PSU PROBABILITIES GIVEN DIFFERENTIAL SAMPLING AT SECOND STAGE

Mansour Fahimi and David Judkins, Westat, Inc.
David Judkins, 1650 Research Blvd., Rockville, MD 20850

Key Words: Admissibility, Between/Within-PSU Variances, Measure of Size, Simulation, Stratification

1.0 Introduction

Since the start of NHIS about 30 years ago, PSUs have been selected with probability proportional to total population. However, there are compelling arguments for considering measures of size for PSUs that reflect the interaction of demographic composition with the type of oversampling that is desired by demographic domain. The various schemes proposed along these lines can reduce both between-PSU variance for targeted demographic domains and variation in interviewer workload from PSU to PSU. Moreover, these positive effects can be realized with minimal impact on within-PSU variance. The reduction in between-PSU variance is caused by: greater number of sample PSUs with nontrivial minority concentrations; less variability in the number of minority sample persons per sample PSU; and higher correlation between the PSU probability of selection and minority statistics.

On the negative side, the proposed alternate measures of size can increase the between-PSU variance for nontargeted domains and for totals, and they can lead to decreased efficiency for surveys who share the same set of PSUs but do not share the same oversampling goals. In particular, utilization of these measures might cause the 1995 NHIS to have fewer PSUs in common with the 1985 redesign which could lead to increased recruiting and training cost for the phase-in of the new design and slightly greater turbidity in time-series. Consequently, the most important trade-off in considering these alternate measures of size appears to be in reduction of between-PSU variance and workload variation on the one hand, and possibility of degrading the utility of the design (NHIS) for other surveys on the other hand.

In this paper we consider the traditional measure of size along with three proposed alternatives. We compare these measures on theoretical grounds, and present the results of a simulation study. Although the focus is on a general demographic survey (NHIS), the results are applicable to other multistage samples where the overall sampling rates vary among ultimate sample units, e.g., sample of students clustered by school where students in different fields are to be sampled at different rates.

For a two-stage sampling scheme, Waksberg (1975) proposed two alternative measures of size, each emphasizing the importance of one of the stages. The first alternative places priority on equalizing the cluster size of the total sample in each PSU, and requires measures of size proportional to:

$$\sum_{k=1}^{K} \frac{N_{jk}}{N_{.k}} = \frac{1}{n} \sum_{k=1}^{K} \frac{n}{N_{.k}} N_{jk}$$

$$= \frac{1}{n} \sum_{k=1}^{K} f_k N_{jk}$$

where N_{jk} is the number of ultimate units in the k-th domain of the j-th PSU, and n represents a constant sample size per each domain.

It should be noted that Hendricks has used a similar measure of size for the National Assessment of Educational Progress (NAEP) in the early 1970's. Folsom (1980) has developed a general multiple domain version of this measure which maintains a fixed workload in each PSU and achieves a self-weighting sample in each domain. In order to allow changes in the sampling rate and domain membership definition, Folsom et al. (1987) have introduced modifications of the above composite measure. While preserving the self-weighting feature of the sample, these modifications result in a varying number of elementary units selected in each PSU.

Waksberg's second alternative measure of size is designed to ensure that first-stage units with strong concentration of any rare domain are selected with higher probabilities. It is of the form:

$$\underset{k}{\text{Max}} \left\{ \frac{N_{jk}}{N_{.k}} \right\}.$$

1.1 Definitions

For illustration purposes we define three demographic domains: households with a Black nonHispanic householder; households with an Hispanic householder; and all other households. Let f_1, f_2 and f_3 be the resulting overall sampling fractions for these three domains of households, and let Q_{ij1}, Q_{ij2} and Q_{ij3} be corresponding estimates of households in the i-th first-stage stratum for the j-th PSU. The traditional or generic, the first alternative (Hendricks/Waksberg/Folsom), and the second alternative measures of size (Waksberg), respectively, are then given by

$$MOS_{ij}^{T} = \sum_{k=1}^{3} Q_{ijk}$$

$$MOS_{ij}^{1} = \sum_{k=1}^{3} (f_k \times Q_{ijk})$$

$$MOS_{ij}^{2} = \max_{k} \left\{ f_k \times \frac{Q_{ijk}}{Q_{.k}} \right\}$$

Furthermore, we wish to consider a third alternative. This alternative comes from a class of size measures due to Don Malec (1990), derived by considering an admissibility concept. M is inadmissible if there is some other measure of size, M*, such that the between-PSU variance for every statistic in a predefined set is at least as good with M* as with M and the between-PSU variance of one statistic is actually better with M* than with M. Given a fixed stratification, Malec states that any measure of size with the following form is admissible:

$$M = \sqrt{\sum_{t} \lambda_t \left(\frac{Q_{ijt}}{Q_{i.t}} \right)^2}$$

where t ranges across the set of statistics and $\lambda_t > 0$. Specifically, the third alternative measure of size considered in this paper is given by:

$$MOS_{ij}^{3} = \sqrt{\sum_{k=1}^{3} f_k \left(\frac{Q_{ijk}}{Q_{.k}} \right)^2}$$

1.2 Operational Effects

1.2.1 Workload

We define the workload for a PSU as the sum of sample sizes from each of the 3 domains of ultimate units from the PSU. With the generic MOS when the overall sampling fractions are not constant, the only way to get constant workloads is to limit the extent of oversampling in some PSUs and accept wider variations in overall probabilities. But this has adverse effects on within-PSU variances; hence this sort of procedure is rarely worth considering.

The first alternate MOS, while self-weighting, has the benefit of inducing constant expected workloads. The second alternative measure of size will also have effects on workloads. By increasing the measure of size for PSUs with unusual concentrations of any of the ultimate domains, it decreases all the within-PSU sampling fractions for such PSUs and results in fewer extremely large workloads than what would result with the traditional MOS. The third alternative appears to fall between the first and second.

1.2.2 Variance Effects

Within-PSU Variance: Let S_{wijk}^2 be the population variance for some characteristic within the k-th ultimate domain of the j-th PSU in the i-th stratum. Then when the sample is self-weighting, the within-PSU variance of an estimate of the total prevalence of that characteristic is

$$\begin{aligned}
\sigma_w^2 &= \sum_i \sum_j \frac{1}{P_{ij}} \sum_k \frac{N_{ijk}^2}{n_{ijk}} \left[1 - \left(f_{ijk} \approx \frac{n_{ijk}}{N_{ijk}} \right) \right] S_{wijk}^2 \\
&= \sum_i \sum_j \sum_k \frac{1}{P_{ij}} \times \frac{N_{ijk}^2}{n_{ijk}} (1 - f_{ijk}) S_{wijk}^2 \\
&= \sum_i \sum_j \sum_k \frac{1}{P_{ij}} \times \frac{N_{ijk}}{f_{ijk}} (1 - f_{ijk}) S_{wijk}^2 \\
&= \sum_i \sum_j \sum_k \frac{f_{ijk}}{f_k} \times \frac{N_{ijk}}{f_{ijk}} (1 - f_{ijk}) S_{wijk}^2 \\
&= \sum_i \sum_j \sum_k \frac{1}{f_k} \times N_{ijk} (1 - f_{ijk}) S_{wijk}^2
\end{aligned}$$

where next to the last statement follows due to self-weighting scheme within domains.

Typically, within-PSU sampling fractions are tiny. Thus, under all but exceptional circumstances,

$$\sigma_w^2 = \sum_i \sum_j \sum_k \frac{N_{ijk} S_{wijk}^2}{f_k}$$

where it is noted that the first-stage probabilities play no role.

Between-PSU Variance: Once the strata are fixed, selection of PSUs with probability proportionate to a measure of size sensitive to targeted populations will result in a further decrease in between-PSU variance for those targeted populations (and a further increase in between-PSU variance for the total population and for nontargeted populations). This is fairly obvious from inspection of the formula for between-PSU variance.

$$\sigma_B^2 = \sum_i \sum_j P_{ij} \left(\frac{X_{ij..}}{P_{ij}} - X_{i...} \right)^2$$

where X_{ijkl} is the value of some statistic on the l-th ultimate unit within the k-th domain within the j-th PSU of the i-th first-stage stratum and a dot signifies summation on a subscript. A strong correlation between $X_{ij..}$ and P_{ij} makes for more accurate projection of PSU totals to stratum totals and thus smaller between-PSU variance. It is easy to see that the correlation between $X_{ij..}$ for targeted domains and P_{ij} will be increased through use of the first alternate measure of size to determine P_{ij}. Although less obvious for the second and third measures of size, the same is thought to hold true.

The second alternate measure of size affects between-PSU variances a little differently. The effect of this measure of size is to increase the probability of selection for PSUs that have unusually high concentrations of any of the ultimate classes. This

provides a guard against missing any PSU that is extremely important for any class. It thus guards against having a large between-PSU variance for any class.

2. Simulation

Given the obduracy of several of the alternative measures of size to intuitive comprehension, it seemed wise to conduct a small simulation study prior to making a final decision on which alternate (if any) to use. There were several steps in the simulation study. The first was to create a set of pseudo PSUs. For each PSU, it was necessary to develop models for population sizes for each of the three domains (Blacks, Hispanics, and others) and models for underlying and realized disease prevalence rates. These were then used to calculate the measure of size for each PSU under each scheme. The second step was to stratify the PSUs. The final step was to calculate between-PSU variances on disease incidence by domain given the probabilities induced by the stratification and the various measures of size.

2.1 Creation of Pseudo PSUs

First, the population specific to each domain was assumed to follow a Gamma distribution across PSUs. The Gamma distribution was selected because of its non-negative and right-skewed properties which are fairly realistic for PSU population totals. A separate gamma distribution was used for each of the three domains. The distribution parameters were chosen such that the nonBlack/nonHispanic population would have the largest mean and the smallest coefficient of variation (cv), the Black population would have a much smaller mean and a somewhat larger cv, and the Hispanic population would have a mean smaller yet with a very large cv. The three gamma variables were generated for each of 100 pseudo PSUs using the standard SAS pseudo random number generator. (The three variables should be nearly independent across pseudo PSUs.) The specific parameters and resulting population means and cv's follow:

Black Population

$Q_{1i} \sim \Gamma(\alpha=1.0, \beta=10{,}000)$, with
$E[Q_{1i}] \equiv 10{,}000$, $cv(Q_{1i}) \equiv 1.00$

Hispanic Population

$Q_{2i} \sim \Gamma(\alpha=0.4, \beta=18{,}000)$, with
$E[Q_{2i}] \equiv 7{,}200$, $cv(Q_{2i}) \equiv 1.58$

Other Population

$Q_{3i} \sim \Gamma(\alpha=4.0, \beta=15{,}000)$, with
$E[Q_{3i}] \equiv 60{,}000$, $cv(Q_{3i}) \equiv 0.50$

Using the resulting set of domain by pseudo-PSU populations and reasonable sampling fractions ($f_1=0.000554$, $f_2=0.000596$ and $f_3=0.000380$ for Blacks, Hispanics and other, respectively), the traditional measure of size along with the three proposed alternative measures of size were computed for each pseudo PSU.

Three different models were used to simulate disease prevalence rates for the three domains across PSUs. (The disease simulated is a generic disease.) The first model assumed that the prevalence of the disease follows a binomial distribution across PSUs with a different mean for each domain. Let X_{1i}, X_{2i} and X_{3i} represent counts of cases for some rare disease in PSU i among Blacks, Hispanics and others, respectively.

Model 1

$X_{1i} \sim B(Q_{1i}, p_1=0.05)$
$X_{2i} \sim B(Q_{2i}, p_2=0.03)$
$X_{3i} \sim B(Q_{3i}, p_3=0.02)$

Furthermore, let $X_{4i} = (X_{1i} + X_{2i} + X_{3i})$ represent the overall count of cases of the disease in the i-th PSU. We generated two replications of Model 1 using different seeds for the pseudo-random number generator. Both PSU populations and disease counts were replicated.

In the second model, we assumed that the prevalence of the disease follows a beta-binomial distribution across PSUs with a different mean for each domain. This distribution was selected because of the larger variance it creates. The resulting "hot spots" in the disease correspond more closely to reality for many diseases.

Model 2

$X_{1i} \sim B(Q_{1i}, \Theta_{1i})$, with
$\Theta_{1i} \sim \text{Beta}(\alpha_1=1, \beta_1=19)$
$X_{2i} \sim B(Q_{2i}, \Theta_{2i})$, with
$\Theta_{2i} \sim \text{Beta}(\alpha_2=1, \beta_2=32.33)$
$X_{3i} \sim B(Q_{3i}, \Theta_{3i})$, with
$\Theta_{3i} \sim \text{Beta}(\alpha_3=1, \beta_3=49)$

Note that the beta parameter for each distribution was chosen such that the mean for the beta-binomial variable for a domain was equal to the corresponding mean in the first model; i.e., $E[\Theta_{ki}]=\alpha_k/(\alpha_k+\beta_k)=p_k$ for every k. As for Model 1, we generated two replications of Model 2 using different seeds for the pseudo-random number generator. Both PSU populations and disease counts were replicated.

The third model was very similar to the second model but assumed that all three domains had equal susceptibility to the disease. This model was examined not because of plausibility but to test out

some theories on the behavior of the alternate measures of size.

Model 3

$X_{1i} \sim B(Q_{1i}, \Theta_{1i})$, with
$\Theta_{1i} \sim \text{Beta}(\alpha_1=1,\beta_1=32.33)$
$X_{2i} \sim B(Q_{2i}, \Theta_{2i})$, with
$\Theta_{2i} \sim \text{Beta}(\alpha_2=1,\beta_2=32.33)$
$X_{3i} \sim B(Q_{3i}, \Theta_{3i})$, with
$\Theta_{3i} \sim \text{Beta}(\alpha_3=1,\beta_3=32.33)$

As for Models 1 and 2, we generated two replications of Model 3 using different seeds for the pseudo-random number generator.

2.2 Stratification

After generation of random variables, PSUs were divided into 10 groups of 10 PSUs each according to proportion Black. These groups were ordered by ascending Black density. Within each group, the PSUs were sorted by proportion Hispanics. However, this sort did not have a constant direction. In the first Black density group, the PSUs were sorted on ascending Hispanic density. In the second Black density group, the PSUs were sorted on descending Hispanic density. This procedure reduces discontinuity for the second sort key at the boundaries of the first sort key.

Each of the four measures of size under consideration was calculated for each PSU for each replicate and model. To make the four measures of size for a PSU in a given replicate more comparable, we standardized the measures of size. After standardization, the sum of each MOS across the 100 PSUs in a replicate was equal to 100,000. This makes it easier to get an intuitive feeling for the impact of the different rules.

For each measure of size and replicate, a stratification was created with 5 component strata. These strata were formed in such a manner as to minimize the differences between stratum measures of size (come as close as possible to 20,000 for each stratum) subject to the constraint that PSUs in the same stratum should never be separated by PSUs from a different stratum with respect to the Black/alternating Hispanic sort. The first stratum was formed for each measure of size by starting at the top of the file with respect to the shown sort and cumulating downwards until the measure of size for the first stratum was close to 20,000. This process was repeated to form the remaining four strata for each measure of size.

2.3 Between-PSU Variances

The variances were calculated using the standard formula given in Section 1.2. To ensure stability of the results, the preceding process was repeated once for each model using a different seed for the creation of all random variates (including PSU populations) before variances were computed. The resulting sets of between-PSU variances are summarized in Table 2.3.1. Corresponding relative between-PSU variances are shown in Table 2.3.2.

Examining the columns for Model 1 in Table 2.3.1, we see that the first and third alternative measures of size (MOS^1 and MOS^3) are better than the generic measure of size for Blacks on both replications. The second alternate, MOS^2, can be better or worse for Blacks than the generic. For Hispanics, any of the three alternates are better than the generic measure of size. Furthermore, the reduction in between-PSU variance appears more substantial than for Blacks. Among the alternates, the second and third appear to be better for Hispanics than the first alternate, with a slight preference for MOS^2. For the other population, the traditional MOS consistently produces the smallest between-PSU variance. The penalty for using the second or third alternate can be extremely severe. For the entire population, the first alternative measure of size, MOS^1 outperforms the others.

The superiority of MOS^2 and MOS^3 for Hispanics is easy to understand since the distribution of the Hispanic population across PSUs is the lumpiest of the three domains and since MOS^2 and MOS^3 are very sensitive to lumps. (This is especially true of MOS^2.) Similarly MOS^1 or MOS^3 is best for Blacks because the Black population distribution is slightly less lumpy. The traditional MOS^T is best for the other population since the correlation is high between Q_3 (the count of other population) and MOS. The superiority of MOS^1 to MOS^T for the entire population is somewhat surprising. We suspect that it is mainly due to the conjunction of the higher disease prevalence assumed for minorities with the oversampling of minorities. Note that improvements achieved for minorities by adopting the second or third alternates over the traditional MOS or first alternate are trivial compared to the degradation for the other and total populations.

Examining the columns for model 2 in Table 2.3.2, we see similar patterns emerge: the MOS^3 is best for Blacks (with MOS^2 a close second), the traditional MOS is best for others, and MOS^1 is best for the entire population. For Hispanics, the second and third alternative measures of size, MOS^2 and MOS^3 seem to be about equally satisfactory. For this model, the advantages of the second and third alternates for the minority populations are closer in magnitude to the disadvantages for the other and total populations.

Finally, examining the columns for model 3 in Table 2.3.2, we see that the findings for Model 1

were replicated for Hispanics, others and overall, while for Blacks the second and third alternatives appear to be equally satisfactory with the first alternate clearly in third-place. Here we see that the traditional MOS and the first alternate are essentially tied for the total population, thereby partially confirming our explanation of the puzzling superiority of MOS^1 to MOS^T for the total population under model 1. Even the fact that they are tied is at first surprising. We suspect that it is due to the greater variance in the distribution of PSU-level populations for minorities. The disease count, X_{ji}, for a PSU and domain is binomial given $n=Q_{ji}$ and $p=\Theta_{ji}$. Θ_{ji} has the same distribution for each domain in model 3, but Q_{ji} does not. The greater relative variance in Q_{2i} and Q_{3i} means that X_{2i} and X_{3i} have greater relative variance than X_{1i}. Since MOS^1 leads to smaller variances on X_2 and X_3 than does MOS^T, perhaps this explains why MOS^1 appears to be as effective in reducing the variance on X_4 as MOS^T even though the correlation is greater between X_4 and MOS^T than between X_4 and MOS^1.

2.4 Summary

i. MOS^1 is consistently better for both minorities than the traditional MOS^T.

ii. MOS^2 and MOS^3 are the best for the Hispanic population (because of the lumpy distribution). In no case does either method produce a worse result than MOS^T or MOS^1. There is little reason to prefer one to the other for the Hispanic population.

iii. MOS^2 and MOS^3 perform well for the Black population, but MOS^2 does occasionally produce a worse result than MOS^T and MOS^1. MOS^3 thus seems preferable for the Black population.

iv. MOS^2 and MOS^3 can produce extremely unfavorable results for the other and total populations with MOS^2 the more dangerous of the two. The degradation is less severe for the beta-binomial models than for the straight binomial model, but it is still troubling.

3. Recommendations

At this point, it is clear that Blacks and Hispanics are to be intensively oversampled for NHIS. Assuming that the use of NHIS sample PSUs for other surveys is a secondary concern, then one of the alternate size measures should be used. Given our past experience with MOS^1 and the results of the simulation study, we recommend it. MOS^2 and MOS^3 are so extremely oriented to optimization for minority statistics, that their use might make NHIS PSUs unsuitable to serve as the basis for an integrated design for several surveys.

This tentative recommendation becomes more emphatic if there is a continued belief at the Census Bureau that PSU-level workloads need to be tightly controlled for efficient administration of the program. In that case, use of MOS^1 at the first stage would allow oversampling at the second stage while simultaneously maintaining workload control and self-weighting samples within the intersections of second-stage strata and ultimate domains.

On the other hand, thinking about total variances, as within-PSU sampling procedures are skewed more and more to reduction of within-PSU variance for minority statistics at the expense of precision on white and other statistics, there is some point at which it would be better to adopt a more radical measure of size, such as MOS^3, for PSUs than to further skew the within-PSU sample. We are not sure where that point is but think that the current NHIS within-PSU sampling plans have not yet reached it. Even if NHIS plans become skewed more to minority statistics than they are currently, it appears that MOS^3 would be a better choice than MOS^2 since the difference between the two are slight for minority statistics and large for other and total statistics under model 1. However, this would probably need further research.

4. References

Folsom, R.E. and Iannacchione,V.G. (1980). "NMCUES State Medicaid Household Survey Sample Design Statement," an RTI final report for Health Care Financing Administration and the National Center for Health Statistics.

Folsom, R.E., Potter, F.J., and Williams, S.K. (1987). "Notes on a Composite Measure for Self-Weighting Samples in Multiple Domains." *Proceedings of the Section on Survey Research Methods, American Statistical Association, 792-796.*

Malec, D. (1991). "Optimal Multiple Objective Survey Design: Admissible Principle Component Designs." *Submitted for Publication.*

Waksberg, J. (1975). "Sample Selection When Equal Size Samples in a Number of Fields are Required." *An Internal Memo to Sampling Staff in Westat.*

Table 2.3.1. Between-PSU variances (in millions)

MOS	Domain	Model 1. Binomial X, Varying means across domians		Model 2. Beta-Binomial X, Varying means across domians		Model 3. Beta-Binomial X, Constant means across domians	
		Seed 1	Seed 2	Seed 1	Seed 2	Seed 1	Seed 2
Traditional	Black	97	93	708	1,256	417	721
Alter. 1	Black	61	68	680	1,132	403	668
Alter. 2	Black	58	101	575	721	221	375
Alter. 3	Black	42	72	544	697	240	383
Traditional	Hispanic	154	141	333	466	507	449
Alter. 1	Hispanic	119	106	286	408	442	394
Alter. 2	Hispanic	51	35	192	239	240	230
Alter. 3	Hispanic	61	40	190	253	255	243
Traditional	Other	100	87	1,812	2,087	3,976	4,362
Alter. 1	Other	159	145	1,936	2,134	4,150	4,484
Alter. 2	Other	1,304	1,189	3,681	3,980	9,669	8,824
Alter. 3	Other	924	778	3,061	3,390	7,065	7,465
Traditional	Total	68	56	3,387	4,675	4,929	5,811
Alter. 1	Total	15	13	3,368	4,376	4,950	5,680
Alter. 2	Total	1,314	1,389	5,207	5,527	10,133	9,649
Alter. 3	Total	735	762	4,293	4,710	7,405	8,101

Table 2.3.2. Relative between-PSU variances

MOS	Domain	Model 1. Binomial X, Varying means across domians		Model 2. Beta-Binomial X, Varying means across domians		Model 3. Beta-Binomial X, Constant means across domians	
		Seed 1	Seed 2	Seed 1	Seed 2	Seed 1	Seed 2
Traditional	Black	0.0344	0.0366	0.3720	0.4935	0.5380	0.7209
Alter. 1	Black	0.0217	0.0267	0.3573	0.4448	0.5200	0.6679
Alter. 2	Black	0.0206	0.0397	0.3021	0.2833	0.2851	0.3749
Alter. 3	Black	0.0149	0.0283	0.2859	0.2739	0.3097	0.3829
Traditional	Hispanic	0.1977	0.2628	0.4319	0.7097	0.5771	0.6804
Alter. 1	Hispanic	0.1527	0.1975	0.3710	0.6214	0.5031	0.5970
Alter. 2	Hispanic	0.0655	0.0652	0.2490	0.3640	0.2732	0.3485
Alter. 3	Hispanic	0.0783	0.0745	0.2464	0.3853	0.2903	0.3682
Traditional	Other	0.0073	0.0059	0.1581	0.1610	0.1482	0.1575
Alter. 1	Other	0.0117	0.0098	0.1690	0.1646	0.1547	0.1619
Alter. 2	Other	0.0956	0.0803	0.3213	0.3070	0.3605	0.3186
Alter. 3	Other	0.0677	0.0525	0.2672	0.2615	0.2634	0.2695
Traditional	Total	0.0017	0.0015	0.1064	0.1296	0.1007	0.1161
Alter. 1	Total	0.0004	0.0003	0.1058	0.1213	0.1011	0.1135
Alter. 2	Total	0.0336	0.0364	0.1635	0.1532	0.2070	0.1928
Alter. 3	Total	0.0188	0.0200	0.1348	0.1306	0.1513	0.1618