Two weighting schemes for combining sample in the Health Behaviors in School-age Children Survey.

Ronaldo Iachan, Pedro Saavedra, William Robb
William Robb, ORC Macro, 126 College St. Burlington, VT 05401

Keywords

Sampling, weighting, variance, cumulating, schools, frame overlap

Introduction

It is not uncommon for a survey to require specified levels of precision for a both the general population and certain low incidence sub-populations, in which the latter specification in turn require extensive oversampling. A strategy based on a single sample focusing on the rare populations may fail to achieve the overall precision requirements due to large design effects resulting from modifications necessary to achieve a high degree of oversampling. Coverage of the general population may also be an issue with a single, focused design.

One solution is to use two distinct samples; the first designed to meet the precision requirements for the general population under study, requiring no oversampling; and the second a supplementary sample which is expected to meet the precision requirements for low incidence subpopulation, with this latter sample typically implemented as a very focused oversample. Both samples are drawn independently from sampling frame, with the general population sample typically being drawn from the whole frame, and the oversample being drawn from a subset of the frame containing a higher concentration of the sub-population of interest. The issue then becomes one of combining two independent samples with frame overlap in order to obtain optimal population estimates.

The more general approach, based on known properties of linear combinations of two estimators, provides for a simple weighing adjustment, and can be used even when the details of the weighting are not known, or the frames are not available. However, Pedlow and O'Muircheartaigh (2002) point out that there is an alternate strategy available when the frames are known – that is, when one can calculate every respondent's probability of selection from either frame.

This paper uses data obtained from the Health Behaviors Among School Children (HBSC) survey, a study with a design based on two independent samples, to compare these two methods of weighting samples drawn from overlapping frames. For this paper, both schemes were used to weight the combined sample. The first calculated weights separately, and created a linear combination of weights with coefficients inversely proportional to the effective sample size due to weighting. The second calculated an overall probability of inclusion in one or both samples. Sample efficiency under both schemes was compared via effective sample sizes, and confidence intervals computed for several items.

Methods

The of Health Behaviors in School-age Children is a national survey of sixth to tenth graders which is in turn part of an international study where samples are obtained from a number of countries. The international requirements for the HBSC prescribed overall confidence intervals for the school age children, while the US requirements prescribe confidence intervals for sub-populations defined by race and ethnicity.

Several constraints on the study preclude the simpler methods of meeting these requirements. For inclusion in the international dataset, the US sample must be self-weighting. This requirement eliminates the use of disproportional allocation to focus the sample in regions containing high concentrations of minority students, and makes the use of school enrollment as a size measure in the PPS sampling attractive. School policies require the sampling of intact classes, and thus do not allow targeted sampling of students by race or ethnicity within schools.

For the latest cycle of the HBSC these conflicting requirements were met with a design consisting of two independent samples, both multi-stage cluster samples, with each using school districts, either singly or grouped, as the first sampling stage, schools as the second sampling stage, and finally, intact classes of students as the third sampling stage.

The first sample, (referred to in this paper as the Main sample) was drawn using student enrollment as the size measure for the probability proportional to size sampling employed in the first two sampling stages. Classes, the third stage, were selected with equal probabilities within a sampled school-grade combination. The first stage sampling units were allocate proportionally across strata. This sample yielded 11,732 responding students, 10,454 of which were used in this analysis.

The second sample (referred to in this paper

as the Supplemental sample), designed to oversample minorities, employed a measure of size weighted to increase the selection probabilities for high minority districts and schools in the first two sampling stages. In addition, the private schools were excluded from the frame for the supplemental sample. As with the Main sample, first stage sampling units were allocated proportionally across strata. This sample yielded 5,440 participating students, 4,437 of wich were used in this analysis.

Weighted for providing estimates incorporating data from both samples were computed using two methods. The first, which we refer to as the Combining Samples (CS) method, makes use of the a well know property of linear estimators: Given x1 and x2, estimates for a linear statistic, and a factor f, ranging between 0 and 1, f(x1)+(1-f)x2 is a valid estimate of that statistic. With two samples and a frame overlap, we make use of this by multiplying all weights in the first sample by f, and all weights in the second sample by (1-f). The result is a weight providing estimates based on data combined at the level of the sample, estimates that are certain to be an improvement to one or both of the separate samples.

The optimum coefficient f is one that is proportional to the variances in the two samples. We take the effective sample size, computed as n/deff, where deff is the design effect due to weighting, as a measure of the variance of each sample, as it is impractical to compute an optimum weighting scheme for each estimate individually. This weighting scheme methodology yields proper estimates, and is the best one can do without access to both the frames.

The second weighing scheme examined is an is an alternative that can be though of as cumulating probabilities across cases, referred to in this paper as Cumulating Probabilities (CP). This method is available when one has access to the sampling frames, and is able to compute the inclusion probability of each sampled unit from either frame. Letting p_1 be a unit's probability of selection from the first sample and p_2 the probability of selection from the second, given independent selection, $p_1 + p_2 - (p_1 p_2)$ is the probability that a given unit be found in one or the other or both samples. The inverse of this probability is then taken as the sampling weight.

We note that the weights computed for this analysis were simplified versions of the weights applied to the HBSC public use file, as they did not include adjustments for non-response, and a coarser set of post-stratification cells were employed. The CS weights were post-stratified prior to the computation of f, and the CP weights were post-stratified following the computation of the combined weight.

Results

We first examine the performance of the weights in terms of the design effect due to weighing, and a related measure, the effective sample size, for all students, Hispanic students, and African American students. The following charts present the un-weighted n, along with these measures for both samples individually and for both methods of combining samples.

Effective Sample – All Cases

Weight Summary & Effective Sample Sizes - All Cases				
	Single Sample		Combined Sample	
	Main	Supp.	Combining Samples (C-S)	Cumulating Across Cases (C-P)
n	3,115	1,737	14,891	14,891
CV(w)	1.53	1.24	1.41	1.11
Deff(w)	3.35	2.55	2.99	2.25
n. Eff.	3,115	1,737	4,979	6,607
Gain: Effective Sample Added from 4,437 cases (% of total)			1,864 (37%)	3,492 (52%)

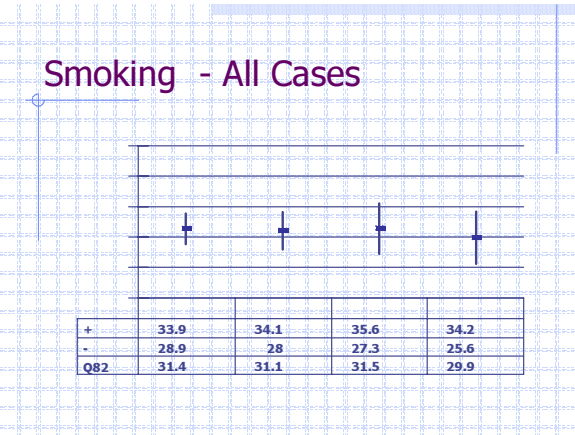
Effective Sample – Hispanics

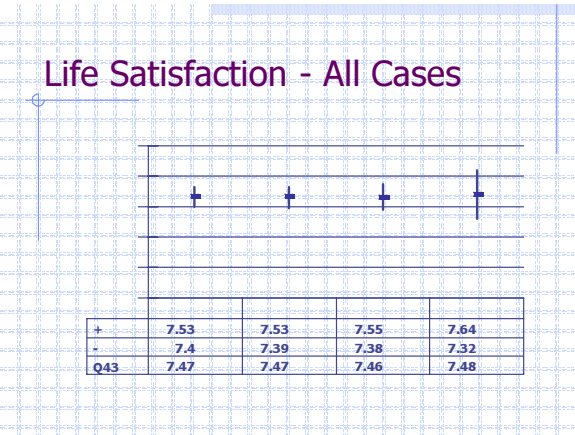
Weight Summary & Effective Sample Sizes - Hispanic				
	Single Sample		Combined Sample	
	Main	Supp.	Combining Samples (C-S)	Cumulating Cases (C-P)
n	1,516	1,405	2,921	2,921
CV(w)	1.47	1.08	1.26	1.18
Deff(w)	3.18	2.15	2.59	2.39
n. Eff.	476	650	1,127	1,221
Gain: Effective Sample Added from 4,437 cases (% of total)			650 (57%)	744 (60%)

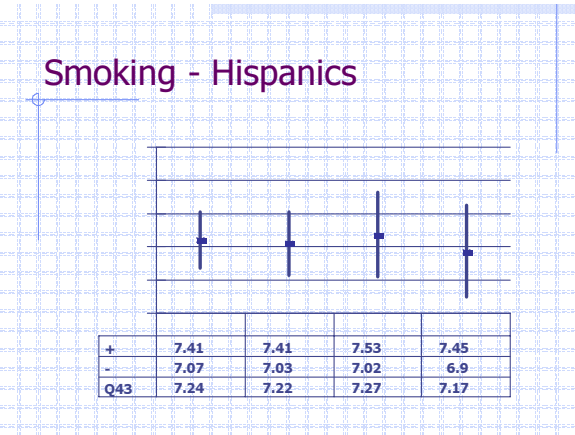
Effective Sample – Afr. Amer.

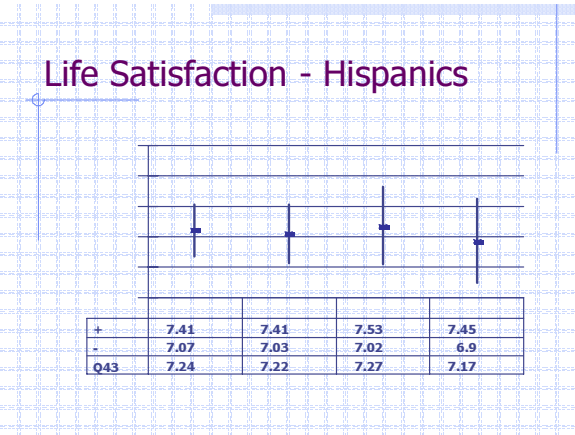
Weight Summary & Effective Sample Sizes - Non-Hispanic African American				
	Single Sample		Combined Sample	
	Main	Supp.	Combining Samples (C-S)	Cumulating Cases (C-P)
n	1,828	1,235	3,063	3,063
CV(w)	1.22	0.97	1.11	0.97
Deff(w)	2.50	1.94	2.24	1.94
n .Eff.	728	634	1,363	1,579
Gain: Effective Sample Added from 4,437 cases (% of total)			634 (46%)	850 (54%)

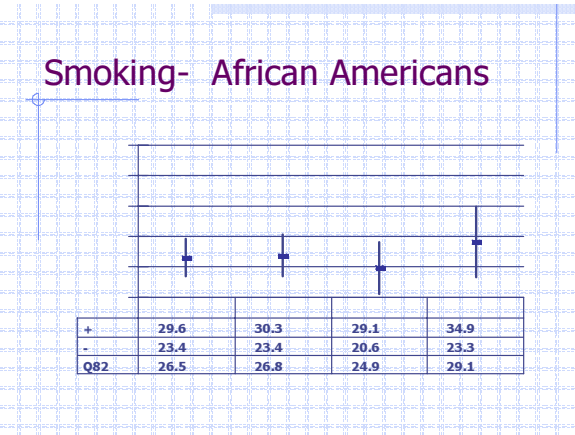
Confidence intervals were computed to examine the effect of both weighting schemes on two estimates, percentage current smokers (defined as 'yes' to an item asking if the student tried a cigarette), mean life satisfaction (a rating on a scale where 0 represents the worst possible life, and 10 represents the best possible life. The following charts present confidence intervals for all students, Hispanic students, and African American students.

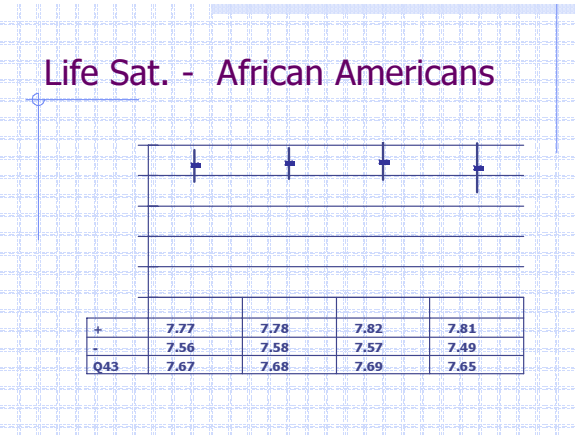












Discussion

The method of weighting the combined sample via cumulating probabilities across cases resulted in a greater increase in effective sample sizes for students as a whole, and for both of the racial/ethnic sub-populations examined. This can be though of in terms of the gain in the overall effective sample size resulting from the addition of the 4,437 cases in the supplemental sample. For all students the resulting increase in effective sample size in the combined file was 1,864 cases for the CS method, and 3,492 cases for the CP method. For Hispanic and African American students the difference in gain was not as dramatic, 651 cases for the CS method, and 745 cases for the CP method for Hispanic students; 635 cases for the CS method, and 851 cases for the CP method. This difference in sample efficiency can be attributed to lower variability in the weights resulting from the CP method. .

Given the large gains in the effective sample sizes due to weighting effects, the differences in the actual confidence intervals for estimates based on survey items were small. The gains varied both with the estimate and sub-population under consideration. While the patterns of confidence intervals varied, in all cases intervals computed for both samples independently and for both combined samples showed a considerable degree of overlap, indicating no statistically significant differences between either estimates methods, or between estimates based on the two samples.

Conclusions

The method of combining independent samples based on overlapping frames via cumulating probabilities across cases appears to make more efficient use of sample via the reduction of variability in the resulting weight, resulting in marginally narrower confidence intervals.

While our results show that this method is preferable from a statistical point of view, operationally it is more complex, requiring detailed knowledge of both frames and the sampling methods employed. Thus it may not always be feasible.

The method of combining samples via a linear combination of individual-sample appropriate weights may be less efficient, but is simpler to implement, and may be employed with any two weighted samples.

We note that this study examines these methods for a few estimates, and in the context of a fairly specialized sample design. Comparisons of the two methods are needed over both a broader range of estimates and sample designs.