Estimation and Reliability Issues of Health Estimates from the Behavioral Risk Factor Surveillance System for U.S. Counties Contiguous to the United States-Mexico Border

Joe Fred Gonzalez, Jr.[1], Machell Town[2], Jay J. Kim[1], Sam Notzon[1], and Juan R. Albertorio[1]
Centers for Disease Control and Prevention
National Center for Health Statistics[1]
National Center for Chronic Disease Prevention and Health Promotion[2]

Summary[1]

The Behavioral Risk Factor Surveillance System (BRFSS) is a State telephone based survey of the civilian non-institutionalized adult (18 years and over) population residing in the United States. Consequently, the BRFSS final weights that are currently available in the data files are designed to produce unbiased estimates of socio-demographic and health characteristics for adults at the State level (Gonzalez, et al, 2005). In addition to State level BRFSS estimates, there is interest in the health status of adults residing in the 25 U.S. counties contiguous to the United States-Mexico Border region (Arizona, California, New Mexico, and Texas.) The purpose of this paper is to investigate alternative ways of arriving at post-stratification factors (ratio adjustments) by collapsing the weighting matrix by age-sex-ethnicity/race for producing final weights/estimates for this border region. An optimal approach which minimizes local (cell) squared bias was applied to BRFSS data (25 contiguous counties). Then, a conditional mean square analysis was used to observe the effect of cell collapsing (in tandem with the optimal bias approach) on the absolute bias and variance estimators for several BRFSS socio-demographic and health characteristics.

Keywords: Cell Collapsing, Post-stratification Factors, Collapsing Adjustment Factors, Bias, Variance

1. Introduction

The BRFSS is a State telephone based survey of the civilian non-institutionalized adult (18 years and over) population residing in the United States. However, there is interest in another geographical subpopulation, the 25 U.S. counties contiguous to the United States-Mexico Border (Arizona, California, New Mexico, and Texas.). The map in Figure 1 displays the "sister cities" along both sides of the United States-Mexico Border region. Figure 2 shows a map of the actual counties that are contiguous to the United States-Mexico Border region.

It was determined that it would be worthwhile to produce BRFSS estimates for the adult population in the border region by certain age-sex-ethnicity/race cells. The desired six age groups were: 18-24, 25-34, 35-44, 45-54, 55-64, and 65 and over. The desired three ethnicity/race groups were: Hispanic, White Non-Hispanic, and Non-Hispanic Black/Multiracial and others. In previous work (Gonzalez, et al, 2005), BRFSS sample counts were tabulated by age-sex-ethnicity/race within each border county. Although sample counts were insufficient for some cells within each border county for the current estimation research, BRFSS county level estimation techniques have been investigated (Jia, et al, 2004) and have been produced (Jia, et al, 2006). For detailed documentation for producing county level estimates, the reader is referred to: BRFSS's SMART (Selected Metropolitan/Micropolitan Area Risk Trends) data from metropolitan/micropolitan statistical areas. The URL for these data is http://apps.nccd.cdc.gov/brfss-smart/SelMMSAPrevData.asp. The SMART home page is http://apps.nccd.cdc.gov/brfss-smart/index.asp. For the current estimation research, sample sizes were aggregated by the desired age-sex-ethnicity/race cells for the 25 counties contiguous to the United States-Mexico Border (Arizona, California, New Mexico, and Texas.). At the border region level, cell sizes were sufficiently large for the desired age-sex-ethnicity/race cells for both Hispanics and White Non-Hispanics, and in a few instances for Non-Hispanic Black/Multiracial and others. This level of geographical aggregation was defined as the United States-Mexico Border for the purpose of our paper. Hereafter, the United States-Mexico Border Region will be simply referred to as the "border region" and is similarly defined for each of the three years 2001-2003. In addition, the same age-sex-ethnicity/race crosstabulation that was used for

[1] **Disclaimer:** This paper represents the views of the authors and should not be interpreted as representing the views, policies or practices of the Centers for Disease Control and Prevention, National Center for Health Statistics, or the National Center for Chronic Disease Prevention and Health Promotion.

determining sample size sufficiency was also used as the weighting matrix for this investigation.

As mentioned earlier, the focus of this paper is to investigate alternative ways of arriving at post-stratification factors (ratio adjustments) by collapsing rows or columns by age-sex-ethnicity/race for producing final weights/estimates for this border region. An optimal approach which minimizes local (cell) squared bias was applied to BRFSS data (25 contiguous counties). Then, a conditional mean square analysis was used to observe the effect of cell collapsing (in tandem with the optimal bias approach) on the absolute value of the bias and variance estimators for several BRFSS socio-demographic and health characteristics.

2. Sample Weighting Procedures for the Border Region

Post-stratification is used for incorporating population distributions of key socio-demographic variables into survey estimates. One reference about post-stratification is Kim (2004) "Effect of Collapsing Rows/Columns of Weighting Matrix on Weights."

For this analysis, the variable _WT2, which is available in the 2001-2003 BRFSS data sets is the initial sample weight as follows:

_WT2 =_STRWT * NAD / NPH

where,

STRWT = within State stratum weight,
NAD = number of adults in household, and
NPH = number of phones in the household.

For purposes of this investigation, the initial sample weight (_WT2) was used to create the "initial poststratification factors (PSF)" which were calculated in the usual manner by age (6 groups)-sex(2)-ethnicity/race (Hispanic, White Non-Hispanic, and Non-Hispanic Black/Multiracial and Others) as follows:

PSF = Census pop. count within an i-th cell / sum of _WT2 within same i-th cell.

Table 1 shows the initial 2003 poststratification factors (PSF) that were multiplied by the basic sampling weights (_WT2) for adults to produce unbiased estimates of adult health characteristics in the border region. Similarly, "initial poststratification factors (PSF)" were calculated for the years 2001-2002, but are not shown in this paper due to space limitations.

The "initial poststratified Final Weights" used in this investigation were calculated in the same fashion for each of the three years 2001-2003 as follows:

"Final_ Weight" = _WT2*PSF

where PSF is as previously defined.

The usual approach, c*onventional cell collapsing* was used. This approach is usually driven by sample size considerations (here, minimum cell count, raw cell count = 20), and maximum ratio criteria (original PSF) by domains, and row adjacency. Table 2 shows the maximum ratio criteria used for conventional collapsing by year (25 contiguous counties) .

The "Final_ Weights" were used to produce BRFSS percent estimates of adult characteristics for all three years (2001- 2003) using the following binary health variables for adults (18+ years of age):

- Ever had Asthma
- Ever had high blood pressure (not available for 2002)
- High cholesterol (not available for 2002)
- Diabetes
- Having health insurance
- Current smoker
- Any exercise.

3. Conditional Bias and Mean Square Error Analysis

First, we will introduce the usual notation involved in doing a mean square error (MSE) analysis as follows:

$$MSE(p) = [Bias(p)]^2 + [se(p)]^2$$

where p= percent estimator of a health characteristic, and

se (p)= standard error estimator for the same health characteristic.

The percent estimates of health characteristics using the "initial poststratified Final_Weights" are unbiased estimates and treated as "*parameters,"* that is, as true values of health characteristics for the adult population in the border region for this mean square error (MSE) analysis. So, in reality, the bias, variance, and the mean square error (MSE) analysis is *conditiona*l. The bias and MSE analysis was performed by comparing these "*parameters"* of health characteristics with corresponding percent estimates of health characteristics generated by: applying the *local (cell)*

minimum bias strategy followed by investigating the effects on the MSE of the same estimates.

"New" PSF, corresponding Final Weights, and corresponding percent estimates were produced by using the above sequential approach.

Table 3 (Kim, 2004) defines the quantities that are involved for producing PSF using *conventional cell collapsing*.

Table 4 shows an example of initial PSFs for row 1 and row 2 where PSF = $f_i = N_i / W_i$ where all quantities are as previously defined in Table 3.

For the sake of illustration, suppose that we collapse row 1 and row 2 and assume that the row population counts are the same, that is, $N_1 = N_2$ (Kim, et al, 2006). What would be the revised PSF for each row? The revised PSF for row 1 would be

$$\frac{N_1 + N_2}{W_1 + W_2} = \frac{2N_1}{5W_1} = \frac{2}{5} f_1 \ .$$

That is, by collapsing rows 1 and 2, row 1 has lost 3/5 of its original population count.

Similarly, the revised PSF for row 2 would be

$$\frac{N_1 + N_2}{W_1 + W_2} = \frac{8N_2}{5W_2} = \frac{8}{5} f_2 \ .$$

That is, by collapsing rows 1 and 2, row 2 has gained 3/5 of its original population counts.

The ratios 2/5 and 8/5 are referred to as the *collapsing adjustment factors (CAFs)*.

A generalization of *CAFs* by Kim (2004) follows. Let

$N_2 = c\, N_1$ where $c > 0$.

The revised PSF in terms of the original PSF (f_1) for row 1 is:

$$\frac{N_1 + N_2}{W_1 + W_2} = \frac{f_2(1+c)}{cf_1 + f_2} f_1$$

where

$$\frac{f_2(1+c)}{cf_1 + f_2} = CAF_1 \ \ for\ row\ 1.$$

Similarly,

$$\frac{f_1(1+c)}{cf_1 + f_2} = CAF_2 \ \ for\ row\ 2.$$

What follows is a possible remedy to avoid shifting potentially large population counts from one row to another when collapsing rows is to apply a local (cell) minimum bias strategy. In this strategy, an expression for the squared bias was developed by multiplying CAF_1 by a constant (k), that is,

$$\left[k \frac{f_2(1+c)}{cf_1 + f_2} N_1 \bar{x}_1 + \frac{f_1(1+c)}{cf_1 + f_2} N_2 \bar{x}_2 - (N_1 \bar{x}_1 + N_2 \bar{x}_2) \right]^2 .$$

Then, this expression was differentiated with respect to k and set equal to zero. Finally, the value of k that minimizes the squared bias was found. For further details of this strategy, see Kim, et al, 2006.

4. Results

Using the t-distribution and the p-value approach, the following right-tailed test of the hypothesis for one population proportion was conducted:

H_0: P = 0.5
Ha: P > 0.5

where P is the proportion of the time the minimum bias approach performed better.

Table 5 shows the overall summary for the performance of |Bias|, that is, better, worse, or the same by the # of sub-domains for each level of performance. The value of the test statistic t=3.074 and the corresponding p-value < 0.005. Therefore, we reject he null hypothesis, H_0 , in favor of the alterative hypothesis, Ha.

A similar type of analysis was performed for investigating MSE. The value of the test statistic t=1.818 and the corresponding p-value < 0.05. Therefore, we reject the null hypothesis, H_0 , in favor of the alterative hypothesis, Ha.

Table 6 shows the overall summary for MSE in the same format as Table 5.

5. Conclusion and Further Research

As mentioned earlier, the purpose of this paper is to investigate alternative ways of poststratifying the BRFSS basic sampling weight (_WT2) for producing health estimates for the 25 U.S. counties (taken as a whole) contiguous to the U.S.-Mexico Border.

Results from this investigation indicate that in general when calculating PSFs, the collapsing adjustment factor (CAF) approach along with the local (cell) minimum bias approach is significantly better than the conventional collapsing approach in terms of bias and MSE. An area for further research is to investigate other approaches for modifying PSFs, for example, local (cell) minimization of MSE, and global (over an entire dataset) minimization of bias and MSE.

References

Gonzalez, Joe F.; Town, Machell; Kim, Jay J. (2005) Mean Square Error Analysis of Health Estimates from the Behavioral Risk Factor Surveillance System for Counties along the United States-Mexico Border Region, Proceedings of the American Statistical Association, Survey Research Methods Section.

Jia, Haomiao; Muennig Peter; Borawski, Elaine. (2004) Comparison of Small-Area Analysis Techniques for Estimating County-Level Outcomes, American Journal of Preventive Medicine; 26 (5):453-60.

Jia, Haomiao; Link, Michael; Holt, James.; Mokdad, Ali H.; Li, Lee; Levy, Paul S. (2006) Monitoring County-Level Vaccination Coverage During the 2004-2005 Influenza Season, American Journal of Preventive Medicine; 31 (4):275-280.

Kim, Jay J. (2004) Effect of Collapsing Rows/Columns of Weighting Matrix on Weights, Proceedings of the ASA Survey Research Methods Section.

Kim, Jay J.; Valliant, Richard; Zha, Wenxing, (2006) *Cell Collapsing Strategies based on Collapsing Adjustment Factor.* Paper presented at the Joint Statistical Meetings, Seattle, Washington, August 8, 2006, and to appear in the Proceedings of the American Statistical Association, Survey Research Methods Section).

Figure 1.

Figure 2.

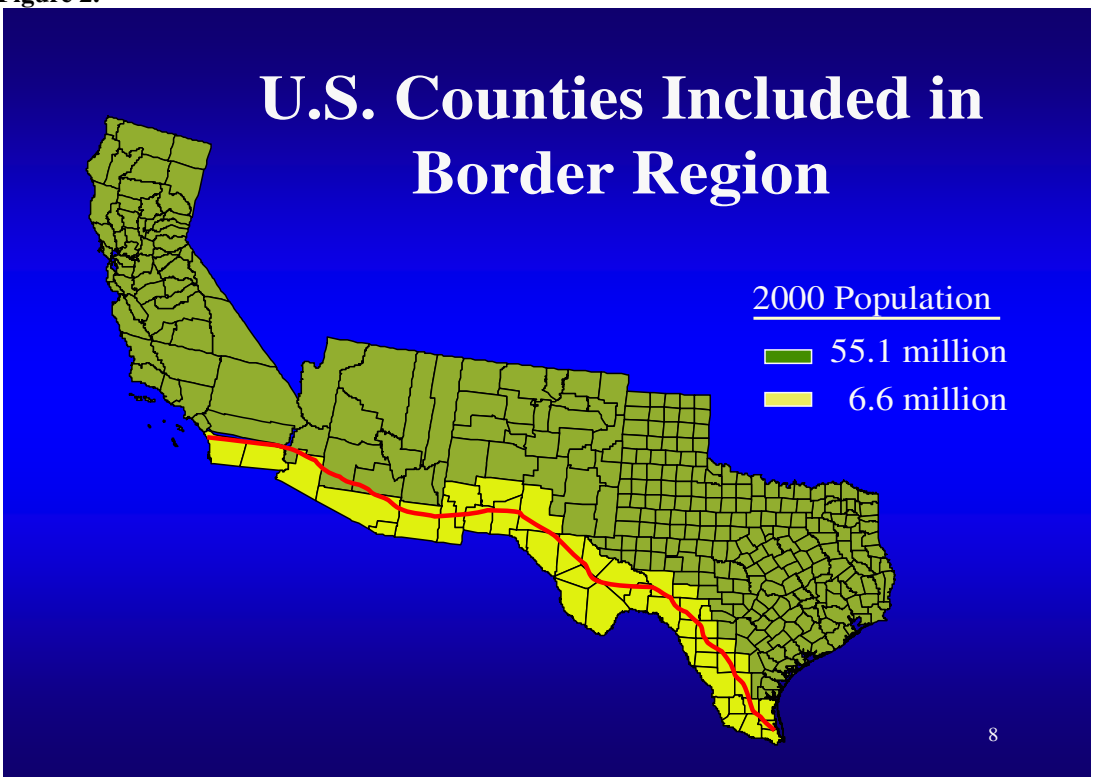

Table 1. Initial 2003 Post-stratification Factors (PSF) by Age-Sex-Ethnicity/Race for the Border Region (25 contiguous counties).

Age and Sex	Ethnicity/Race		
	Hispanic	White Non-Hispanic	Non-Hispanic Black/Multiracial/Other
Male			
18-24	6.1	5.2	7.7
25-34	5.1	5.1	9.6
35-44	5.5	3.5	2.7
45-54	4.7	4.3	2.9
55-64	3.8	4.6	2.4
65+	9.3	3.1	4.3
Female			
18-24	4.4	3.8	4.1
25-34	2.7	3.2	6.4
35-44	3.2	3.3	4.1
45-54	2.6	2.8	7.7
55-64	3.6	2.9	4.2
65+	4.4	3.6	5.6

Table 2. Maximum Ratio Criteria Used for Conventional Collapsing by Year (25 contiguous counties)

Ethnicity/race	Year		
	2001	2002	2003
Hispanic	4	5	7
White (Non-Hispanic)	4	4	5
Non-Hispanic Black/Multiracial/Others	8	7	6

Table 3. Weighting Matrix for Calculating Usual PSF.

Rows	Raw Sample Count	Initially Weighted Sample Count	Control Count
Row 1	N_1	W_1	N_1
Row 2	N_2	W_2	N_2

Table 4. Conventional Collapsing Example (PSF = $f_i = N_i / W_i$)

Row 1	$f_1 = 4$
Row 2	$f_2 = 1$

Table 5. Overall Summary for |Bias| Analysis

Performance	# of Sub-Domains
Better	59
Worse	30
Same	14

Table 6. Overall Summary for MSE Analysis

Performance	# of Sub-Domains
Better	58
Worse	40
Same	5