An Application of Alternative Weighting Matrix Collapsing Approaches for Improving Sample Estimates

Linda Tompkins[1], Jay J. Kim[2]

[1]Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Room 3115, Hyattsville, MD, 20782

[2]Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Room 3111, Hyattsville, MD, 20782

Abstract

When creating sample weights, most U.S. government agencies combine small race groups such as the American Indians and Asians with Whites disregarding the different coverage ratios of the groups. This paper examines this methodology using the 2003 National Health Interview Survey (NHIS) data of the National Center for Health Statistics (NCHS) and reports the effect on the sample weights and estimates, specifically for Whites, American Indians (AI) and Asians. Two alternative weighting approaches will be used in an effort to reduce the bias.

KEY WORDS: coverage ratio, sample weighting, cell collapsing.

1. Introduction

Before final weights are developed for survey data, a poststratification (ratio or initial adjustment) factor (PSF) is calculated for each cell (row or column) of a weighting matrix and applied to the cell. However, for some cells, poststratification factors cannot be computed. For example, if the sample count is zero for a cell, it is impossible to calculate the PSF because the denominator of the involved fraction is zero. Also if the raw sample count for a fraction is small, the fraction would be considered unstable. Because of these occurrences in many surveys, cells are checked as to whether they have enough raw sample cases to stand by themselves.. Additionally, for most surveys, the cells are checked to see whether its PSF lies within an acceptable range. This ratio criterion assures that the final weights are not too large or too small. It should be noted that very large or small weights can inflate the variance of estimates. If a cell fails either of the above tests, it is combined with another cell.

The cell collapsing strategy described above has merits. However, Kim (2004) raised a potential problem of combining cells which are different in coverage ratios. Let N_i be the control count for cell i, $\hat{N}_i$ the initially weighted sample count for cell i, i =1, 2 and $f_i = \frac{N_i}{\hat{N}_i}$, i =1, 2, the Initial Adjustment Factor (IAF) for cells 1 and 2. Then $\frac{1}{f_i}$, i = 1, 2, is the coverage ratio for cell i, i = 1, 2. Let $N_2 = c\,N_1$. The PSF for the combined cell was expressed by Kim (2004) as:

For cell 1: $$\frac{f_2(1+c)}{cf_1 + f_2} f_1, \qquad (1)$$

and

for cell 2: $$\frac{f_1(1+c)}{cf_1 + f_2} f_2. \qquad (2)$$

Before collapsing, the PSF for cell 1 is f_1. However, because of collapsing, as shown in equation (1), f_1 is modified by $\frac{f_2(1+c)}{cf_1 + f_2}$, which is called the Collapsing Adjustment Factor (CAF) for cell 1 by Kim, et all (2005). Similarly, for cell 2, the CAF is $\frac{f_1(1+c)}{cf_1 + f_2}$. Using the above formulas, we can make the following observations: when c = 10 and $\frac{f_1}{f_2}$ = 4.0, cell 1 will lose 73 percent of its own weight to cell 2. For the same c, if $\frac{f_1}{f_2}$ = .25, cell 1 will gain an additional 214 percent of its own weight from cell 2. Note that this weight shift is artificial. Thus, Kim (2004) and Kim and Tompkins (2007) claimed that the current approach of cell collapsing can introduce bias, which can often be large.

Most surveys collapse a cell (row or column) with another if the PSF (ratio) for the cell is greater than 2. This standard collapsing procedure allows the PSF of the poorly covered cell to decrease below 2. Hence, Kim (2004) proposed to truncate (censor) the PSF for the cell at 2 to make sure that the PSF for that cell is 2 or at least 2, depending on the method. Kim, et al (2007) implemented these two approaches of weight truncation in their simulation studies and found that the latter

outperforms the former and the standard collapsing procedure.

When creating sample weights, most U.S. government agencies combine small race groups such as the American Indians and Asians with Whites disregarding the different coverage ratios of the groups. This paper examines this methodology using the 2003 National Health Interview Survey (NHIS) data of the National Center for Health Statistics (NCHS) and reports the effect on the sample weights and estimates, specifically for Whites, American Indians (AI) and Asians. Two alternative weighting approaches will be used in an effort to reduce the bias.

2. Cell Collapsing and Alternative Weighting Approaches

The NHIS uses the following weighting matrix:

Table 1. Weighting Matrix

	Hispanic		Non-Hispanic Black		Non-Hispanic Other	
	M	F	M	F	M	F
< 1 yr						
1 - 4						
5 - 9						
10-14						
15–19						
.						
.						

In the above table, M stands for male and F for female.

The non-Hispanic other category, as mentioned before, includes all non-Hispanic races other than non-Hispanic Blacks, (i.e., it includes Whites, American Indians, Asians, Native Hawaiian and Pacific Islanders and all multiple race groups).

It is interesting to see how much the coverage ratios differ among the race groups in the "others" race category. Tables 2a and 2b present coverage ratios for Whites, American Indians (AI) and Asians by age categories from the 2003 NHIS.

Table 2a. Coverage Ratios for 2003 NHIS - Males

Age Group	White	AI	Asian
< 1	.85	.17	.33
1 - 4	.80	.44	.66
5 - 9	.79	.88	.59
10 - 14	.80	.65	.54
15 - 17	.84	.46	.75
18 – 19	.61	.26	.55
20 – 24	.59	.55	.51
25 - 29	.60	.44	.31
30 - 34	.67	.54	.65
35 - 44	.67	.32	.53
45 - 49	.65	.51	.63
50 - 54	.67	.54	.57
55 - 64	.70	.53	.47
65 - 74	.75	.44	.44
75+	.71	.51	.65

Table 2b. Coverage Ratios for 2003 NHIS - Females

Age Group	White	AI	Asian
< 1	.82	-	.38
1 - 4	.80	.43	.71
5 - 9	.84	.70	.78
10 - 14	.76	.95	.70
15 - 17	.77	.25	.67
18 – 19	.72	.10	.50
20 – 24	.59	.57	.50
25 - 29	.68	.39	.56
30 - 34	.75	.46	.57
35 - 44	.76	.59	.59
45 - 49	.76	.36	.67
50 - 54	.80	.31	.53
55 - 64	.78	.62	.45
65 - 74	.75	.12	.48
75+	.76	.36	.64

In Table 2a, except for one age group (5 – 9 years), White males always have higher coverage ratios than American Indian males. Also, White males always have higher coverage ratios than Asian males, without exception. One extreme case is age group less than 1, where the coverage ratio for White males is .85, while that for American Indians is .17. The coverage ratio for American Indian males age < 1 is only 1/5 of that for Whites. For the same age group, the Asian coverage rate is less than half that of Whites. Of 15 male age groups, 7 age groups have coverage ratios less than .5 for American Indians. For the 18 – 19, 20 – 24 and 25 – 29 years age groups, coverage ratios for Whites are also low, but those for American Indians and Asians are even lower, sometimes less than half of that for Whites.

As for females in Table 2b, Whites always have higher coverage ratios than American Indians, with one exception (10 – 14 years of age). Also, Whites are better covered than Asians for all age groups. For the 18 – 19 years age group, Whites have a coverage rate which is more than 7 times better than that of American Indians. For the 65 – 74 year age group, the coverage ratio for Whites is more than 6 times that of American Indians. Quite often the White coverage rate is much better than that of American Indians.

The following example demonstrates the effect on weights and estimates when two cells with very different coverage ratios are combined.

Example 1. Suppose we have the following initially weighted sample counts, control counts and the initial adjustment factors for 2 cells, one for Whites and the other for American Indians in Table 3.

Table 3. Sample Weighting Data

	$\hat{N}_i$	N_i	f_i
AI	50	300	6
White	17,000	20,000	1.17647

When White and American Indian cells in the above table are combined, the new PSF for the combined cell is

$$\frac{300+20,000}{50+17,000}=1.1906158 \qquad (3)$$

The original PSF for American Indians was 6, but the new PSF is 1.1906158. Hence, the new weighted total for American Indians is $1.1906158\times 50\approx 60$. Since the control count is 300, we observe an underestimation of 240, which equates to an 80 percent underestimation of American Indians in this cell. On the other hand, the original PSF for Whites is 1.18, but the new PSF is 1.1906158. Thus, the new weighted total is 20,240, which is greater than the control count (20,000). In other words, Whites picked up an additional weight of 240 due to collapsing. This amount is 1.2 percent of the control count (20,000). Note that a 1.2 percent overestimation for Whites is negligible, but an 80 percent underestimation for American Indian is large.

In fact, the Collapsing Adjustment Factors (CAFs) for cells 1 and 2 from equations (1) and (2) have been implicitly applied to f_1 (6) to reach 1.1906158 in equation (3). That is, the CAF for cell 1 is:

$$\frac{1.17647(20,000/300+1)}{6(20,000/300)+1.17647}=.1984358 \qquad (4)$$

The new PSF for cell 1 is

$$6(.1984358)=1.1906148.$$

There is a slight difference between the values in equations (3) and (4), which is due to rounding error.

As mentioned before, the category of White males age <1 has a much higher coverage ratio than American Indians and Asians. The same observation can be made for females. Consequently, both White males and females age <1 were overestimated by 7 percent in 2003. For both genders, in all except two age groups, Whites are better covered than American Indians, which causes the former to absorb weights from the latter. As a result, American Indians, overall, were underestimated by 29.7 percent, as will be seen in section 3. Similarly, Asians were underestimated by 20.7 percent.

To rectify this problem, we propose two alternative weighting procedures. The first is to weight American Indians and Asians independently. American Indians had 197 raw sample cases, which is enough for independent sample weighting. The number of sample persons is 1,200 for Asians, which is more than enough for independent sample weighting.

The second procedure is to artificially inflate to .5 the coverage ratios which are originally lower than .5. This is to "protect" the sample cases in the cells whose coverage ratios are too low, or whose PSF is too high. This approach is to ensure that the final weighted total in the cell is at least half the control count. According to this approach, the PSF can sometimes go much higher than 2. This approach is somewhat consistent with the weight truncation approach by Kim, et al (2007). They considered two approaches of weight truncation: one allows PSF to go over the threshold (2), but the other does not. The approach proposed here is similar in spirit to the former. The protection of the weights in the poorly covered cells is greater in the approach proposed here because the PSF for this new approach can increase much more than that considered by Kim, et al.

Example 2 (Table 4) numerically illustrates the approach proposed here.

Table 4. Sample Weighting Data

	$\hat{N}_i$	N_i	f_i
AI	~~50~~ 150	300	~~6~~ 2
White	17,000	20,000	1.17647

In Table 4, we set f_i for American Indian equal to 2, instead of 6 as in Table 3. To do so, we had to multiply $\hat{N}_i$ (50) by 3 to make it 150. In other words, to make sure that f_i = 2, we had to artificially inflate $\hat{N}_i$ by a factor of 3. If the original f_i were 3 (this means $\hat{N}_i$ = 100), then we had to artificially inflate $\hat{N}_i$ by a factor of 1.5, instead of 3.

When White and American Indian cells in the above table are combined, the new PSF for the combined cell is

$$\frac{300+20,000}{150+17,000}=1.18367 \qquad (5)$$

The new PSF for Whites is 1.18367, but that for American Indians is 3(1.18367) = 3.55101. Compare 1.1906158 to 3.55101 for the American Indian cell's PSF. The new cell estimate for American Indians is 50(3.55101) = 177.5505. Since the control count is 300, we observe an underestimation of 122, which equates to an approximate 41 percent underestimation of American Indians in this cell. This is a big improvement in comparison to the result of the original cell collapsing approach.

3. Alternative Sample Weighting

When independently weighting the sample for American Indians and Asians, a minimum raw sample count of 20 was used for cell collapsing. That is, starting with the age group <1 cell, if a raw sample count was less than 20 for a cell, it was combined with the next nearest cell. It should be noted that no artificial inflation of the weights was done while combining cells in each of the race groups. Artificially inflating the weights was, however, employed in collapsing American Indians and Asians with Whites. After weighting was completed, weights for each sample unit were accumulated for American Indians and Asians, where the results are shown in Tables 5 and 6, respectively.

Table 5. American Indian Weighting (in 1,000's)

	Total Weight	Control Count
Current	1,496 (-29.7%)	2,127
Inflated	1,752 (-17.4%)	2,127
Independent	2,127	2,127

As the Table 5 shows, when we rely on the current weighting procedure, i.e., when American Indians are collapsed with Whites for weighting, the weight total for American Indians is 29.7 percent lower than its control count. On the other hand, when a special measure was taken to protect the weights in the cells whose coverage ratios were lower than .5, the weight total improved over the current approach by 12.3 percent. However, the inflation approach still underestimates the control count by 17.4 percent. There are two reasons for this. First, we did not take any measure to protect the cells whose coverage ratios were higher than .5, even if coverage ratio for American Indians was lower than that for Whites. Second, even if we gave higher PSF's to cells whose coverage ratios were lower than .5, we did not raise the ratio all the way to the same level as that for Whites.

As can be predicted, when the independent weighting approach was used, the total weight is the same as control.

Table 6. Asian Sample Weighting (in 1,000's)

	Total Weight	Control
Current	9,369 (-20.7%)	11,817
Inflated	9,753 (-17.5%)	11,817
Independent	11,817	11,817

As shown in Table 6, when Asian cells are collapsed with Whites for weighting, as in the current approach, Asians are underestimated by 20.7 percent. Note that this underestimation rate is better than that for American Indians. This is because Asians, in general, have better coverage ratios than American Indians for both genders.

When the inflation approach was used, the weighted total improved over the current approach by only 3.2 percent. This improvement is much lower than that observed for American Indians. The difference is due to the fact that 16 out of 30 American Indian age groups have coverage ratios less than .5, but for Asians, the same observation could be made for only 7 age groups.

Prevalence rates were calculated for 4 health characteristics based on the three cell collapsing approaches: diabetes, health insurance coverage, overnight hospital stay and asthma. It should be noted that one rate for each race was computed just as in published survey reports.

Table 7 presents prevalence rates for American Indians.

Table 7. Prevalence Rates for American Indians – Weighted Total as Denominator

	Current	Inflated	Independent
Diabetes	9.22	9.43	10.28
Health Insurance Coverage	64.90	63.72	65.33
Overnight Hospital Stay	7.73	8.67	8.25
Asthma	17.41	16.32	18.04

In Table 7, for all 4 health characteristics, the prevalence rate for the independent weighting approach is higher than that for the current weighting approach. The biggest difference can be observed for diabetes. The independent weighting approach provides the prevalence rate for diabetes more than 1 percentage (in absolute term) higher than the current approach. It is 11 percent higher in relative term. The inflation approach's rate is higher for 2 characteristics than the current approach's rate, but

it is lower than the independent approach's rate. However, for 2 other characteristics, the prevalence rate for the truncation approach is lower than that of the current approach.

Table 8 presents prevalence rates for Asians.

Table 8. Prevalence Rates for Asians – Weighted Total as Denominator

	Current	Inflated	Independent
Diabetes	4.35	4.50	4.70
Health Insurance Coverage	83.49	83.44	83.70
Overnight Hospital Stay	4.85	5.05	5.09
Asthma	5.96	5.84	5.83

As shown in Table 8, the prevalence rate for the independent weighting approach is higher than that for the current weighting approach except for asthma. The truncation approach provides prevalence rates closer to that of the independent weighting approach for all variables, except for health insurance.

The difference for the prevalence rates between the current and the independent weighting approach for Asians is much smaller than that for American Indians. This may be due to the fact that the coverage ratios for Asians are much more stable than those for American Indians.

Note that in calculating the prevalence rates in Tables 7 and 8, estimated counts were used for both numerators and denominators. However, control (population) counts instead of estimated counts (weighted totals) can be used for the denominator, while estimated counts are still used for numerator. For example, suppose researchers want to calculate the prevalence rates for American Indians or Asians residing in certain age groups regions of the nation, since NCHS' report does not show the rates for regions. To do so, they can cumulate weights of, for example, diabetic people in the regions and compute the prevalence rates using the cumulated weights as the numerator and the population count as the denominator. The following two tables show the prevalence rates calculated in that manner:

Table 9. Prevalence Rates for American Indians – Control Count as Denominator

	Current	Inflated	Independent
Diabetes	6.48	7.77	10.28
Health Insurance Coverage	45.65	52.49	65.33
Overnight Hospital Stay	5.44	7.14	8.25
Asthma	12.25	13.44	18.04

Tables 7 and 9 show the prevalence rates for American Indians. The rates in Table 7 are computed with the weighted total in the denominator and those in Table 9, with the population count in the denominator.

The rates in Table 9 are much lower than those in Table 7, except for those for the independent weighting method, which are the same. The rate for the current approach in Table 9 is 29.7 percent lower than that in Table 7 for each of the four health characteristics. Similarly, the rates for the inflation approach in Table 9 are 17.6 percent lower than those in Table 7.

In Table 9, the rates for the current approach are almost one third lower than those for the independent weighting approach. The rates for the inflation approach are between the two approaches.

Table 10. Prevalence Rates for Asians – Control Count as Denominator

	Current	Inflated	Independent
Diabetes	3.45	3.71	4.70
Health Insurance Coverage	66.19	68.87	83.70
Overnight Hospital Stay	3.85	4.17	5.09
Asthma	4.73	4.82	5.83

Both Tables 8 and 10 show the prevalence rates for Asians. The relationship between Tables 8 and 10 is the same as that between Table 7 and Table 9.

The rates in Table 10 are much lower than the rates in Table 8, except for those for the independent weighting method, which remains the same. The rate for the current approach in Table 10 is 20.7 percent lower than that in Table 8 for each of the four health characteristics. Similarly, the rates for the inflation approach in Table 10 are 17.4 percent lower than those in Table 8. Again, these differences are due to the different denominators, that is, the weighted total or the control count.

Comparisons between the rates in Table 7 and the rates in Table 9 and between the rates in Table 8 and the rates in Table 10 show that when the prevalence rates are calculated it is better to use the weighted totals as the denominator for American Indians and Asians.

4. Concluding Remarks

Thus far, we have observed that combining cells with varying coverage ratios results in under- and over-estimation of population (control) counts. In order to

alleviate this problem, we proposed independent weighting and weight inflation approaches for collapsing cells, implemented these approaches using NHIS data and compared them with the current weighting procedure. Currently, American Indians and Asians are combined with Whites for sample weighting. However, coverage rates for Whites are better, often much better, than those for American Indians in 28 out of 30 age groups. Coverage rates for 3 age groups for American Indians are extremely low, i.e., they are in the 10 – 17 percent range, while they are at least 72 percent for Whites. Because of this, the current weighting approach underestimated American Indian by 29.7 percent. Also Whites consistently had better coverage ratios than Asians, and as a result, the current weighting approach underestimated Asians by 20.7 percent.

We also estimated the prevalence rates for diabetes, health insurance coverage, overnight hospital stay and asthma using the weights developed by three different independent weighting approach, except for health insurance.

The prevalence rate can be calculated using two methods. One is to use weighted counts for both numerator and denominator, and the other is to use weighted counts for the numerator, but population counts for the denominator. The first approach was used for the tables above. However, if the second approach were to be used, the rates would be underestimated by 29.7 percent for American Indians and by 20.7 percent for Asians with the current collapsing approach and 17.7 percent and 17.4 percent, respectively, with the inflation approach. This is because their weighted totals are lower than their respective population counts. Thus, the first approach is recommended for computing the prevalence rates.

The public use micro data (PUM) file from the survey data we used for this study has been released to the general public. Note that the PUM file contains sample weights for sample persons in the file. Some data users of the PUM file might want to accumulate weights for American Indians or Asians, say with diabetes, to come up with the number of diabetic American Indians or Asians in the nation or some region of the nation. However, the result would be a gross underestimation of the true values for the reason mentioned above. A better approach of getting the number of diabetic American Indians or Asians in the nation or a region would be to calculate the prevalence rate using weighted counts for both the numerator and the denominator and to then multiply the rate by the American Indian or Asian population count, respectively.

In conclusion, the independent weighting approach for American Indians and Asians may produce more realistic weights, and therefore, more accurate estimates. In addition, the current approach appears to underperform when compared to the inflation approach, even though the latter can be further fine tuned.

cell collapsing approaches. For all 4 health characteristics, American Indians show higher prevalence rates when they are weighted independently than when they are weighted as a part of the "Other" race category (i.e., when they are weighted while combined with Whites). The American Indian diabetes prevalence rate is more than 1 percent higher when the independent weighting approach is used (10.28 %) than when current weighting approach is used (9.22 %). The weight inflation approach shows mixed results for American Indians. For 2 characteristics, the weight inflation approach showed higher prevalence rates than the current weighting approach, whereas for 2 others, the reverse was observed.

For Asians, the prevalence rate for the independent weighting approach is higher than that for the current weighting approach, except for asthma. The inflation approach provides prevalence rates closer to that of the

5. References

Kim, J. J. (2004). Effect of collapsing rows/columns of weighting matrix on weights. P*roceedings of the Section on Survey Methods Research*, American Statistical Association CD.

Kim, J.J., Li, J., and Valliant, R. (2007). Cell collapsing in poststratification, to be published in *Survey Methodology*.

Kim, J.J. and Tompkins, L. (2007). Comparisons of current and alternative collapsing approaches for improved health estimates. Paper presented at the *11th Biennial CDC/ASTDR Symposium on Statistical Methods*, in Atlanta, Georgia, April 17-18, 2007.

DISCLAIMER: The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.