PSU varied from a maximum of 46 eligible ninth grades to a minimum of 5 9th, 10th, 11th and 12th grades for the smallest PSU. These figures are presented in detail Table 1. Note that the mean number of grades per school was 2.48 for this set of PSUs.

In the case of Keyfitz sampling, the 11th grade sample was taken first and the samples from the other grades were derived relative to the 11th grade sample. An alternative approach would have been to derive the 12th grade sample from the 11th, then the 10th from the 12th, and finally the 9th from the 10th (to use one arbitrary sequence of grades), but that sequential approach tends to result in less overlap across the samples.

Weighting and Estimation: Each grade was assigned a sampling weight that was the inverse of the desired probability of selection used in the sampling. The weighted enrollments for Hispanic students were computed for each PSU and summed over each draw to produce an estimate of total Hispanic enrollments the draw. A weighted estimate of the number of eligible grades containing each grades in the sampling frame, in other words, an estimate of the frame count in terms of schools at each grade level frame was computed as well.

Let E_{IJK} be the enrollment for the I^{th}, PSU, J^{th} school, K^{th} grade and let I_{hijk} be an indicator variable equal to 1 if the K^{th} grade in the J^{th} school in the I^{th} PSU is selected in the H^{th} sample; it is zero otherwise. With the sampling weight is defined as

$$w_{ijk}=\left(\frac{1}{3}\right)\left(\frac{\sum_{j=1}^{m_i} E_{ijk}}{E_{ijk}}\right)$$

The estimate of Hispanic enrollment for the draw for grade K is:

$$\hat{H}_{hk}=\sum_{i=1}^{n}\sum_{j=1}^{m_i} I_{hijk}H_{ijk}w_{ijk}$$

The estimate of the number of 9th, 10th, 11th and 12th grades at each level in the sample is PSU is simply the sum of the weights for that grade for the draw:

$$\hat{C}_{hk}=\sum_{i=1}^{n}\sum_{j=1}^{m_i} I_{hijk}w_{ijk}$$

The population mean and the mean estimate over

draws for each sampling method (with standard deviation) for these quantities are displayed in Table 2.

These measures were compared to the corresponding population values to allow for a comparison of the three methods in terms of estimation. For each sample, the squared difference between estimated and actual total enrollment, Hispanic enrollment and school count for each grade were computed, averaged over samples and compared. These results are presented in table 3, with a minimum significant difference for means computed using a Bonferroni multiple comparison test.

Overlap: Two measures were used to compare the methods in terms of the amount of grade overlap within a school, and therefore in terms of overall sample size and cost. The mean number of unique schools taken in the four grade draws per iteration and the mean number of grades per sampled school are shown, with 95% confidence limits in Table 4.

These comparisons indicate that the Keyfitz procedure shows no advantage over the Sequential Poisson or Odds Ratio Sequential Poisson in this sampling situation, either in the accuracy of the estimates or in terms of the amount of grade overlap at the school level The gross number of schools sampled using the Keyfitz procedure is on the average over 37 percent higher than either of the other methods, and, on the average, less than two grades were sampled per school.

The ORSPS and SPS methods were then tested using paired comparisons. This is feasible as the sampling methods, unlike the Keyfitz procedure, use a permanent random number assigned to each school for each draw. In our study, the same sequence of permanent random numbers was for both methods in the simulation, allowing for a comparisons to be made on a draw by draw basis. By using a paired comparison, any difference in estimates or amount of overlap between methods for a draw is due entirely to the sampling method. For each draw the summary statistics mentioned above were computed and differenced. The differences were then tested for a mean of zero. The results are presented in Table 5.

Conclusions

In terms of estimation, the Keyfitz method produces tighter estimates, than both ORSPS and SPS methods. However, the estimates tend to be more biased, leading to higher mean squared error overall in two of the four grades considered for Hispanic enrollment and three of four grades for estimating frame counts. ORSPS and SPS tend to perform similarly when tested in aggregate. In terms of overlap, the Keyfitz procedure drew samples containing on the average 36 percent more schools than either ORSPS or SPS. In terms of grades per school the Keyfitz method averaged less than two, whereas both ORSPS and SPS averaged over two grades per school.

We therefore conclude that in these sampling situations, the Keyfitz method presents no advantages.

In paired comparisons no substantial differences were found between the ORSPS and SPS methods.

Bibliography

Keyfitz, N. (1951) "Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities" J. Amer. Statist. Assoc. 46, pp.105-109.

Ohlsson, E. (1995a) "Coordination of Samples Using Permanent Random Numbers" in Survey Methods for Business, Farms and Institutions, edited by Brenda Cox, New York: Wiley.

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

Saavedra, P.J. (1988) "Linking multiple stratifications: Two petroleum surveys." Proceedings of the 1988 Joint Statistical Meetings, American Statistical Association Survey Section, 777-781.

Saavedra, P.J., Errecart, M.T and Robb, W. (1996) "Odds Ratio Sequential Poisson Sampling, a fixed Sample Size PPS Approximation." Joint Statistical Meetings, American Statistical Association, Chicago.

Saavedra, P.J. (1995) "Fixed sample size approximations with a Permanent Random Number." Joint Statistical

Meetings, American Statistical Association, Survey Section, Orlando, Florida.

Table 1

	Number of Schools Per PSU in Frame			
	Max	**Min**	**Mean**	**Total**
Ninth Grades	46	5	14.5	493
Tenth Grades	27	5	11.1	377
Eleventh Grades	26	5	11.0	372
Twelfth Grades	26	5	11.0	371
Schools	50	11	19.5	662

Table 2

	Mean Estimated Hispanic Enrollment				Number of Sections (in K)			
Grade	**9**	**10**	**11**	**12**	**9**	**10**	**11**	**12**
SP	7043 (468)	6493 (524)	5771 (477)	5083 (429)	485 (46)	367 (40)	361 (37)	359 (35)
ORSPS	7029 (469)	6504 (530)	5791 (475)	5107 (432)	493 (47)	374 (44)	369 (39)	368 (36)
KEF	6960 (410)	7007 (437)	5772 (409)	5135 (377)	444 (38)	351 (27)	370 (38)	425 (35)
Population	7019	6482	5771	5091	493	377	372	371

Table 3

	Hispanic Enrollment (in K)				Number of Sections (in K)			
Grade	**9**	**10**	**11**	**12**	**9**	**10**	**11**	**12**
SP	219	275	228	184	2.15	1.72	1.59	1.34
ORSPS	220	281	226	186	2.21	1.72	1.53	1.31
KEF	164	476	167	144	2.03	3.99	2.58	6.98
Min. Sig. Difference	30	46	43	28	0.42	0.30	0.29	0.39

Table 4

	Number of Schools				Grades Per School			
	Mean	**Std.**	**Lower CLM**	**Upper CLM**	**Mean**	**Std.**	**Upper CLM**	**Lower CLM**
SP	187.0	4.25	186.8	187.3	2.18	0.049	2.17	2.18
ORSPS	186.9	4.08	186.7	178.2	2.18	0.047	2.18	2.18
Keyfitz	255.1	5.65	254.8	255.5	1.58	0.350	1.57	1.58

Table 5

	Section Count				Unique Schools	Mean Grades Per School
	9	**10**	**11**	**12**		
Mean	-60	-0.41	-1.88	27	0.06	-0.0006
Std. (in K)	1.836	1.540	1.527	1.296	2.40	0.028
P Value	0.299	0.994	0.969	0.492	0.4305	0.4919

	Hispanic Enrollment			
	9	**10**	**11**	**12**
Mean	-915	-6146	1757	-2608
Std. (in K)	125	157	143	117
P Value	0.818	0.217	0.689	0.4816