OUTLIER WEIGHT ADJUSTMENT IN REACH 2010

Steven Pedlow, Javier Porras, Colm O'Muircheartaigh, Hee-Choon Shin, NORC
Steven Pedlow, National Opinion Research Center, 55 E. Monroe, Chicago, IL 60603

Key Words: Extreme weights, Winsorization, Capping weights.

1. Introduction to REACH. Racial and Ethnic Approaches to Community Health: 2010 (REACH) is a project sponsored by the Centers for Disease Control (CDC) with the goal to eliminate racial and ethnic disparities in health by 2010. REACH is a community-based program: local community groups across the United States applied for funds to design and implement a local health intervention. These interventions target one or more health priority areas (diabetes, cardiovascular disease, breast and cervical cancer, HIV/AIDS, and adult and childhood immunization) and one or more race-ethnicity groups (African-American, Hispanic, Asian, Native American, and Pacific Islander). The goals of the interventions are to increase community awareness and knowledge about the health priority issue and how to prevent and combat these health problems, as well as to improve medical care access for the targeted race-ethnicity group.

The role of the National Opinion Research Center is to conduct interviews in each community within REACH (twenty-one in the first year) to collect information on health outcomes and behavioral risk factors. This information can be used to measure and monitor the progress of the interventions.

The study designs among the twenty-one communities in year one (six more were added in year two) varied greatly. While some used a very simple list-assisted random-digit dial (RDD) sampling method, others used stratification, supplementation from a list sample, or in-person interviewing. Some communities involved over-sampling in order to target the rare race-ethnicity group of interest. These designs sometimes resulted in large probability and weight differentials, which in turn reduced the effective sample size of analyses. Five of the communities had enough variability in the weights to reduce the effective sample size by a factor of three.

Since the variability in the weights is so large, some trimming of them seems desirable. Trimming weights is assumed to cause bias, but will reduce variability. Of course, like a lot of choices in statistics, this is a delicate balancing act. In this paper, we have actual data from year one of REACH, as well as the sampling weights calculated with no outlier weight adjustment. We adjust these sampling weights with many variations of simple outlier adjustments of two types: winsorization (capping all weights at the level of a certain percentile) and capping the ratios from the median weight. Using thirteen different variables, we then calculate the bias, variance, and mean-squared error under each variation. We then compare the mean-squared error for each adjustment method to find the best outlier adjustment for each community and overall, as well as search for general patterns among the communities.

This paper gives a very brief look at some previous (recent) work in outlier weight adjustment in Section 2. Sections 3 and 4 describe more details about the REACH community samples and weights. Section 5 then describes our approach and methodology. Results are then presented and discussed in Section 6, while Section 7 gives conclusions and speculations.

2. Previous (recent) work. Whether to use weights in statistical analyses is an issue that often arises in practice. The imperfect and applied nature of weighting is probably the reason that there is not a rich literature on how to prevent extreme weights from damaging statistical analyses. However, there are several recent papers that do study outlier weight adjustment.

Potter (1988) gives an overview of some of the early work done to control extreme sampling weights, including the trimming of weights and weight adjustments at each step of the weighting process. More recent work by Chantala (2001) and Roey et al. (2001) uses other simple techniques to optimize the mean-squared error of an estimator. Chantala compared winsorization at each percentile point between the 70^{th} and 99^{th} to conclude that the mean-squared error was minimized by setting the maximum weight to the 85^{th} percentile for the

National Longitudinal Study of Adolescent Health. Roey et al. (2001) used a method to constrain outlier weights by trimming the overall weight of any school in the 1998 High School Transcript Study that contributed more than a specified proportion of the estimated variance.

Other researchers have used more complex model-based outlier weight adjustment techniques. Deville and Sarndal (1992) use a logit-type model to use calibration to trim weights, Folsom and Singh (2000) use a generalized exponential model instead, while Elliott and Little (1999) build a random effects model framework.

This paper, however, will consider only the simplest approaches to understand and improve the REACH weights. We intend to further this research with more complex models such as those given above in the future.

3. REACH Community Samples. REACH consists of twenty-one separate samples. Most communities were phone surveys using list-assisted random-digit dial (RDD) sampling or some modification of list-assisted RDD sampling.

Many different factors were considered to determine the most appropriate sampling approach for a community. In order to qualify for telephone data collection, a community had to have phones in at least eighty percent of its households. Almost all communities satisfied this condition. Cultural considerations also played a role, in consultation with the local community organization. Three communities did use in-person interviewing. One of these three communities (North Carolina) used a list sample (provided by the Native American tribe itself) while the other two communities used traditional area probability methods to do the in-person interviewing. Making the point that cultural considerations were important, one of the two traditional area probability samples switched to telephone data collection for year two.

Among the telephone sampling communities, straightforward list-assisted RDD sampling was used where appropriate (six communities). However, in many communities, the targeted race-ethnicity group was sparse. Using RDD methods alone would have required a large amount of screening calls, so several modifications were made. In four communities, we were able to separate area codes and/or exchanges into higher and lower density strata. For seven communities, we purchased list samples to supplement the RDD sample (and reduce the number of RDD calls to make). These seven communities are referred to as "Dual Frame" communities. One community (Santa Clara, CA) used only a list sample because the screening (for Vietnamese) would have been too costly to use any RDD numbers.

For the stratified and dual frame communities, decisions had to be made regarding how many phone calls/completed interviews would come from the different strata or frames. Increasing the sample from the high-density strata or list frame would reduce the number of phone calls needed, and therefore the costs. However, increasing these sample sizes would also increase the differential selection probabilities and therefore, the variance of the weights. The amount of oversampling was determined for each community on a case-by-case basis. This resulted in some communities with small weight differentials and some with larger differentials. The twenty-one year one REACH communities are listed in Table 1 along with the sampling method used.

4. REACH Weights. REACH weights were calculated using a standard set of nine steps. The base weight is the inverse of the selection probability. We then make adjustments for the working residential number rate (telephone communities only), screener nonresponse, number of telephone lines (telephone communities only), household eligibility, within household selection, and person nonresponse. The eighth step is the outlier adjustment step. Finally, the ninth step was a scale adjustment so that the sum of the weights for each community is equal to the sample size for that community. Scale adjustment is done for the ease of use by the individual community groups, who are provided with the data for their community. Many standard statistical packages often treat weights as if that many observations have that value, which overstates the precision (i.e., the number of degrees of freedom) of the analysis (without scale adjustment).

We first calculated the weights without an outlier adjustment step in order to assess their variability. One measure of variability in the weights is the coefficient of variation (CV), which is the standard deviation of a quantity divided by the mean (which, in REACH, is one because of the scale adjustment step). A well-known property of weights (Kish, 1965) is that arbitrary weights increase the variance

of estimates by a factor 1+L where:

$$L = \frac{Var(W_i)}{\overline{W}^2} = [CV(W_i)]^2$$

1+L is commonly referred to as the Design Effect (DEFF) due to weighting. The effective sample sizes (n_{eff}) due to weighting can also be defined as:

$$n_{eff} = \frac{n}{1+L}$$

The above effective sample sizes only account for the variability in the weights, but the other main factors in reducing effective sample sizes (sample design and clustering issues) would remain the same under different outlier weight adjustment strategies, and thus, we ignore them here.

For REACH, five of the twenty-one communities have L > 2. This implies that the effective sample size is cut to a third of the observed sample size (e.g., a sample size of 900 interviews would have an effective sample size of only 300).

Table 1 shows the squared coefficient of variation (L) for each of the twenty-one REACH communities.

5. Approach and Methodology. With the actual year one REACH interview data, we were able to examine the bias caused and variance reduced by various outlier weight adjustment choices. We used two basic outlier weight adjustment strategies and computed twenty-four variations of the REACH weights. One of these variations was, of course, to use the full REACH weight with no outlier weight adjustment. We used twelve different variations of winsorization, which essentially caps a weight at a certain percentile (e.g,, for each community, set all weights larger than the 95th percentile to be equal to the 95th percentile value and/or set all weights smaller than the 5th percentile to be equal to the 5th percentile value). We used upper end caps (only) at the 99th, 95th, 90th, and 75th percentiles. We also used lower end caps at the 1st, 5th, 10th, and 25th percentiles. Finally, we used two-sided balanced winsorization at these four levels (1st and 99th; 5th and 95th; 10th and 90th; and 25th and 75th). We also used ten different variations of capping weight ratios from the median weight (e.g., for each community, set all weights more than five times the median weight to be equal to five times the median weight, and/or set all weights less than one-fifth of the median weight to be one-fifth of the median weight). We capped weights at the upper end at three, five, ten, and twenty times the median weight. We also capped weights at one-third, one-fifth, and one-tenth of the median weight (there were no communities with weights less than one-twentieth of the median). Finally, we enforced both ratios to be three, five, and ten. For all twenty-three of these outlier adjustment strategies, a scale adjustment was performed to make sure the weights were all comparable (all sum to the sample size). The twenty-fourth variation of the REACH weights tested was to use no weight at all (all cases have a weight equal to one). It is important to note that there were several communities in which many of the outlier weight adjustments had no effect because the variability among the weights was small.

We calculated means and standard errors for each community under each outlier weight adjustment strategy on thirteen important binary variables CDC has identified as "performance measures." The thirteen performance measures break down into two immunization questions ("Are you immunized for flu?" and "Are you immunized for pneumonia?"), three questions asked only of diabetics ("Have you had an HbA1C test/Feet checked/Eyes dilated within the last year?"), two questions on knowing warning signs ("Do you know the signs of myocardial infarction?" and "Do you know the signs of a stroke?"), two questions asked only of females at least fifty years old ("Have you had a mammogram in the last two years?" and "Have you had a Pap smear in the last three years?"), and four miscellaneous questions ("Are you a current smoker?"; "Are you eating five or more servings of fruits and vegetables per day?"; "Are you taking hypertension medication?"; and "Are you under a doctor's care for elevated cholesterol?").

Given the means and standard errors, we calculated the bias and variance for each of our twenty-four weights for each performance measure in each community. We made the assumption that the full weight with no outlier weight adjustment was unbiased. The bias for every other outlier weight adjustment was the difference between that weight's estimates and the estimates for the full weight. We then calculated the mean-squared error as the sum of the squared bias and the variance.

Given the mean-squared errors for each weight

on each variable in each community, we proceeded with two simple analyses. First, for each community, we summed the mean-squared errors across the thirteen performance measures. The weights could then be ranked for each community. For the second analysis, we took the mean-squared error sums from the first analysis and summed across the communities. This provided us with an overall ranking of the twenty-four weights.

6. Results. Table 1 shows which weight performed best for each community across the thirteen performance measures. The communities are ranked from the largest value of L (Lowell, MA) to the smallest (Texas). Thus, the full weights (with no outlier weight adjustment) had the most variability in Lowell, MA, and the least variability in Texas. The effective sample size (due to weighting) in Lowell, MA is less than one-fourth of the number of interviews completed. The sample type is given for each community, as well as the weight that had the lowest mean-squared error. For Lowell, MA, the best-performing weights were those that capped the highest weights to be ten times the median value, and the weight that capped both ratios at ten. These two weights were actually equivalent because the median is only 1.02 times the smallest weight (no capping was therefore done at the lower end). In Texas, because the weights had almost no variability (L=0.03), there was a 21-way tie for the best weight. The final column in Table 1 is the improvement in the mean-squared error of the best performing weight(s), as compared to the full weight with no outlier weight adjustment. For Lowell, MA, capping the highest weights to be only ten times the median weight improves the mean-squared error by 13.99%. This implies that in Lowell, MA, the effective sample size can be improved by 16.27% (.1399/.8601) simply by performing a particular outlier weight adjustment. In Texas, no such improvement is possible among the choices studied here.

The sorting on L in Table 1 allows us to generalize for three separate groups of communities. Communities with the largest values of L ($L > 1.5$) show the largest gains in mean-squared error. All of them show at least a double-digit gain in percentage terms. It is also noticeable that the best weight for each of these communities is one that caps the highest weights to be some multiple of the median.

For medium values of L ($0.3 < L < 1.0$), the gains can still be large, but not all are. Five of the ten such communities show a double-digit gain. For these communities, winsorization seems to be preferred to capping the weight ratios from the median weight. The best weight is one of the winsorization weights for eight of the ten communities.

Finally, the possible improvement for communities with small values of L ($L < 0.3$) is minimal. In fact, all five of these communities show that many of the variables are equivalent to no outlier weight adjustment at all, and none of the outlier weight adjustments can improve the mean-squared error.

It is important to note that the sample type is related to the values of L, but not consistently. For example, two of the three in-person interviewing communities have the lowest two values of L, but the third has the highest. Except for Alabama, the stratified RDD communities are among the highest in L. The exception of Alabama can be explained by the fact that Alabama did not oversample the higher-density cases as much as the others. The dual frame communities have generally higher values of L than the (list-assisted) RDD communities.

Table 2 shows which weights performed best across all twenty-one REACH communities. Those that show the smallest overall sum of mean-squared errors appear at the top of the table. Overall, it appears that capping the highest weights at a multiple of the median weight is the best strategy. Capping the highest weights at five times the median weight is marginally better than at ten, twenty, and three times the median weight. Capping the lowest weights to be at least some fraction of the median weight was totally ineffective, as is shown by the three low ratio weights occupying the three places right above the full weight with no outlier weight adjustment.

Across all communities, winsorization was not as effective as capping the highest rates at a multiple of the median weight. While capping the highest weights to be five times the median weight reduced the sum of the mean-squared errors by 10.29%, the best of the winsorizations only reduced the sum of the mean-squared errors by 5.24%. Again, the low outlier weights were much less important than the high outlier weights. In every case, the weight capping at the higher percentile (e.g., 95^{th}) finished much higher than the weight capping at the equivalent lower percentile (e.g., 5^{th}).

The top line of Table 2 indicates that if we chose, for each community, the top-performing outlier

weight adjustment in that community, the sum of the mean-squared errors could be reduced by 16.65%. This is not an elegant solution and is difficult to implement quickly because the weights need to be created while the questionnaire data is being cleaned for simultaneous delivery.

A key message of Table 2 is that weights are important. Analysts who ignore weights do so at their own risk. For REACH 2010, the bias caused by ignoring the weights outweighs the variance saved by not having any variability in the weights. Ignoring the weights entirely increased the sum of the mean-squared errors by 10.74%. This finding also points out a problem for complicated survey analyses done with software that does not accommodate weights. The bias in such cases could be severe.

A second important message of Table 2 is that any outlier weight adjustment is as good as or better than no outlier adjustment. Other than capping at the 1st percentile (which has a mean-squared error sum 0.07% worse than no outlier weight adjustment) and, of course not using any weight, the full REACH weight with no outlier weight adjustment was at the very bottom of the list.

A final message of Table 2 is that worrying about low outliers is not worth it. Every weight that only addresses low outliers was outperformed by every weight that only addresses high outliers. In addition, these weights addressing only low outliers showed very little improvement over the full REACH weight with no outlier weight adjustment.

7. Conclusions. One caveat to this paper is that only the simplest two techniques for outlier weight adjustment were considered: winsorization and the capping of weight ratios from the median weight. Nevertheless, capping the largest weights to be only five times the median weight performed the best, and improved the sum of the mean-squared errors by almost ten percent. This shrinks the design effect due to weighting by ten percent and adds eleven percent to the effective sample size. For analysts, this is like having an extra eleven percent of interviews.

We are just completing the year two weights for REACH 2010. Since the data is still being cleaned, we cannot yet repeat this analysis for year two data. However, we are eager to see if these results are replicated. For year two, we are using a default outlier weight adjustment of capping the highest weights to be five times the median weight. We have already had to make an exception for a community for which this adjustment would strongly reduce the stratification we built into the sample. In this case, we adopted a conservative approach in order to prevent significant bias. Another option being considered is to use the CV of the full weight with no outlier weight adjustment to trigger either the above adjustment or winsorization at the 95th percentile. We are also eager to consider more complex options in the future.

References:

Chantala, K. (2001). Constructing Weights to Use in Analyzing Pairs of Individuals from Add Health Data. www.cpc.unc.edu/projects/addhealth/files/pweights.pdf.

Deville, J. C. and Sarndal, C. E. (1992) Calibration Estimating in Survey Sampling. Journal of the American Statistical Association. Volume 87, pp. 376-382.

Elliott, M. R. and Little, R. J. A. (1999). Weight Trimming in a Random Effects Model Framework. Proceedings of the American Statistical Assocition Section on Survey Research Methods, pp. 365-370.

Folsom, R.E. and Singh, A. C. (2000). The General Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification. Proceedings of the American Statistical Assocition Section on Survey Research Methods, pp. 598-603.

Kish, L (1965) *Survey Sampling*. New York: John Wiley and Sons.

Potter, F. (1988). Survey of Procedures to Control Extreme Sampling Weights. Proceedings of the American Statistical Assocition Section on Survey Research Methods, pp. 453-458.

Roey, S., Caldwell, N., Rust, K., Blumstein, E., Krenzke, T., Legurn, S., Kuhn, J., Waksberg, M., Haynes, J., and Brown, J. (2001). The 1998 High School Transcript Study User's Guide and Technical Report. National Center for Health Statistics. NCES 2001-477, Washington, D. C.

Table 1. The twenty-one REACH year one communities, and the weight that worked best for each.

Community	Sample Type	$L = (CV)^2$	Best performing Weight(s)	MSE Improvement
Lowell, MA	Field - Area Prob	3.14	High ratio 10/Both ratios 10	13.99%
Atlanta	Dual Frame	2.60	High ratio 5/Both ratios 5	32.64%
San Diego	Stratified RDD	2.39	High ratio 10/Both ratios 10	30.95%
Seattle	Stratified RDD	2.25	High ratio 3	32.45%
Boston	Stratified RDD	2.00	High ratio 10/Both ratios 10	15.88%
Nashville	RDD	1.66	High ratio 5/Both ratios 5	19.02%
Los Angeles	Dual Frame	0.98	Cap at 90th percentile	14.69%
Chicago	Dual Frame	0.81	Cap at 75th percentile	16.13%
Chicago	Dual Frame	0.65	Winsorize 25-75	4.08%
New Orleans	Dual Frame	0.57	Winsorize 10-90	23.18%
Charleston	RDD	0.49	Cap at 5th/10th/25th/90th/95th	8.40%
Charlotte	Dual Frame	0.47	Winsorize 25-75	6.10%
Santa Clara	Phone List only	0.41	High ratio 3/Both ratios 3	7.14%
Los Angeles	Dual Frame	0.39	Cap 75th/Winsor 25-75	16.03%
Oklahoma	RDD	0.38	Low ratio 3	18.29%
Detroit	RDD	0.33	Cap at 25th	5.30%
Lawrence	RDD	0.29	11-way tie	0.00%
Alabama	Stratified RDD	0.29	9-way tie	0.00%
Bronx	RDD	0.22	13-way tie	0.00%
North Carolina	Field - List	0.04	17-way tie	0.00%
Texas	Field - Area Prob	0.03	21-way tie	0.00%

Table 2. A ranking of all twenty-four weights considered in this paper.

Weight	MSE sum	MSE Improvement
BEST BY COMMUNITY	0.028654	16.65%
Both ratios 5	0.030838	10.29%
High ratio 5	0.030985	9.87%
High ratio 10	0.031292	8.97%
High ratio 20	0.031292	8.97%
Both ratios 10	0.031292	8.97%
Both ratios 3	0.031708	7.76%
High ratio 3	0.031823	7.43%
Winsorize 5-95	0.032577	5.24%
Cap at 95th percentile	0.032600	5.17%
Winsorize 25-75	0.032677	4.95%
Cap at 75th percentile	0.033008	3.98%
Winsorize 10-90	0.033277	3.20%
Cap at 90th percentile	0.033362	2.95%
Cap at 99th percentile	0.033877	1.45%
Winsorize 1-99	0.033900	1.39%
Cap at 25th percentile	0.033938	1.28%
Cap at 10th percentile	0.034154	0.65%
Cap at 5th percentile	0.034169	0.60%
Low ratio 3	0.034169	0.60%
Low ratio 10	0.034377	0.00%
Low ratio 5	0.034377	0.00%
NO OUTLIER ADJ	0.034377	0.00%
Cap at 1st	0.034400	-0.07%
NO WEIGHT	0.038069	-10.74%