THE WEIGHTING AND VARIANCES OF SURVEY DATA

Jai Won Choi. The National Center for Health Statistics,
6525 Belcrest Road, Hyattsville MD 20782.

1. INTRODUCTION

Sample units are often weighted a number of times in an attempt to estimate the original size of population and to make the sample compatible with the characteristics of population. For instance, the National Health Interview Survey data obtained by the National Center for Health Statistics, were weighted at least five times (w1-w5): The basic weight w1 is decided by the sampling design such as simple random , stratified, cluster, pps, equal or unequal, one or multistage stages, and with or without replacement design. W2 is the weight for non-response adjustment. The weight w3 is the 1st ratio adjustment for the residential areas in Non-Self- Representing (NSR) PSUs. The 2nd ratio adjustment w4 is the age-sex-race cell adjustment. W5 is the expansion of the two-week reference period to 13 weeks. Such weighting may be extended to more steps to reflect other features of population and to correct sample biases.

We want to have the final estimate as close to population as possible in every aspect. Despite our good efforts, often the estimates are not close to those of the population after such weighting. For instance, the cells of an age-sex-race table ha ve been changed after each weighting, although we wanted cell estimates close to those of the population, the final table is still quite different from that of the population as seen in Section 2.

The weighting may reduce the difference, variance and/or bias. It depends on each particular situation in sampling and weighting. The weighting may be repeated until the difference is minimized between the population and final cell distributions.

It is always necessary to have an accurate estimation of a variance to have proper inferences on data. It is rather tedious to find the variance of weighted data as the data is often weighted by the ratio estimation. As the ratio estimation is biased, sometimes we use the linear form of a ratio for the variance calculation. Others use a resampling method such as Balanced Half Samples (BHS). We may also use a correlation model defining the relationship between variables to find a variance.

In Section 2, the weights are estimated five times for a sample of 12 persons. The changes of cell distribution are shown after each weighting. The impacts of weighting are discussed in Section 3. The variances of the weighted data are discussed in Section 4.

2. EXAMPLE

A population is created for illustration, from which a sample of 12 persons is taken to show the five steps of weighting in Table 1. The population of 1,600 is divided into four strata as shown in the first column, the first stratum of 300, the second of 300, the third stratum of 600, and the fourth stratum of 400.

The first two are self-representing PSUs, while the last four are the NSR-PSUs as indicated in Column 2. No sampling is involved in the first two PSUs. Two PSUs are selected by equal probability out of the three PSUs, 300 each, in the stratum 3. Two PSUs are taken from the four PSUs, 100 each, in the stratum 4.

The weights given to these six PSUs are 1, 1, 3/2, 3/2, 4/2 and 4/2, respectively, in the first stage selection of PSUs.

A sample of 12 persons is selected from the six sample PSUs by simple random sample, 3 from each of the first two PSUs, 2 from each of the third and fourth PSUs, and 1 from each of the last two PSUs. The weights are 300/3, 300/3, 200/2, 200/2, and 100/1, 100/1 for the selection of a person from the respective sample PSU.

The column 3 in Table 1 shows the basic weights, the multiplications of the two weights arising from the selection of PSUs and persons.

Table 1: Weighting of Doctor Visits

	psu	Bas	non res	R1 w3	R2 w4	v 2w
1	A 300	100 100 100		1 100 1 100 1 100	0.8 80 1 100 1 100	2 160 0 0 0 0
2	B 300	100 100 100	3/2 No 3/2	1 150 No 1 150	1 150 No 1 150	1 150 No 0 0
3	A 200	150 150		1 150 1 150	1 150 1 150	0 0 3 450
	B 200	150 150	2/1 No	.9 270 No	.8 216 No	1 216 No
4	100	200		1 200	1 200	2 400
	100	200		1 200	1 200	0 0
K	1.2	1.6		1.57	1.496	9 1.376

The non-responses are adjusted within the PSU by the ratio, the sample number divided by the number of respondents within the PSU. Non-responses ratios of the 5th and 10th samples are shown in Column 4.

The living areas in NSR-PSU are divided into three cells of city, urban and rural places. Table 2 shows the six sample persons, 7, 8, 9, 10, 11 and 12 from the NSR-PSUs in the strata 3 and 4 in column 5 of Table 1. The first stage ratios R1, population to the sample estimates, are 1.0, 0.9, and 1.1 as seen in the last row of Table 2. Since the only ratio different from one is 0.9 in the second cell for the 9th sample, requiring the adjustment, while no adjustment is needed for the remaining 5 sample persons in the first cell

Table 2. 1st ratios in NSR-PSUs

	1 city	2 urban	3 rural
pop	510	90	100
est	510	100	90
1st ratio	1(7,8,10,11,12)	0.9(9)	1.1(0)

Each of the 12 persons belongs to one of the 8 age-sex-race cells. The Table 3 is consisted of the three subtables, each with eight cells of the 2 ages, 2 sex, and 2 races. The ratios in the last row are population divided by estimates, which also shown in Column 7 of Table 1. The figures in the parenthesis are the sample numbers.

Table 3. 2nd ratio between population and estimation

	age	MW	MB	FW	FB
pop	1-49	350	40	350	60
	50+	350	60	350	40
est	1-49	350	50	350	50
	50+	350	50	350	50
rate	1-49	1(5,8,10)	0.8(1)	1(2,3,7)	1.2(-)
	50+	1(4,11)	1.2(-)	1(6,12)	.8(9)

Column 7 in Table 1 shows the Doctor visits for the past two weeks (V), weighted visits for the 2 weeks recall (2W). The multiplication by 26 would give the weighted visits for 52 weeks.

Table 4 shows the weighting process of the 9th sample person. The weight of this person is changed five times, W1 through W5.

Table 4 The Changes of Weight

wl	w2	w3	w4	w5
150	300	270	216	5,616

W1 is the product of the two weights for the first and second stage selection. For instance, two PSUs are selected from three NSR-PSUs, and two persons are sampled from each sampled PSU of 200 in the third stratum. The basic weight is the product of these two weights, i.e., 150 = (3/2) x (200/2).

The W2 is the number adjusted for the non-response. Since one of the two sampled persons in the same PSU did not respond, the weight of the respondent is doubled (300 = 2 x 150) to cover the nonrespondent.

The W3 is 270 from the first ratio weighting (270 = 0.9 x 300). As this sample person lives in urban area, her first stage ratio is 0.9 as shown in Table 2.

The W4 is 216 by the second ratio weighting (216 = 270 x 0.8). Since this sample belongs to the cell (2,4), the black female of 50+ years, her second ratio is 0.8 for her age-sex-race. She represents 216 people for her stratum, PSU, residential area, and age-sex-race class.

The W5 is the estimated number of visits for 52 weeks or one year. She visited doctors' office once during the past two weeks, and her one visit became 5,616 visits (= 26 x 216) for 52 weeks as shown in the column seven.

Each sample person in Table 1 is weighted the same way. The nine visits from 12 sample persons would be 35,776 visits after the sample visits were weighted five times.

The numbers in the eight cells of the age-sex-race table are the basic weights (1), and they have been changed three times (2-4) through the three weighting processes as seen in Table 5.

Table 5. Four Weight changes of Eight cells

	1	2	3	4	5	6	7	8
p	350	40	350	60	350	60	350	40
1	400	100	350	0	250	0	300	150
2	300	100	350	0	300	0	350	150
3	270	100	350	0	300	0	350	150
4	216	80	350	0	150	0	350	150

After these weightings, the last row of w4 is still quite different from that of the population. This difference is mainly due to sampling, non-responses of the samples 5 and 10, and the first ratio adjustment of the sample nine, and the second ratio weighting of the samples 1 and 9.

Similarly the residential cells of population in the second row differ from the estimation in the last row of Table 6. This difference is also due to the sampling, empty cell, and first and second ratio adjustments of the ninth sample.

Table 6. Three cells of residential areas

cell	1 city	2 urban	3 rural
pop	300	90	100
est.	300	100	90
w1	300(8,10)	150(9)	0(-)
w2	300(8,10)	150(9)	0(-)
W3	270(8,10)	150(9)	0(-)
W4	216(8,10)	150(9)	0(-)

3. REMARKS IN WEIGHTING

In the process of ratio weighting, we observed that each step of weighting may reduce or increase the differences between the estimates and population. The weighting may also increase the relative bias and variance, depending on the specific situations in sampling and weighting, each step contributing to the estimation.

(1) The Basic Weight W1

There are many ways to select a sample. For instance, if population were structured in three stages, and the sample taken by pps design, the variance would be minimized. The basic weight of a unit is decided by the design to select it.

If a sample is randomly selected, the basic weighting may reduce the relative variance, while a non-random design might increase the variance and bias dramatically.

When the persons of a small sample are distributed over the cells in a large table, there may be some empty cells as seen in the previous example. No ratio estimation for cells is possible for those empty cells.

(2) Non-response Adjusted Weight W2

Non-response may be adjusted at a proper stage or stages. If non-responses arise randomly and the nonresponse rate is low, the ratio adjustment may be valid especially for a large sample, and bias an and/or variance reduced.

On the other hand, if the nonresponse rate is more than 30 percent, the ratio estimate may cause severe biases even for a large sample.

Alternative methods may be used in order to reduce bias in the presence of high rates of nonresponse. Other methods such as regression and Bayesian methods are often useful for

non-response estimation. But such methods usually bring problems later at the stage of data analysis.

(3) The First Stage Ratio Adjusted Weight W3

We often do not have enough sample persons in sparsely populated area or subpopulation such as black or older people. Consequently, a small sample may not be able to reflect the characters of population. Thus, we may use the ratio between a population and its estimation to have better representation in such areas or subpopulation.

(4) The Second Stage Ratio Adjusted Weight W4

The weights from the previous adjustment may not reflect the age-sex-race cells of the population. We may multiply the ratio, population to its estimate, to the previous weights. This is done for each person in the age-sex-race cell. But the resulting cells may differ from those of the population due to the empty cells, small sample, and the previous weighting. Although this process reduces the difference between the age-sex-race cells in the tables of population and estimate, it makes the difference greater between the cells of living areas.

(5) The Recall Adjusted Weight W5

The number of Dr. visits past two weeks is only 1/26 of one year, hence we multiply 26 to make the previous weights to be the visits for one year.

The resulting number of visits per year may mislead readers for a calendar year as two weeks could be extended to the future or past 52 weeks from the day of interview. In this case, the visits may be counted to a different year depending on the interview date.

If the nonresponse was already biased, the recalls adjust may further increase the bias.

(6) Comments

The ratio method does not create new estimates for empty cells in an age-sex-race table. Unless we use the estimates for empty cells, no improvement can be made. However, we may put one in an empty cell for estimation or we may increase sample size to avoid empty cells if it is possible.

The high rates of non-responses may cause larger biases especially when the units in a cell are different.

The ratios may be unstable for a small sample. Since a small sample may leave more empty cells, large biases may be introduced, and the problem of non-responses may be magnified.

The order of weighting also has influence on the final outcome of a table. If the order of weighting were changed in the previous example, the results would be quite different. For example, the age-sex-race was adjusted first and then residential area in NSR-PSUs next, the cell distribution of the tables would be very different. It is desirable that one may do the most important weighting at the last stage.

Above examples illustrate the difficulty to estimate population by ratio weighting to satisfy all of its aspects. In order to reduce the difference between the population and estimate in the age-sex-race as well as in residential areas, we may repeat steps from the first ratio W3 to the second ratio W4, leaving W1, W2, and W5 out after initial weighting, and stop when the difference between population and final estimates is minimum for both tables. Each time a new ratio table is created from the ratios between the population and new estimate of W3 or W4.

The ratio estimation may work better if no cells were empty, response rates high, sample size reasonably large, and cell members homogeneous.

4. VARIANCE

We used a few methods to derive the variances for weighted data: 1. Design based method, 2. Balanced Half Sample, 3. Survey data analysis (SUDAAN). The results are varied, depending on the variance method used, assumptions we make, weighting, sampling, manipulations required in BHS, model, handling of missing data, and others.

Table 7. Data for Variance Calculation

	N	n	wl	R wl	v vw4 wl(v-Rp)
1	A 300	3	100	0.8 80	2 160 100(1.16)
			100	1.0 100	0 0 100(-0.77)
			100	1.0 100	0 0 100(-0.77)
2	B 300	3	100	1.0 100	1 100 100(0.23)
			100	1.0 100	0 0 100(-0.77)
			100	1.0 100	0 0 100(-0.77)
3	A 200	2	150	1.0 150	0 0 150(-0.77)
			150	1.0 150	3 450 150(2.23)
	B 200	2	150	0.8 120	1 120 150(0.16)
			150	1.0 150	0 0 150(-0.77)
4	100	1	200	1.0 200	2 400 200(1.23)
	100	1	200	1.0 200	0 0 200(-0.77)
k	1.2	12	1.6	1.55	9 1.23

Table 7 is the revised form of Table 1 for easy calculation. We calculate variance without the first ratio estimation and nonresponse. But we have the second ratio estimation of the age-sex-race cells for the 12 sample persons, their doctor visits (v) for 2 weeks recall, and weighted visits (vw4) for 2 weeks recall in the last column..

The basic weights in the 4th column are the multiplication of two weights for selecting PSUs and persons, which are the same as seen in Table 1.

Denote E for expectation operator and V for the variance operator. The last row in Table 8 at the end of this paper shows the linear form for the fourth row, which are also shown in the last column of Table 7. The Using these two tables, we obtained the variances based on design-based method, SUDAAN, and BHS.

(1) Design Based Variance.

Define symbols as:

m for the average doctor visits per person,

L for the number of strata indexed by h (h=1,4)

N_h for the numbers of psu in the population in the hth stratum.

n_h for the number of sample psu indexed by i (i=1,nh)

M_{hi} for the number of subunits in the hi-th PSU indexed by j (j=1,Mhi)

m_{hi} for the subunits sampled from Mhi.

M_o for the sum of all subunits.

The upper case P_a for the population in the a-th age group, the lower case p_a for the sample estimate of it.

Y_{hija} for the number of doctor visits of hij-th person belonging to the a-th age group.

The number of doctor visits per person is given by

$$m_1 = \frac{1}{M_o} \Sigma_h \Sigma_{i=1,n_h} \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} \Sigma_{j=1,m_{hi}} \frac{y_{hija} P_a}{p_a}$$

where $y_{hi} = y_{hi+}/m_{hi}$. P_a and p_a remain the same for all persons in the ath age cell. When two PSUs are selected from each stratum (i.e. $n_h = 2$), the variance is given by

$$v(m_1) = \frac{1}{M_o^2} \Sigma_{h=1,L} [\ (\Pi_{h1} \Pi_{h2} \Pi_{h12}{}^{-1} - 1) \\ (\frac{M_{h1} y_{h1}}{\Pi_{h1}} - \frac{M_{h2} y_{h2}}{\Pi_{h2}}) \\ + \Sigma_{i=1,n_h} \frac{M_{hi}^2 (1 - f_{hi}) s_{hi}^2}{m_{hi}\ \Pi_{hi}}]$$

where s_{hi}^2 is the usual sum of squares divided by $(m_{hi} - 1)$. The pai is the selection probability of PSU, which can be defined differently. (Cochran, 1977).

When the parameter pai with subscript h1 is replaced by n_h/N_h and pai with subscript h12 by $n_h(n_h-1)/ N_h(N_h-1)$, above variance is given by

$$v(m_2) = \frac{1}{M_o^2}\ \Sigma_h\ [\ (1 - \frac{2}{N_h}) \frac{N_h^2}{4} \\ (M_{h1} y_{h1} - M_{h2} y_{h2})^2 \\ + \Sigma_{i=1,n_h} \frac{M_{hi}^2 (1 - f_{hi}) s_{hi}^2}{m_{hi}\ \Pi_{hi}}]$$

(2) SUDAAN

As seen previously, the sample data is weighted by the ratio of two variables. Thus, weighted estimates are biased. The SUDAAN is the method that we take the linear form of ratio to reduce such biases. We then apply the usual variance formula to the linearized variable.

In the previous example, Since the weighted visits y is adjusted by P_a/p_a, where P is known and p is a variable, y/p_a is approximated by linear expansion. When $P_a/E(p_a)$ is one, the mean m1 is written as

$$m_2 = \frac{1}{M_o} \Sigma_{h=1,L} \Sigma_{i=1,n_h} \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} \Sigma_{j=1,m_{hi}} (y_{hija} - R_{hij} p_a)$$

with the variance v(m2) below.

$$v(m_2) = \frac{1}{M_o^2} \Sigma_{h=1,L} \left[(1 - \frac{2}{N_h}) \frac{N_h^2}{4} (M_{h1} y_{1h1} - M_{h2} y_{1h2})^2 + \Sigma_{i=1,n_h} \frac{M_{hi}^2 (1-f_{hi}) s_{1hi}^2}{m_{hi} \pi_{hi}} \right]$$

Only difference between $v(m_1)$ and $v(m_2)$ is that y_{hi} and s_{hi}^2 are replaced by y_{lhi} and s_{lhi}^2, where the $y_{hija} P_a/p_a$ is replaced by its linear form $(y_{hija} - R_{hij} p_a)$.

(3) Balanced Half Sample

Each stratum was divided into the two parts, A and B. If it included two PSUs, one is randomly assigned to A and the other to B. Otherwise, we often make pseudo pairs. In this example, the first and the second are the pseudo-pair and the third stratum already included the two PSUs, each with 200 subunits, and the fourth stratum also included two PSUs, each 100 subunits.

The PSUs in the pair should be as similar as possible, so the two random groups in the pair would provide the random difference in the sample.

The replications are indexed by k (=1,4) and strata by L (=1,3). Denote the mean visits in the two PSUs of a stratum by m_{ikL} (i=A,B). Then the mean of the mean visits for the kth replication is $m_k = (m_{ik1} + m_{ik2} + m_{ik3})/3$, and the overall mean $m_o = (m_1 + m_2 + m_3 + m_4)/4$.

The variance is given by

$$v(m_o) = \frac{1}{4} \Sigma_{k=1,4} (m_k - m_o)^2$$
$$= \frac{1}{4} \frac{1}{9} \Sigma_{L=1,3} (m_{AL} - m_{BL})^2$$

Note that the two forms of an above variance are producing exactly the same numbers even when m_0 is a nonlinear estimate.

This result is illustrated with the data of 3 strata and 4 replicates in Table 9 below.

Table 9a Mean Visits per person for 2 weeks Recall

	strat 1		Strat 2		strat 3	
	A	B	A	B	A	B
Wtd Vists	160	100	450	120	400	0
Est pop	380	300	300	270	200	200
Wvisits/Est p	0.42	0.33	1.5	0.44	2	0

Table 9b Orthogonal Replications

repl	3 Strata with A or B PSU and its visits/est		
1	A 0.42	A 1.5	A 2.0
2	A 0.42	B 0.44	B 0
3	B 0.33	A 1.5	B 0
4	B 0.33	B 0.44	A 2.0

When one PSU of pair is empty either by nonresponse or by no sample, the variance may distort a true variance. Unless these defects are corrected by putting one or by estimation, biases would remain in the estimation. However, the actual measurements are zeroes, then they would not be any problem.

Using the last row of Table 9a and its orthogonal matrix, we made the replications Table 9b. From each stratum, we take one of A or B by orthogonal matrix as shown in Table 9b, and get four replications to obtain the mean and variance.

(4) Model Method

A model may be to define the correlation between members when variables are dependent. Let the positive correlation between sample units expressed by r (0 < r <1). It is known that the variance for correlated data is larger than that for independent variables by the factor of w [1 + r(m-1)], where m is the number of members in a cluster when the weight W of individuals is constant.

(5) Comments on variances

The variances of the means from the first three methods are shown in Table 10. The variance of design-based method is a little larger than that of the SUDAAN. It is because the variations are smaller in linear approximations than in the original variables. The subtractions by Rp reduce the size of absolute differences between v and v-Rp in the last column of Table 7.

It appears that, when the constants P/Ep and/or R in the linear form of variable is fluctuating, the variances are becoming larger. A linear approximation does not work well when the non-linear terms add up to a larger number.

When nonresponses are estimated by ratio in the presence of a high nonresponse rate, the resulting variance may not be reliable as the biases may remain in the variance.

Table 10 Variance comparison

var	Design Based	SUDAAN	BHS
Bet	0.0318	0.047	-
Within	0.129	0.0996	-
Total (mean)	0.1607 (0.79)	0.1463 (0.79)	0.1425 (0.78)

The variance of BHS depends on how to form pseudo-strata and how to pair the two pseudo-PSUs in a stratum. The difference of pairs should reflect the random difference of original random variables. Often the units in a pair have to be reweighted for each replication to reflect the overall mean or total.

However, the variances of the three different methods are remarkably close when the data are properly handled as seen in Table 10.

REFERENCES

Cochran, W. G. (1977). Sampling Technique, 3rd Edition. John Wiley and Sons. New York.

Kendall, M. G. and Stuart, A. (1968). The Advanced Theory of Statistics, Vol. 3, 2nd ed. Hafner Publishing Co., New York.

Choi, J. W. and McHugh, R. (1989). A Reduction Factor in Goodness of fit and Independence Tests for Clustered and Weighted Observations. Biometrics 45, September, pp 979-996.

Table 8. Linear form of ratio estimates for age-sex-race cells

cell (no)	1(3) 5,8,10	2(1) 1	3(3) 2,3,7	4(-)	5(2) 4,11	6(-)	7(2) 6,12	8(1) 9
r=P/p (r w1)	1(100,150 150)	0.8(80)	1(100, 100,150)	1.2(-)	1(100, 200)	1.2(-) -	1(100, 200)	0.8(120)
v cell p* Ev=9P* Es=12p*	0, 3, 0 0.219 2 2.6	2 0.025 0.2 0.3	0, 0, 0 0.219 2 2.6	- 0.038 0.3 0.5	1, 2 0.219 2 2.6	- 0.038 0.3 0.5	0, 0 0.219 2 2.6	1 0.025 0.2 0.3
r w1 v	0, 450, 0	160	0, 0, 0	-	100, 400	-	0, 0	120
Ev/Es=A	0.77	0.67	0.77	0.6	0.77	0.6	0.77	0.67
v - Rp R=A/Ep	-0.77 2.23 -0.77	1.16	-0.77 - 0.77 - 0.77	-	0.23 1.23	-	-0.77 -0.77	0.16

weighted visits = w1.(P/p).v = 150.(350/350).3 = 450 for the 5th sample

Linear expansion: w1{(P/Ep)[v - A/E(p).p]}; last cell= 150(40/40)[1 - (0.67/40) 50] = 150[0.16]