Proper Variance Estimation When Adjusting for Both Unknown Eligibility and Unit Nonresponse

Dhuly Chowdhury[1] Phillip S. Kott[1]
[1]RTI International; 6110 Executive Blvd; Rockville, MD 20852

Abstract
It is common practice to reweight for sampled elements with unknown eligibility within mutually exclusive weighting cells and then reweight among the eligible sampled elements in each cell for unit nonresponse. Even when the weighting cells are identical to the design strata under stratified simple random sampling, it is incorrect to treat the respondent sample as if it were a single-phase sample for variance estimation purposes. We show how variances should be estimated and then, using realistic data, compare the proposed approach to what is often done.

Key Words: Two-phase sample, completes, finite-population-correction factor

It is common practice to reweight for sampled elements with unknown eligibility within mutually exclusive weighting cells and then reweight among the eligible sampled elements in each cell for unit nonresponse. Even when the weighting cells are identical to the design strata under stratified simple random sampling, it is incorrect to treat the respondent sample as if it were a single-phase sample for variance estimation purposes as we shall see

Suppose we have a stratified simple random sample and treat the strata as the reweighting cells. To simplify the analysis, we concentrate on a single stratum which we treat as the population.

Let N be the population size (number of elements in the stratum/population)
- r be the number of respondents, defined here as sampled elements for which eligibility can be determined
- e be the number of sampled eligibles
- c be the number of completes among the eligibles
- $d = N/r$.
- $a = e/c$
- $w_k = da$ when element k is a complete, 0 otherwise. This is the final weight for k.

We assume that every element is equally likely to respond, and every eligible respondent is equally likely to be a complete. Under these assumptions are estimator for a population total for a variable y is

$$t = \sum_{k=1}^{c} d\, a y_k = \sum_{k=1}^{r} w_k y_k = de\bar{y},$$

where $\bar{y} = \frac{1}{c}\sum_{k=1}^{c} y_k$ (mean y-value among the completes).

The variance of the product of de, the estimated number of eligibles in the population, and $\bar{y}$, the estimated average y-value of an eligible, is

$$\text{Var}(de\bar{y}) \approx \bar{y}^2 Var(de) + (de)^2 \text{Var}(\bar{y}).$$

A good estimator for this variance is

$$v = \left(1-\frac{r}{N}\right) N^2 \bar{y}^2 \frac{p(1-p)}{r-1} + \left(1-\frac{c}{pN}\right)\left(\frac{c}{c-1}\right) d^2 a^2 \sum_{k=1}^{c} (y_k - \bar{y})^2$$

$$= \left(1-\frac{r}{N}\right) N^2 \bar{y}^2 \frac{p(1-p)}{r-1} + \left(1-\frac{c}{pN}\right)\left(\frac{c}{c-1}\right) d^2 a^2 \left(\sum_{k=1}^{c} y_k^2 - c\bar{y}^2\right),$$

where $p = e/r$ and $\bar{y} = \frac{1}{c}\sum_{k=1}^{c} y_k$. Note that $da = Np/c$.

One commonly used *ad-hoc* variance estimator for t ignore the first term of v:

$$v_2 = \left(1-\frac{c}{pN}\right)\left(\frac{c}{c-1}\right) d^2 a^2 \left(\sum_{k=1}^{c} y_k^2 - c\bar{y}^2\right).$$

This treats the c completes as the sample size and pN as the population size under simple random sampling without replacement (note that $pN = de$ estimates the population size of eligibles), so that $1 - c/(pN)$ is the finite-population-correction factor.

A more conservative *ad-hoc* variance estimator is

$$v_C = \left(1-\frac{c}{pN}\right)\left(\frac{r}{r-1}\right) \left[\sum_{k=1}^{c} w_k^2 y_k^2 - \frac{\left(\sum_{k=1}^{c} w_k y_k\right)^2}{r}\right]$$

$$= \left(1-\frac{c}{pN}\right)\left(\frac{r}{r-1}\right) d^2 a^2 \left(\sum_{k=1}^{c} y_k^2 - c^2 \frac{\bar{y}^2}{r}\right).$$

This estimator treats r as the sample size (and incompletes as 0s), but again treats pN as the population size, and $1 - c/(pN)$ as the finite-population-correction factor.

To see how the variance estimators can differ, we generate a respondent samples of 20 elements (k = 1, …, 20) from a χ_1^2 distribution. Let that value be y_k. Note that for our purposes, the sample size is the respondent sample size (r = 20).

For each k, we generate a ρ from a uniform [0, 1) distribution. We call a sampled respondent eligible when $y_k > b$ (note that eligibility is determined by the size of y_k). Call an eligible respondent complete when $\rho \geq q$.

We set b at 0, .1, .2, .3, .4;
q at 0, .1, .2, .3, .4; and
N at 50, 100, 150, 200.
100 (5 x 5 x 4) settings in all. Note that e and c depend on b and q.

In the following tables, we report e and c for each value of q and b, and we report the relative variance estimate using v (i.e., v/t^2), v_2, and v_C (i.e., v_C/t^2) for each value of e, c, and N.

Our proposed variance estimator, although always at least as large as the *ad-hoc* estimator v_2, returns a smaller value that the conservative v_C except when all responding eligibles are complete.

Table 1: Frame size (N) = 200 and Number of respondents with known eligibility (r) = 20

Total Eligible e	Total Complete c	Estimated Total t	First Variance Term	Second Variance Term v_2	Standard Error $\sqrt{v}$	Relative Bias of $\sqrt{v_2}$ $(\sqrt{v_2}/\sqrt{v} - 1) \times 100\%$	Relative Bias of $\sqrt{v_C}$ $(\sqrt{v_C}/\sqrt{v} - 1) \times 100\%$
20	20	214.0	0.0	4363.6	66.1	0.0	0.0
20	17	227.9	0.0	5757.6	75.9	0.0	3.3
20	14	203.9	0.0	5968.2	77.3	0.0	6.0
20	13	215.7	0.0	6847.8	82.8	0.0	7.3
20	10	196.4	0.0	10680.5	103.3	0.0	6.2
13	13	211.4	1140.1	3369.1	67.2	-13.6	-1.1
13	12	207.1	1094.4	3996.0	71.3	-11.4	1.3
13	10	182.2	847.1	4172.6	70.9	-8.8	5.3
13	9	198.9	1008.8	4937.5	77.1	-8.9	8.4
13	7	178.3	810.6	8285.3	95.4	-4.6	7.0
12	12	210.0	1392.5	3158.7	67.5	-16.7	-1.2
12	11	207.0	1353.5	3815.9	71.9	-14.1	1.5
12	9	185.0	1080.9	4127.8	72.2	-11.0	6.4
12	8	204.4	1319.0	4904.9	78.9	-11.2	10.5
12	6	189.1	1129.4	9121.3	101.2	-5.7	9.0
10	10	204.5	1981.9	2692.9	68.4	-24.1	-1.5
10	9	204.8	1987.0	3403.5	73.4	-20.5	2.3
10	7	190.4	1718.1	4022.6	75.8	-16.3	9.8
10	7	190.4	1718.1	4022.6	75.8	-16.3	9.8
10	5	183.2	1590.6	8564.1	100.8	-8.2	9.8
9	9	201.2	2343.8	2397.1	68.9	-28.9	-1.6
9	8	203.6	2400.2	3113.3	74.3	-24.9	3.0
9	6	195.0	2200.5	3853.0	77.8	-20.2	13.0
9	6	195.0	2200.5	3853.0	77.8	-20.2	13.0
9	4	198.6	2284.2	9819.2	110.0	-9.9	13.9

Table 2: Frame size (*N*) = 150 and Number of respondents with known eligibility (*r*) = 20

Total Eligible e	Total Complete c	Estimated Total t	First Variance Term	Second Variance Term v_2	Standard Error $\sqrt{v}$	Relative Bias of $\sqrt{v_2}$ $(\sqrt{v_2}/\sqrt{v}-1)\times 100\%$	Relative Bias of $\sqrt{v_C}$ $(\sqrt{v_C}/\sqrt{v}-1)\times 100\%$
20	20	160.5	0.0	2363.6	48.6	0.0	0.0
20	17	171.0	0.0	3138.4	56.0	0.0	3.3
20	14	152.9	0.0	3272.9	57.2	0.0	6.0
20	13	161.8	0.0	3762.6	61.3	0.0	7.3
20	10	147.3	0.0	5902.4	76.8	0.0	6.2
13	13	158.6	617.6	1824.9	49.4	-13.6	-1.1
13	12	155.4	592.8	2171.6	52.6	-11.4	1.3
13	10	136.7	458.9	2281.9	52.4	-8.8	5.4
13	9	149.2	546.4	2708.5	57.1	-8.8	8.5
13	7	133.7	439.1	4572.1	70.8	-4.5	7.1
12	12	157.5	754.3	1710.9	49.7	-16.7	-1.2
12	11	155.3	733.1	2074.2	53.0	-14.0	1.6
12	9	138.8	585.5	2259.1	53.3	-10.9	6.5
12	8	153.3	714.5	2693.3	58.4	-11.1	10.6
12	6	141.8	611.8	5040.7	75.2	-5.6	9.1
10	10	153.4	1073.5	1458.7	50.3	-24.1	-1.5
10	9	153.6	1076.3	1851.4	54.1	-20.5	2.4
10	7	142.8	930.6	2206.0	56.0	-16.1	10.0
10	7	142.8	930.6	2206.0	56.0	-16.1	10.0
10	5	137.4	861.6	4732.8	74.8	-8.0	10.0
9	9	150.9	1269.6	1298.5	50.7	-28.9	-1.6
9	8	152.7	1300.1	1694.3	54.7	-24.8	3.1
9	6	146.2	1191.9	2115.7	57.5	-20.0	13.3
9	6	146.2	1191.9	2115.7	57.5	-20.0	13.3
9	4	149.0	1237.3	5437.6	81.7	-9.7	14.1

Table 3: Frame size (*N*) = 100 and Number of respondents with known eligibility (*r*) = 20

Total Eligible e	Total Complete c	Estimated Total t	First Variance Term	Second Variance Term v_2	Standard Error $\sqrt{v}$	Relative Bias of $\sqrt{v_2}$ $(\sqrt{v_2}/\sqrt{v}-1)$ $\times$ 100%	Relative Bias of $\sqrt{v_C}$ $(\sqrt{v_C}/\sqrt{v}-1)$ $\times$ 100%
20	20	107.0	0.0	969.7	31.1	0.0	0.0
20	17	114.0	0.0	1305.7	36.1	0.0	3.3
20	14	101.9	0.0	1379.8	37.1	0.0	6.0
20	13	107.8	0.0	1592.9	39.9	0.0	7.3
20	10	98.2	0.0	2529.6	50.3	0.0	6.2
13	13	105.7	253.4	748.7	31.7	-13.6	-1.1
13	12	103.6	243.2	897.4	33.8	-11.3	1.4
13	10	91.1	188.3	956.2	33.8	-8.6	5.6
13	9	99.4	224.2	1142.6	37.0	-8.6	8.8
13	7	89.1	180.1	1953.4	46.2	-4.3	7.3
12	12	105.0	309.4	701.9	31.8	-16.7	-1.2
12	11	103.5	300.8	857.7	34.0	-14.0	1.7
12	9	92.5	240.2	948.3	34.5	-10.7	6.8
12	8	102.2	293.1	1138.6	37.8	-10.8	11.0
12	6	94.6	251.0	2160.3	49.1	-5.3	9.3
10	10	102.3	440.4	598.4	32.2	-24.1	-1.5
10	9	102.4	441.6	766.7	34.8	-20.3	2.5
10	7	95.2	381.8	930.0	36.2	-15.8	10.5
10	7	95.2	381.8	930.0	36.2	-15.8	10.5
10	5	91.6	353.5	2028.3	48.8	-7.7	10.4
9	9	100.6	520.8	532.7	32.5	-28.9	-1.6
9	8	101.8	533.4	702.4	35.2	-24.6	3.3
9	6	97.5	489.0	894.5	37.2	-19.6	13.9
9	6	97.5	489.0	894.5	37.2	-19.6	13.9
9	4	99.3	507.6	2340.6	53.4	-9.3	14.6

Table 4: Frame size (N) = 50 and Number of respondents with known eligibility (r) = 20

Total Eligible e	Total Complete c	Estimated Total t	First Variance Term	Second Variance Term v_2	Standard Error $\sqrt{v}$	Relative Bias of $\sqrt{v_2}$ $(\sqrt{v_2}/\sqrt{v} - 1) \times 100\%$	Relative Bias of $\sqrt{v_C}$ $(\sqrt{v_C}/\sqrt{v} - 1) \times 100\%$
20	20	53.5	0.0	181.8	13.5	0.0	0.0
20	17	57.0	0.0	259.6	16.1	0.0	3.3
20	14	51.0	0.0	288.8	17.0	0.0	6.0
20	13	53.9	0.0	338.7	18.4	0.0	7.3
20	10	49.1	0.0	562.1	23.7	0.0	6.2
13	13	52.9	47.5	140.4	13.7	-13.6	-1.1
13	12	51.8	45.6	173.6	14.8	-11.0	1.7
13	10	45.6	35.3	195.6	15.2	-8.0	6.3
13	9	49.7	42.0	239.7	16.8	-7.8	9.7
13	7	44.6	33.8	429.4	21.5	-3.7	8.0
12	12	52.5	58.0	131.6	13.8	-16.7	-1.2
12	11	51.8	56.4	166.3	14.9	-13.6	2.1
12	9	46.3	45.0	195.2	15.5	-9.9	7.8
12	8	51.1	55.0	240.9	17.2	-9.8	12.3
12	6	47.3	47.1	480.1	23.0	-4.6	10.2
10	10	51.1	82.6	112.2	14.0	-24.1	-1.5
10	9	51.2	82.8	149.6	15.2	-19.8	3.3
10	7	47.6	71.6	194.6	16.3	-14.5	12.2
10	7	47.6	71.6	194.6	16.3	-14.5	12.2
10	5	45.8	66.3	450.7	22.7	-6.6	11.7
9	9	50.3	97.7	99.9	14.1	-28.9	-1.6
9	8	50.9	100.0	137.6	15.4	-23.9	4.3
9	6	48.7	91.7	189.2	16.8	-17.9	16.3
9	6	48.7	91.7	189.2	16.8	-17.9	16.3
9	4	49.7	95.2	528.1	25.0	-8.0	16.4