ISSUES WITH PROSPECTIVE SAMPLING FROM A SAMPLING FRAME WITH FLUCTUATIONS ACROSS TIME

Margaret Byron and Jun Liu, Research Triangle Institute
Margaret Byron, RTI, P.O. Box 12194, Research Triangle Park, NC 27709

Key Words: Prospective Sampling, Unequal Weighting Effects, Sampling Rate Adjustment

Introduction

Prospective sampling is a sampling method that selects samples as the eligible subjects appear over time. The sampling frame for the whole target population is typically not constructed prior to the start of the sampling and data collection. One very important consideration in designing a prospective sampling method is how to minimize the variation in the sampling weights, and still achieve the prescribed total sample yield at the end of the data collection. On one hand, the study design typically has a predetermined sample size requirement for achieving certain analytical goals. Field operation requirements also dictate that the number of interviews to field within each regular time period does not fluctuate very much. On the other hand, the frame count changes across time and is most likely beyond the sampler's control. Depending on the design, such fluctuation brings many challenges: an unpredictable number of selections for field interviewers, difficulties in achieving the predetermined sample sizes, and a wide variation in sampling weights.

Very often, when information on frame count fluctuations over a period of time is available or if the fluctuations can be predicted, one can design the sample selection algorithm to take the temporal variations in frame counts into consideration. When the sampling frames experience unpredictable fluctuations during the sample selection period, maintaining a minimum sampling weight variation is an especially challenging problem. This is the dilemma that we faced in the HIV Risk and Psychiatric Disorder of Women Probationers study. In this paper, we will discuss the issues and options relating to sampling rate adjustments and sample allocation for this type of sample design in the context of this study.

The Women Probationer Study

The HIV Risk and Psychiatric Disorder of Women study was conducted on the target population of women aged 18 and over who were sentenced to probation in North Carolina from June through December, 1998. Women who were put on probation after serving prison terms, or split sentences, were not eligible for the study. The goal of the study was to produce 1,000 completed interviews. The first stage of the sampling design involved the selection of PSUs that were formed by clusters of no more than two adjacent counties in North Carolina. There were a number of counties which could not be combined with an adjacent county to form a PSU with at least thirty probationers, so these very small counties were dropped from the study. Each PSU was assigned a size measure equal to the total number of women who were sentenced to probation in the county or counties contained in the PSU during the 12-month period from July, 1996 to June, 1997. The PSUs were allocated to one of the three selection strata: certainty PSUs, non-certainty large PSUs and non-certainty small PSUs. There were eight PSUs that had a probability of selection, based on the size measure, greater than one and were assigned to the certainty PSU stratum. The thirty-nine non-certainty PSUs which had at least the minimum size requirement of eighty-nine probationers were placed in the non-certainty large PSU stratum. The remaining fifteen PSUs composed the non-certainty small PSU stratum. The non-certainty PSUs were selected from each of the two non-certainty strata using PPS from a sampling frame sorted by region and urbanicity. This sampling frame order ensured a sample of PSUs from each region that included both urban and rural counties. Twenty-one PSUs were selected from the non-certainty large PSUs and two were selected from the non-certainty small PSUs, which combined with the certainty PSUs resulted in thirty-one PSUs selected in the first stage of the sample design.

With assumed response and eligibility rates, a sample yield of 1,588 women probationers was required in order to obtain the desired 1,000 completed interviews. The required sample yield was allocated proportionately to the certainty and non-certainty strata and the initial sampling rates were calculated using proportional-to-size for the certainty PSUs and fixed equal sample selections within each stratum for the non-certainty PSUs.

Because of the time sensitive restrictions of the instrument and the transient nature of the population, a prospective sampling method was used for the second stage unit selection. Eligible women probationers were selected each week for approximately six months. Every week, we received a list of women from the North Carolina Department of Corrections (NCDOC) containing all of the women who were placed on probation during a one week period in the counties that were in our thirty-one selected PSUs. From the sampling frame, women were selected for the study from each PSU according to the following systematic sampling algorithm:

1. Sort the list in random order

2. Select the first k_i names on the list, where k is calculated from the following formula:

$$k_i = [f_i * N_i] + b_i$$

where f_i is the sampling rate for the i^{th} PSU, N_i is the number of women on the sampling frame for the i^{th} PSU, $[f_i * N_i]$ denotes the integer part of the product of the sampling rate and the frame count, and b_i is determined by a uniform random variate u with the following rules:

$$\begin{cases} b_i = 0 \;\; if\ u \geq d_i \\ b_i = 1 \;\; if\ u < d_i \end{cases}$$

where d_i is the fractional part of the sampling rate, or

$$d_i = f_i * N_i - [f_i * N_i]$$

Over repeated samples, this sampling algorithm will realize the desired sampling rate f_i for the i^{th} PSU.

Due to unforseen fluctuations in the weekly sampling frames that we received, we found it necessary to invoke a sampling rate adjustment scheme, but we were unsure of when to make the adjustments in order to keep the unequal weighting effects at a minimum. Should we make one or two adjustments over the six months of data collection or should we make more adjustments at shorter time intervals?

Simulation Results

To examine the effects of various plans for weight adjustments, we looked at four different possible weight adjustment schemes through a simulation:

A. Two weight adjustments in the second half of the sampling period

B. Four weight adjustments in the second half of the sampling period

C. Four weight adjustments evenly spaced throughout the sampling period

D. Weight adjustments every two weeks throughout the sampling period

Since we are interested in how the weight adjustment schemes affect the unequal weighting effects of the second stage sampling, we kept the PSUs and weekly sampling frames fixed. Starting with the initial sampling rates, we selected samples from each weekly frame using the same sampling algorithm given in the previous section. When it came time to adjust the sampling rates, we used the following formula for the proposed adjusted sampling rate:

$$adj\ rate_i = \frac{\left(s_i - \sum_{j=1}^{w} y_{ij}\right)\left(\sum_{j=1}^{w} x_{ij}\ /w\right)}{n - w}$$

where s_i is the total number probationers to be selected in the i^{th} PSU, x_{ij} is the number of women on the sampling frame for PSU i and week j, y_{ij} is the number of women selected in PSU i and week j, n is the number of weeks in the sample selection period and w is the number of weeks for which samples have already been selected. If the proposed sampling rate was greater or less than the current sampling rate by at least 0.05, then the current sampling rate was updated to the proposed sampling rate. Otherwise, the current sampling rate remained the same until the next possible rate adjustment period. We performed the sampling rate adjustments for each PSU according to one of the above adjustment schemes and repeated the sampling process for 150 repetitions using the same weight adjustment scheme. Since the number of women selected from the weekly frame in each PSU is determined by the sampling rate and the random variate u, we used the same seed at the start of the first repetition so that each scheme used the same sequence of pseuo-random numbers. This seed selection allowed the sample sizes for the PSUs to remain the same across the schemes until the first weight adjustment is made. For weight adjustment schemes A and B, the simulation selects the same sample sizes for each PSU from each weekly sample frame until the sampling rates are changed after 10 weeks of samples have been selected. All of the weight adjustment schemes had the last weight adjustment performed after 18 weeks of sample selections. The results of the simulation are shown in **Table 1**.

All of the weight adjustment schemes produced very similar mean sample sizes, with only scheme A having a significantly smaller mean sample size than the other three schemes. For the design effects, schemes A, B and C produced similar mean design effects and only scheme D was significantly smaller than any of the other schemes. This shows that scheme D, performing weight adjustments every two weeks throughout the sample selection period, produces significantly smaller design effects, while performing equally as well as the others in producing the desired sample size at the conclusion of the sample selection.

Discussions on Implementation

During the sample selection period, the selected samples were monitored each week. To check the progress of the sample selections, each week we looked at the sampling frame that had arrived, the number of women that should be selected from the new frame, the cumulative sampling frame totals, and the cumulative sample totals in comparison to the corresponding values that we had expected for that particular week.. We also looked at the difference between the current sampling rate and the proposed adjusted sampling rate should the sampling rates be changed. The proposed adjusted rate for PSU i was calculated by the formula given in the previous section. As we started seeing larger differences between the current and the proposed adjusted rates in more and more PSUs, we discovered that we needed to consider a weight adjustment plan in order to achieve our required sample yield.

Another possible adjustment to the sample design that could have been considered was sample re-allocation. If the number of women in the sampling frames had become disproportionate across the certainty and non-certainty strata, we could have re-allocated the samples to account for the change in the population distribution. However, since we were not selecting the second stage sampling units according to other sampling strata, such as race categories, there was no need for us to make that type of adjustment to our sampling rates.

Summary and Conclusions

The weight adjustment scheme that resulted in a significantly lower unequal weighting effect was the one in which the sampling rates were evaluated for adjustments every two weeks during the sample selection period. A plot of the weekly population counts for the selected PSUs is shown in **Graph 1** and a smoothed plot of the weekly population counts is shown in **Graph 2**. There appears to be a trend in the data according to the smoothed line, even though the population counts for these PSUs fluctuate quite a bit from week to week. However, this trend does carry through to the population fluctuations in each PSU. Time series plots for four PSUs are shown in **Graph 3**.

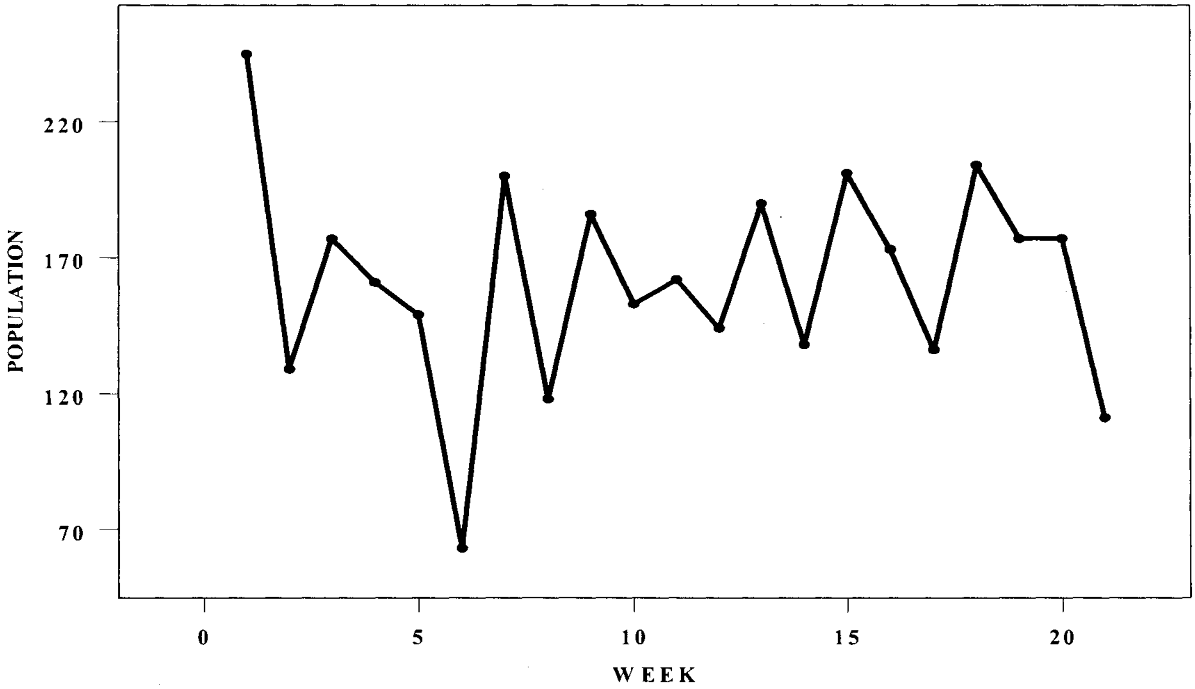
The sampling rate adjustments made frequently throughout the sample selection period performed better than the rate adjustments made less frequently. This rate adjustment scheme also performed better than those in which the adjustments were made in the second half of the sample selection period. More research needs to be done to examine why the bi-weekly rate adjustment scheme performed better for the selected PSUs in this particular population and whether the same results will occur in other prospective sampling situations.

Table 1. Simulation Results

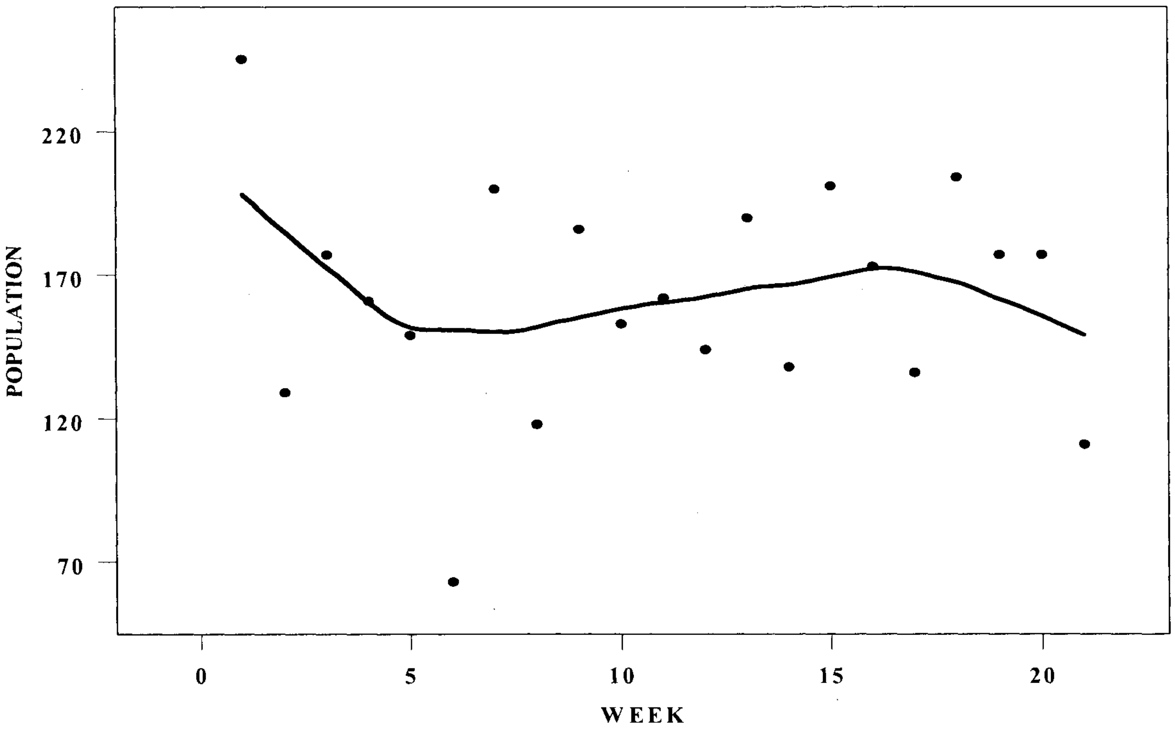
Weight Adjustment Scheme	Achieved Sample Size[1]					Achieved Design Effects				
	Mean	SE	Median	Min	Max	Mean	SE	Median	Min	Max
A. 10th and 18th wk	1527	0.535	1527	1512	1551	1.327	0.0075	1.302	1.203	1.695
B. 10th, 13th, 15th, 18th wk	1531	0.626	1531	1514	1560	1.302	0.0069	1.276	1.188	1.599
C. 4th, 8th, 12th, 18th wk	1532	0.477	1532	1520	1545	1.297	0.0099	1.254	1.168	1.912
D. Every two weeks up to 18th wk	1532	0.505	1532	1520	1548	1.233	0.0063	1.206	1.164	1.502

[1] This simulation used only 21 weeks of data, as opposed to the 26 weeks that were originally planned for sample selection at the beginning of the study. The sample selection period was cut short due to time constraints.

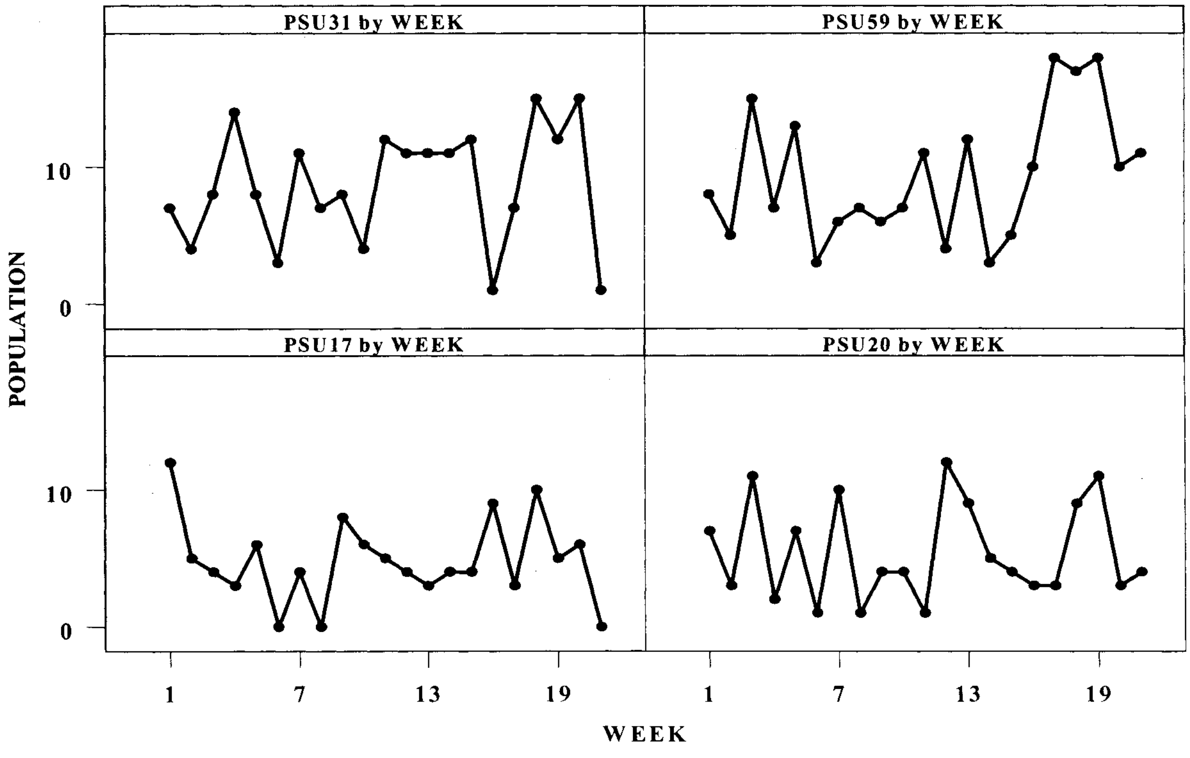
Graph 1. Population of Women Probationers in Sample Selection Period



Graph 2. Smoothed Population of Women Probationers in Sample Selection Period



Graph 3. Plots of Select PSUs


PSU31 by WEEK
PSU59 by WEEK
PSU17 by WEEK
PSU20 by WEEK
POPULATION
WEEK