More Outlier Weight Issues in REACH 2010

Steven Pedlow, Yongyi Wang, Ellen Scheib, and Hee-Choon Shin
National Opinion Research Center, University of Chicago

Abstract

Weights are often used in sample surveys to create unbiased estimates. These weights typically adjust for differential selection probabilities, differential response rates, and variations from control totals among other things. These adjustments can create a lot of variability in the weights, which increases the variability of estimates (reducing the effective sample size). Sometimes, this variability is large, and is caused by extreme weights very different from most of the weights; we refer to these weights as outlier weights. Whether to trim outlier weights in statistical analyses is an issue that often arises in practice. This paper investigates the effects that trimming outliers has on the survey estimates and estimated variances for the REACH 2010 project.

This paper includes some earlier results from Pedlow et al (2003), but expands them. This earlier paper only analyzed year one data for REACH, and compared only simple methods (capping weight ratios and winsorization). This current paper includes analyses for years one and two and adds compound weight pooling to the methods analyzed. Beyond a look at which methods work best, this paper also considers the theoretical question of whether it is proper to curtail weight variability that is driven by sample design decisions.

Keywords: Extreme weights, winsorization, capping weights, simple weight pooling, compound weight pooling.

1. Introduction to REACH 2010

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.

Over forty communities have been given grants, and NORC has collected data from twenty-seven communities (only twenty-one in year one). NORC's interview collects 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-seven communities 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 stratification and over-sampling in order to target a rare race-ethnicity group of interest. These designs sometimes resulted in large probability and weight differentials, which reduce the effective sample size of analyses $\left(n_{eff}\right)$. The reduction in effective sample size is often referred to as the Design Effect (DEFF), which measures the inflation in standard errors when compared to a simple random sample with the same sample size:

$$n_{eff} = \frac{n}{DEFF}$$

Design effects can be approximately divided into two components: one based on weight differentials, and one due to clustering of cases. This talk concentrates only on the weight differential component.

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. 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 ($n_{eff,w}$) due to weighting can also be defined as:

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

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.

In year one, five of the twenty-one communities had enough variability in the weights to reduce the effective sample size by a factor of three. 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). These large design effects motivated us to explore trimming the weights. We were able to improve the sample designs in year two, which resulted in less variability in the weights. However, some of the weighting design effects were still significant in year two. Table 1 shows the squared coefficient of variation (L) for each of the twenty-seven REACH communities in years one and two.

2. Theory versus Practice in Outlier Weight Issues

Since the variability in the weights is so large, some trimming of them seems desirable. Trimming weights is assumed to cause bias if the extreme weight cases (that are trimmed) are different from the other cases, but will reduce the variability of the weights and therefore the variability of the survey estimates. Of course, like a lot of choices in statistics, this is a delicate balancing act. Our question is whether we can reduce the variance enough to outweigh any bias created. We use as our criterion the mean squared error (MSE), which is the bias squared plus the variance:

$$MSE = Bias^2 + Variance$$

Before we put into motion our comparison of weight-trimming strategies, we need to step back to consider the theoretical desirability of weight trimming. The mean squared error combines the bias and variance into a formula, but many argue that bias is much more important than variance. Any biased estimator is tainted, and some would argue that is not worth any variance savings. In REACH, the weight differentials arise by design. Our stratification strategy resulted in large differential probabilities. Theoretically, it is undesirable to tinker with weight differentials caused by our own sample design decisions.

3. Approach and Methodology

Keeping these theoretical issues in mind, we went ahead to study different weight trimming methods. With the actual year one and two 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 (winsorization and capping weight ratios from the median weight) and one more advanced method (compound weight pooling) to compute twenty-seven variations of the REACH weights. One of these variations was, of course, to use the full REACH weight with no outlier weight adjustment. Another was to not use the weights at all (all cases have a weight equal to one). We used twelve different variations of winsorization, which essentially caps the weights 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 (only) 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. The more advanced method we used is the compound weight pooling method given in Elliott and Little (2000). This method is described in the next two sections.

For all twenty-seven of these outlier adjustment strategies, a scale adjustment was performed to make sure the weights were all comparable (all sum to the sample size). 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 full 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." Here are the 13 performance measures:

- Immunization questions (2):
  - "Immunized for flu?"
  - "Immunized for pneumonia?"
- Diabetic within the last year questions (3):
  - "Have you had an HbA1C test?"
  - "Have you had your feet checked?"
  - "Have you had your eyes dilated?"
- Warning sign knowledge (2):
  - "Signs of myocardial infarction?"
  - "Signs of a stroke?"
- Females over 50 questions (2):
  - "Mammogram in last two years?"
  - "Pap smear in the last three years?"
- Hypertension (2):
  - "Taking hypertension medication?"
  - "Under doctor's care for it?"
- High Cholesterol (1):
  - "Under doctor's care for it?"
- Fruits and Vegetables (1):
  - "Five or more servings per day?"

Given the means and standard errors, we calculated the bias and variance for each of our twenty-seven weights for each performance measure in each community. We made the assumption that the full weight with no outlier weight adjustment was unbiased. We realize that this is an important assumption that gives an advantage to not trimming; in that sense, it is a conservative assumption. The bias for every other outlier weight adjustment was then the difference between that weight's estimate and the estimate 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 followed two simple steps to get a simple comparison of the twenty-seven different weights. First, for each community, we summed the mean-squared errors across the thirteen performance measures. The second step was then to take the mean-squared error sums by community and sum across the communities. This provided us with an overall score (sum of all mean-squared errors) for each of the twenty-seven weights.

4. Simple Weight Pooling

While winsorization and capping weight ratios are simple and well-known techniques, this is not the case for compound weight pooling. To understand compound weight pooling, we first explain simple weight pooling.

The first step in simple weight pooling is to determine a finite set of H cut points to divide the weights into cells. Since we had a finite number of weighting steps, we had a finite number of different weight values, so we used the distinct weight values as cut points. For any one cut point, we leave the smaller weights (those below the cutpoint) unchanged, while we pool all weights above the cutpoint and replace them with the average. This results in a new "weight" for each of the H cutpoints. Take a simple example with three cases, with weights of 1, 3, and 5. There are three cutpoints: 0, 2, and 4. Here are the resulting simple pooling weights:

Cutpoint 0: we pool all three cases;	3, 3, and 3.
Cutpoint 2: we pool 3 and 5;	1, 4, and 4.
Cutpoint 4: we do not pool;	1, 3, and 5.

Please note that one of the cutpoints results in "no weighting" (cutpoint 0) and one results in "no outlier adjustment" (cutpoint 4).

We can compute the mean squared error for each possible simple weight pooling method, and determine the "best" cutpoint for each community and overall. However, we found no pattern in the best cut-points by community (sometimes more pooling is better and sometimes less pooling is better), and simple weight pooling can be impractical because you need to analyze the actual data before calculating weights. Sometimes, this is impossible (weights are needed before data is ready) and it does seem dangerous to have the choice of what the final estimates will depend on a weighting decision that uses parameter estimates in the choice. These disadvantages can be overcome by the use of composite weight pooling.

5. Composite Weight Pooling

The idea behind composite weight pooling is to combine the simple weight pooling weights through a Bayesian method. Notationally, we let each cutpoint be represented by C = 1, …, c, … , H. Then, the Bayesian method is equivalent to assigning a probability to each of the H cutpoints c. There are, of course, many options for what probabilities to assign, but we considered three:

Option 1. Take the average of all H cut-point weights.

$$P(C=c)=\frac{1}{H}$$

Option 2. Give less weight to more pooling:

$$P(C=c)=\frac{c}{\sum_{i=1}^{H} c}$$

Option 3. Consider the distance between weights.

Option 1 is the very simplest idea, and Options 2 and 3 were attempts to improve upon the simplest combination method. Since we felt that less pooling might perform better, we explored Option 2 to give less weight to simple weight pooling methods that perform more pooling. Option 3 considers the difference between weight values. More weight is given to cutpoints that are very different from the next smaller weight.

Figure 1 shows an example of compound weight pooling. In this example, there are five different cases. Each of the five cases has a unique weight ranging from 0.2 to 3.0 in the Full Weight column. Since there are five distinct weight values, there are five different simple weight pooling cut-points, resulting in five different simple pooling weights. The first cut-point is below the smallest weight (0.2), so all five weights are averaged (W_1). The first cut-point is the unweighted option. The second cut-point is between the first (0.2) and second (0.3) weight values, so all but the lowest weight is averaged, and so on to create the five simple pooling weights. The fifth cut-point is just below the highest weight value, so no averaging is done for W_5. The last cut-point is the no outlier adjustment option. The different compound weight pooling options combine these five pooling weights (W_1 – W_5) differently. Option 1 simply averages the five pooling weights (each given a weight of 1/5). Option 2 gives more emphasis to W_5 because it involves no pooling; W_1 is given the least emphasis. Option 3 also gives more emphasis to W_5, but because it is so different from the other weights. The example shows that Option 1 trims the weights the most and Option 3 trims the weights the least.

6. Results

Table 2 shows which weights performed best across all twenty-one REACH communities in Year 1. The MSE improvement is the amount of MSE reduction compared to the weight with no outlier adjustment. Those that show the smallest overall sum of mean-squared errors appear at the top of the table. Among the compound weight pooling options, Option 1 performs the best while Option 3 performs the worst. However, Options 1 and 2 are the two best strategies used, beating all of the simpler methods. Both options reduce the mean squared error by over 12%. Among the simpler methods, the ratios from the median outperform winsorization. These ratios also outperform Compound Option 3. The full weight with no outlier weight adjustment is very close to the bottom of the table. This implies that trimming the weights creates very little bias while reducing variability a lot. However, looking at the very bottom of the table, the worst performer is the unweighted option, which implies that while trimming improves the mean squared error, ignoring the weights entirely (i.e., trimming all weights to be equal) causes an explosion of bias that outweighs the variance gain and increases the mean squared error by over 10%.

Table 3 shows the similar results for Year 2. Compound Options 1 and 2 are again the two best weights. The gains in mean squared errors are much smaller in Year 2; for the best two, the gains are 9% and 7% in Year 2 versus 16% and 13% in Year 1. The gains are smaller because sample design improvements were made that decreased the variability in selection probabilities and therefore the weights. Compound Option 3 is outperformed by fewer of the simpler methods in Year 2. Among the simpler methods, the ratios from the median are not as clearly superior to winsorization in Year 2. The lessons from the bottom of the distribution are the same. All of the trimming methods outperform the full weight, but ignoring the weights explodes the bias. The unweighted option increases the mean squared error by over 12.5%.

7. Summary

In summary, trimming resulted in more improvement during Year 1 than in Year 2 because the sample designs were improved for Year 2. The improved sample designs resulted in smaller variability among the weights, as shown in Table 1. The mean squared error improvements are significant. Compound Option 1 reduces the mean squared error by over 15% in Year 1, and by almost 9% in Year 2. These gains are just like creating a larger sample size by simply adjusting which weight is used in analyses. These gains are possible because very little bias is created in trimming compared to the significant reduction in variability.

Of all the methods tried, the simplest compound weight pooling option (Option 1) is best. Giving more emphasis to simple weight pooling cut-points with less averaging is outperformed by the simple average of all cut-points. More generally for the REACH data, all of the methods used outperform the full unadjusted weight.

One of our most important findings, though, is that you should not ignore the sampling weights. The bias explodes compared to the variance savings, which we can see is significant by the savings shown for all the various outlier weight adjustments. In both REACH years, ignoring the weights increases the mean squared error by over 10%. This has serious implications for complicated methods of analysis that must be performed with specialized software that cannot use the weights.

Among the simpler methods, capping ratios from the median weight perform better in Year 1 than winsorization. In Year 2, the best simple methods are still ratio caps, but the overall superiority is not as clear. Capping the high outliers is most important, but both tables show that capping low outliers has little effect on either the bias or variance.

8. Theory versus Practice Final Decision

So our empirical results indicate that significant improvement is possible; 15% in Year 1 and 9% in Year 2 using Compound Option 1. However, let’s step back a bit to consider whether it is theoretically justifiable to trim the weights to achieve these gains. Our analysis shows that the large gains seen are almost entirely achieved by reducing the weight differentials between strata. Recall that many of the REACH community sample designs used stratification to target high-density areas. It is the sampling decision of how much to oversample that caused large probability (and therefore weight) differentials. We re-ran our analysis to only adjust weights within strata and almost no gains could be had by outlier weight adjustments. No gains were possible without tampering with the sample design decisions. Since the gains have been shrinking (this paper only shows Year 1 and 2 data), NORC has decided not to adjust the weights, but we realize this is an open and critical question for further research.

References

Elliott, M. R. and Little, R. J. A. (2000). Model-Based Alternatives to Trimming Survey Weights. Journal of Official Statistics Vol. 16, pp. 191-209.

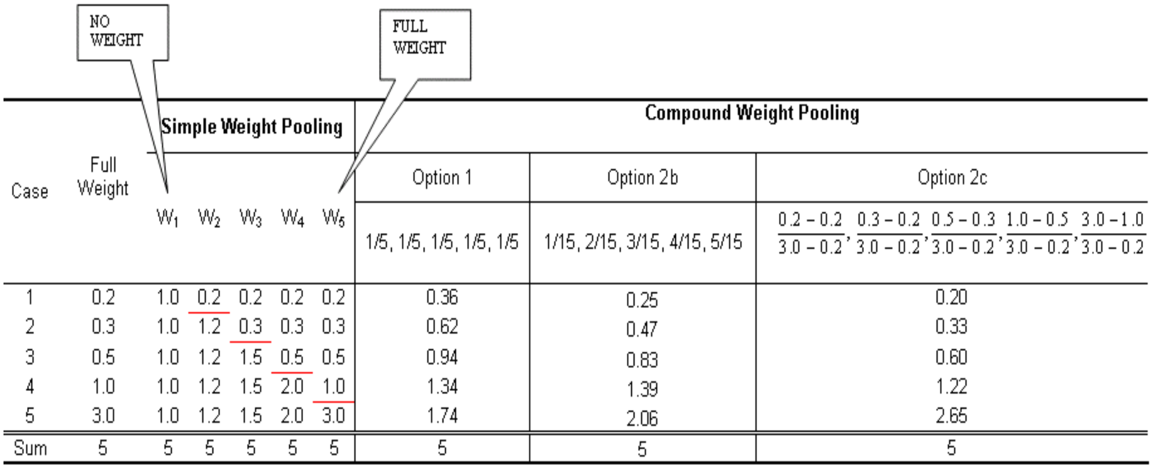
Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

Pedlow, S., Porras, J., O’Muircheartaigh, C., and Shin, H. (2003). Outlier Weight Adjustment in REACH 2010. 2003 Proceedings of the American Statistical Association, Survey Research Methods Section, Alexandria, VA: American Statistical Association: pp. 3228-3233.

Table 1. The twenty-seven REACH communities and the variability of the weights.

Community	Sample Type	L = (CV)^2	
		Year One	Year Two
Lowell, MA	Field - Area Prob	3.14	0.60
Atlanta	Dual Frame	2.60	0.05
San Diego	Stratified RDD	2.39	0.47
Seattle	Stratified RDD	2.25	0.86
Boston	Stratified RDD	2.00	0.49
Nashville	RDD	1.66	0.38
Los Angeles	Dual Frame	0.98	0.27
Chicago	Dual Frame	0.81	0.36
Chicago	Dual Frame	0.65	1.25
New Orleans	Dual Frame	0.57	0.40
Charleston	RDD	0.49	0.47
Charlotte	Dual Frame	0.47	0.73
Santa Clara	Phone List only	0.41	0.12
Los Angeles	Dual Frame	0.39	1.00
Oklahoma	RDD	0.38	0.69
Detroit	RDD	0.33	0.31
Lawrence	RDD	0.29	0.10
Alabama	Stratified RDD	0.29	0.88
Bronx	RDD	0.22	0.23
North Carolina	Field - List	0.04	0.02
Texas	Field - Area Prob	0.03	0.07
Portland	Stratified dual	n/a	0.70
Nevada	Dual Frame	n/a	0.64
New Mexico	Dual Frame	n/a	0.36
Missouri	Dual Frame	n/a	0.20
New Hampshire	Phone List Only	n/a	0.11
Chicago	Dual Frame	n/a	0.10

Figure 1. Example of Compound Weight Pooling.


NO WEIGHT
FULL WEIGHT

		Simple Weight Pooling					Compound Weight Pooling		
Case	Full Weight	W_1	W_2	W_3	W_4	W_5	Option 1	Option 2b	Option 2c
							1/5, 1/5, 1/5, 1/5, 1/5	1/15, 2/15, 3/15, 4/15, 5/15	$\frac{0.2-0.2}{3.0-0.2}, \frac{0.3-0.2}{3.0-0.2}, \frac{0.5-0.3}{3.0-0.2}, \frac{1.0-0.5}{3.0-0.2}, \frac{3.0-1.0}{3.0-0.2}$
1	0.2	1.0	0.2	0.2	0.2	0.2	0.36	0.25	0.20
2	0.3	1.0	1.2	0.3	0.3	0.3	0.62	0.47	0.33
3	0.5	1.0	1.2	1.5	0.5	0.5	0.94	0.83	0.60
4	1.0	1.0	1.2	1.5	2.0	1.0	1.34	1.39	1.22
5	3.0	1.0	1.2	1.5	2.0	3.0	1.74	2.06	2.65
Sum	5	5	5	5	5	5	5	5	5

Table 2. A ranking of all twenty-seven weights in Year 1

Weight	MSE sum	MSE Improvement
Compound Option 1	0.3762	15.83%
Compound Option 2	0.3906	12.59%
Both ratios 5	0.4009	10.29%
High ratio 5	0.4028	9.87%
High ratio 10	0.4068	8.97%
High ratio 20	0.4068	8.97%
Both ratios 10	0.4068	8.97%
Both ratios 3	0.4122	7.76%
High ratio 3	0.4137	7.43%
Compound Option 3	0.4147	7.21%
Winsorize 5-95	0.4235	5.24%
Cap at 95th percentile	0.4238	5.17%
Winsorize 25-75	0.4248	4.95%
Cap at 75th percentile	0.4291	3.98%
Winsorize 10-90	0.4326	3.20%
Cap at 90th percentile	0.4337	2.95%
Cap at 99th percentile	0.4404	1.45%
Winsorize 1-99	0.4407	1.39%
Cap at 25th percentile	0.4412	1.28%
Cap at 10th percentile	0.4440	0.65%
Cap at 5th percentile	0.4442	0.60%
Low ratio 3	0.4442	0.60%
Low ratio 10	0.4469	0.00%
Low ratio 5	0.4469	0.00%
NO OUTLIER ADJ	0.4469	0.00%
Cap at 1st	0.4472	-0.07%
NO WEIGHT	0.4949	-10.74%

Table 3. A ranking of all twenty-seven weights in Year 2

Weight	MSE sum	MSE Improvement
Compound Option 1	0.425848	8.93%
Compound Option 2	0.435966	6.77%
Both ratios 3	0.443181	5.22%
High ratio 3	0.444171	5.01%
Compound Option 3	0.447409	4.32%
Winsorize 25-75	0.454046	2.90%
Winsorize 10-90	0.456618	2.35%
Cap at 25th percentile	0.457647	2.13%
Winsorize 5-95	0.457716	2.12%
Cap at 75th percentile	0.459735	1.68%
Both ratios 5	0.460313	1.56%
Cap at 90th percentile	0.460315	1.56%
Cap at 95th percentile	0.460441	1.53%
High ratio 5	0.460481	1.52%
Cap at 10th percentile	0.463028	0.98%
Winsorize 1-99	0.464364	0.69%
Cap at 5th percentile	0.464426	0.68%
Cap at 99th percentile	0.464695	0.62%
Low ratio 3	0.466265	0.29%
Both ratios 10	0.466811	0.17%
High ratio 10	0.466814	0.17%
Cap at 1st percentile	0.467269	0.07%
Low ratio 5	0.467439	0.04%
Low ratio 10	0.467605	0.00%
High ratio 20	0.467608	0.00%
NO OUTLIER ADJ	0.467608	0.00%
NO WEIGHT	0.526201	-12.53%