Non-response Follow-up Allocation for Domains

Harold Mantel and Mike Hidiroglou
Statistics Canada, Tunney's Pasture, Ottawa, ON K1A 0T6

Abstract

In this paper we consider the problem of how to allocate a sub-sample of non-respondents from an initial sample for follow-up, when we want to estimate totals or means for domains. We consider different scenarios for what is known. At one extreme, we may know the domain of every unit in the initial sample, whether it responds or not. Alternatively, we may only know the domain of the respondent units, but have some auxiliary information such as the domain sizes. We assume that follow-up is intensive, so that all of the non-respondents in the follow-up sub-sample will become respondents, although this is unrealistic in practice. We then derive expressions for the variance which we use to find a follow-up allocation that satisfies CV criteria for domains of interest. Finally, we show how the assumption of complete response to the follow-up can be relaxed.

Key Words: non-response, follow-up, allocation, domains

1. Introduction

The purpose of this paper is to consider sampling with non-response and follow-up of a subsample of non-respondents. It is assumed that follow-up is intense so that all units in the subsample become respondents, however, that assumption may be relaxed if we assume that the propensity to respond at the follow-up phase is known or can be modeled and estimated.

The objective is to determine the allocation of the follow-up subsample to estimate totals for a given set of domains with a pre-specified level of precision, *i.e.* we assume that the population is partitioned into a set of domains of interest. We first find the variance of the estimator of a domain total under this setup, and determine the size of the follow-up sample needed and how it should be allocated.

For simplicity, we first consider, in Section 2, simple random sampling without replacement (SRSWOR) at both the initial and follow-up phases. In Section 3 we then extend this stratified SRSWOR at the initial phase. Section 4 briefly considers stratified two-stage sampling, and more general designs for the first phase are considered in Section 5. Finally we will relax the assumption of complete response to the follow-up phase in Section 6.

We can consider different scenarios with respect to our knowledge of the domains and the non-respondents. At one extreme we can assume we know the domain of all units in the sample, including non-respondents. Alternatively we may not know the domain of the non-respondents, but may have some auxiliary information such as the total size of the domain that can be incorporated into the estimation. At the other extreme we may not have any information about the domain sizes or their membership, and only be able to identify the domain of respondents.

2. Simple Random Sampling

Suppose we have a simple random sample s_1 of size n_1 from a population of size N. The population is partitioned into D domains of size N_d. We want to estimate the domain totals $Y_d = \sum_{i=1}^{N} y_{di}$ where $y_{di} = y_i$ if $i \in d$ and 0 otherwise.

Assuming no non-response we have $\hat{Y}_d = N \sum_{i \in s_1} y_{di} / n_1$ with variance given by $V\left(\hat{Y}_d\right) = N^2 \left(1/n_1 - 1/N\right) S_{dy}^2$, where $S_{dy}^2 = \sum_{i=1}^{N} \left(y_{di} - Y_d/N\right)^2 / \left(N-1\right)$.

For the remainder of this paper we will assume that there is some non-response, and sub-sampling of non-respondents for follow-up.

2.1 Domain known for entire sample

First suppose we know the domain d of all units in s. Suppose we select m_d of the $n_d - n_{rd}$ non-respondents for follow-up, and assume that we obtain responses from all of these second-phase units. So we have

$$s, s_d, n, n_d, s_r, s_{rd}, n_r, n_{rd}, m, m_d$$

Now assuming that the second phase units from domain d represent the non-respondents from d, we have

$$\hat{Y}_d = \frac{N_d}{n_d} \sum_{i \in s_{rd}} y_i + \frac{N_d}{n_d} \frac{n_d - n_{rd}}{m_d} \sum_{i \in s_{2rd}} y_i \tag{1}$$

$$\begin{aligned} Var\left(\hat{Y}_d\right) &= E\left\{Var\left(\hat{Y}_d \middle| n_d, n_{rd}, m_d\right)\right\} + Var\left\{E\left(\hat{Y}_d \middle| n_d, n_{rd}, m_d\right)\right\} \\ &= E\left\{Var\left(\hat{Y}_d \middle| n_d, n_{rd}, m_d\right)\right\} \end{aligned}$$

and

$$E\left\{Var\left[\left(\frac{N_d}{n_d} \sum_{i \in s_{rd}} y_i + \frac{N_d}{n_d} \frac{n_d - n_{rd}}{m_d} \sum_{i \in s_{2rd}} y_i\right) \middle| n_d, n_{rd}, m_d\right]\right\}$$

$$= E\left\{\begin{array}{l} E\left[Var\left\{\left(\frac{N_d}{n_d} \sum_{i \in s_{rd}} y_i + \frac{N_d}{n_d} \frac{n_d - n_{rd}}{m_d} \sum_{i \in s_{2rd}} y_i\right) \middle| s_d, s_{rd}, m_d\right\} \middle| n_d, n_{rd}, m_d\right] \\ +Var\left[E\left\{\left(\frac{N_d}{n_d} \sum_{i \in s_{rd}} y_i + \frac{N_d}{n_d} \frac{n_d - n_{rd}}{m_d} \sum_{i \in s_{2rd}} y_i\right) \middle| s_d, s_{rd}, m_d\right\} \middle| n_d, n_{rd}, m_d\right] \end{array}\right\}$$

$$= E\left\{ \begin{array}{l} E\left[\left(\frac{N_d^2}{n_d^2}\left(n_d - n_{rd}\right)^2 \left(\frac{1}{m_d} - \frac{1}{n_d - n_{rd}} \right) s_y^2 \left(s_d - s_{rd}\right) \right) \middle| n_d, n_{rd}, m_d \right] \\ + Var\left[\left(\frac{N_d}{n_d} \sum_{i \in s_{rd}} y_i + \frac{N_d}{n_d} \sum_{i \in s_d - s_{rd}} y_i \right) \middle| n_d, n_{rd}, m_d \right] \end{array} \right\}$$

$$= E\left\{ \frac{N_d^2}{n_d^2}\left(n_d - n_{rd}\right)^2 \left(\frac{1}{m_d} - \frac{1}{n_d - n_{rd}} \right) \sigma_y^2(d) + N_d^2 \left(\frac{1}{n_d} - \frac{1}{N_d} \right) \sigma_y^2(d) \right\}$$

$$= N_d^2 \sigma_y^2(d) \left\{ \left(\frac{1}{n_d} - \frac{1}{N_d} \right) + \frac{\left(n_d - n_{rd}\right)^2}{n_d^2} \left(\frac{1}{m_d} - \frac{1}{n_d - n_{rd}} \right) \right\} \qquad (2)$$

Here $\sigma_y^2(d)$ denotes the variance of *y* within domain *d*. Note that the first component in the braces of (2) is just the variance due to the first phase of the sampling (*i.e.* what it would have been if there were no non-response) and the second component is the additional variance due to sub-sampling of the non-respondents.

Now suppose we want to allocate the sub-sample to satisfy a maximum CV criterion, say $CV(\hat{Y}_d) \le K$. Then from (2) this is equivalent to

$$\frac{N_d^2 \sigma_y^2(d)}{Y_d^2} \left\{ \left(\frac{1}{n_d} - \frac{1}{N_d} \right) + \frac{\left(n_d - n_{rd}\right)^2}{n_d^2} \left(\frac{1}{n_d} - \frac{1}{n_d - n_{rd}} \right) \right\} \le K^2$$

or

$$\frac{\sigma_y^2(d)}{\overline{Y}_d^2} \left\{ \left(\frac{1}{n_d} - \frac{1}{N_d} \right) + \frac{\left(n_d - n_{rd}\right)^2}{n_d^2} \left(\frac{1}{m_d} - \frac{1}{n_d - n_{rd}} \right) \right\} \le K^2$$

or

$$m_d \ge \left\{ \frac{1}{n_d - n_{rd}} + \frac{n_d^2}{\left(n_d - n_{rd}\right)^2} \left(\frac{K^2}{cv_y^2(d)} - \frac{1}{n_d} + \frac{1}{N_d} \right) \right\}^{-1}$$

where $cv_y^2(d) = \sigma_y^2(d) \big/ \overline{Y}_d^2$. The minimum subsample size m_d that satisfies this constraint would depend on an assumed value for $cv_y^2(d)$. Note that we must also have $m_d \le n_d - n_{rd}$, so such an m_d may or may not exist. Presumably the first phase sample was designed to meet the CV criterion, at least in the case of no non-response, but we may have obtained a bad sample (small n_d) or the first phase data may suggest that $cv_y^2(d)$ is larger than anticipated.

2.2 Domain known only for respondents

Now suppose we know the domain *d* only for the responding units, but that we also know the domain population sizes N_d. We select m of the $n-n_r$ non-respondents for follow-up, and assume that we obtain responses from all of these second-phase units. So we have

$$s, n, s_r, s_{rd}, n_r, n_{rd}, m, m_d$$

Then, again assuming that the second phase units from domain *d* represent the non-respondents from *d*, we have

$$\hat{Y}_d = \frac{N_d}{n_{rd} + m_d \dfrac{n-n_r}{m}} \left(\sum_{i \in s_{rd}} y_i + \frac{n-n_r}{m} \sum_{i \in s_{2rd}} y_i \right) \tag{3}$$

$$Var\left(\hat{Y}_d\right) = E\left\{Var\left(\hat{Y}_d \middle| n_r, n_{rd}, m, m_d\right)\right\} + Var\left\{E\left(\hat{Y}_d \middle| n_r, n_{rd}, m, m_d\right)\right\}$$

$$= E\left\{Var\left(\hat{Y}_d \middle| n_r, n_{rd}, m, m_d\right)\right\}$$

And

$$E\left\{Var\left[\left(\sum_{i \in s_{rd}} y_i + \frac{n-n_r}{m} \sum_{i \in s_{2rd}} y_i\right) \middle| n_r, n_{rd}, m, m_d\right]\right\}$$

$$= E\left\{\begin{array}{l} E\left[Var\left\{\left(\sum_{i \in s_{rd}} y_i + \dfrac{n-n_r}{m} \sum_{i \in s_{2rd}} y_i\right) s_d, s_{rd}, m, m_d\right\} \middle| n_r, n_{rd}, m, m_d\right] \\ +Var\left[E\left\{\left(\sum_{i \in s_{rd}} y_i + \dfrac{n-n_r}{m} \sum_{i \in s_{2rd}} y_i\right) s_d, s_{rd}, m, m_d\right\} \middle| n_r, n_{rd}, m, m_d\right] \end{array}\right\}$$

$$= E\left\{\begin{array}{l} E\left[\left(\dfrac{\left(n-n_r\right)^2}{m^2} m_d^2 \left(\dfrac{1}{m_d} - \dfrac{1}{n_d - n_{rd}}\right) s_y^2 \left(s_d - s_{rd}\right)\right) \middle| n_r, n_{rd}, m, m_d\right] \\ + Var\left[\left(\sum_{i \in s_{rd}} y_i + \dfrac{n-n_r}{m} \dfrac{m_d}{n_d - n_{rd}} \sum_{i \in s_d - s_{rd}} y_i\right) \middle| n_r, n_{rd}, m, m_d\right] \end{array}\right\}$$

$$= E\left\{\begin{array}{l}\dfrac{\left(n-n_r\right)^2}{m^2} m_d^2\left(\dfrac{1}{m_d}-\dfrac{1}{n_d-n_{rd}}\right)\sigma_y^2(d) \\ + Var\left[\left(\sum_{i\in s_d} y_i + \left(\dfrac{n-n_r}{m}\dfrac{m_d}{n_d-n_{rd}}-1\right)\sum_{i\in s_d-s_{rd}} y_i\right)\middle| n_r, n_{rd}, m, m_d\right]\end{array}\right\}$$

$$= E\left\{\begin{array}{l}\dfrac{\left(n-n_r\right)^2}{m^2} m_d^2\left(\dfrac{1}{m_d}-\dfrac{1}{n_d-n_{rd}}\right)\sigma_y^2(d) \\ +n_d^2\left(\dfrac{1}{n_d-n_{rd}}\right)\sigma_y^2(d)+\left(\dfrac{n-n_r}{m}\dfrac{m_d}{n_d-n_{rd}}-1\right)^2\left(n_d-n_{rd}\right)^2\left(\dfrac{1}{n_d-n_{rd}}-\dfrac{1}{N_d}\right)\sigma_y^2(d) \\ +2\left(\dfrac{n-n_r}{m}\dfrac{m_d}{n_d-n_{rd}}-1\right) Cov\left(\sum_{i\in s_d} y_i, \sum_{i\in s_d-s_{rd}} y_i\middle| n_r, n_{rd}, m, m_d\right)\end{array}\right\}$$

Now

$$\begin{aligned}
&Cov\left(\sum_{i\in s_d} y_i, \sum_{i\in s_d-s_{rd}} y_i\middle| n_d, n_{rd}, m, m_d\right) \\
&= Cov\left(\sum_{i\in s_d} y_i, E\left[\sum_{i\in s_d-s_{rd}} y_i\middle| s_d, n_{rd}\right]\middle| n_d, n_{rd}, m, m_d\right) \\
&= Cov\left(\sum_{i\in s_d} y_i, \frac{n_d-n_{rd}}{n_d}\sum_{i\in s_d} y_i\middle| n_d, n_{rd}, m, m_d\right) \\
&= \frac{n_d-n_{rd}}{n_d} Var\left(\sum_{i\in s_d} y_i\middle| n_d, n_{rd}, m, m_d\right) \\
&= \frac{n_d-n_{rd}}{n_d} n_d^2\left(\frac{1}{n_d}-\frac{1}{N_d}\right)\sigma_y^2(d)
\end{aligned}$$

Putting the parts together we get

$$E\left\{Var\left(\hat{Y}_d \,\middle|\, n_d, n_{rd}, m, m_d\right)\right\} = \tag{4}$$

$$\frac{N_d^2}{\left(n_{rd} + m_d \dfrac{n-n_r}{m}\right)^2}\sigma_y^2(d)\left[\begin{array}{l} \dfrac{\left(n-n_r\right)^2}{m^2} m_d^2\left(\dfrac{1}{m_d} - \dfrac{1}{n_d - n_{rd}}\right) + n_d^2\left(\dfrac{1}{n_d} - \dfrac{1}{N_d}\right) \\ +\left(\dfrac{n-n_r}{m}\dfrac{m_d}{n_d - n_{rd}} - 1\right)^2 \left(n_d - n_{rd}\right)^2\left(\dfrac{1}{n_d - n_{rd}} - \dfrac{1}{N_d}\right) \\ +2\left(\dfrac{n-n_r}{m}\dfrac{m_d}{n_d - n_{rd}} - 1\right) n_d\left(n_d - n_{rd}\right)\left(\dfrac{1}{n_d} - \dfrac{1}{N_d}\right) \end{array}\right]$$

Note that in the special case $m_d = \dfrac{n_d - n_{rd}}{n - n_r} m = E\left(m_d \,\middle|\, n_d, n_r, n_{rd}, m\right)$ this simplifies to

$$E\left\{Var\left(\hat{Y}_d \,\middle|\, n_d, n_{rd}, m, m_d\right)\right\} = N_d^2\sigma_y^2(d)\left(\frac{\left(n_d - n_{rd}\right)^2}{n_{rd}^2}\left(\frac{1}{m_d} - \frac{1}{n_d - n_{rd}}\right) + \left(\frac{1}{n_d} - \frac{1}{N_d}\right)\right) \tag{5}$$

Here it is clear that the second component is just the variance due to the first phase of the sampling (*i.e.* what it would have been if there were no non-response) and the first component is the additional variance due to sub-sampling of the non-respondents.

Now suppose we want to satisfy a maximum CV criterion, say $CV(\hat{Y}_d) \le K$. The follow-up sample size in domain *d*, m_d, cannot be directly controlled. If we replace m_d by its expected value, then satisfying the CV criterion is equivalent to

$$\frac{N_d^2\sigma_y^2(d)}{Y_d^2}\left(\frac{\left(n_d - n_{rd}\right)\left(n - n_r\right)}{n_{rd}^2}\left(\frac{1}{m} - \frac{1}{n - n_r}\right) + \left(\frac{1}{n_d} - \frac{1}{N_d}\right)\right) \le K^2$$

for all domains *d*, or

$$m \ge \left\{\frac{1}{n - n_r} + \frac{n_{rd}^2}{\left(n_d - n_{rd}\right)\left(n - n_r\right)}\left(\frac{K^2}{cv_y^2(d)} - \frac{1}{n_d} + \frac{1}{N_d}\right)\right\}^{-1} \tag{6}$$

for all domains *d*, where $cv_y^2(d) = \sigma_y^2(d) \big/ \bar{Y}_d^2$. Note that if $cv_y^2(d)$ is large then there may not be any follow-up sample size *m* that satisfies (6) since we must also have $m_d \le n_d - n_{rd}$.

3. Stratified SRS at the first phase

We now consider again the setup of Section 2.2, but with stratified sampling at the first phase. That is, we suppose we know the domain *d* only for the responding units, but that we also know the domain population sizes N_d.

If we know the domain sample sizes N_{hd} within each stratum *h*, then we can reproduce the development of Section 2.2 within each stratum, using a separate ratio estimator, and things proceed quite similarly except that we would find a minimum follow-up sample size m_h for each stratum *h*.

For the rest of this section we will consider the combined ratio estimator, *i.e.* $\hat{Y}_d = \hat{Y}_{d,strat}\, N_d \big/ \hat{N}_{d,strat}$, where $\hat{Y}_{d,strat} = \sum_h \frac{N_h}{n_h}\left\{\sum_{i\in s_{rhd}} y_i + \frac{n_h - n_{rh}}{m_h}\sum_{i\in s_{2rhd}} y_i\right\}$ and $\hat{N}_{d,strat}$ is defined similarly. For this estimator the derivation of the variance becomes more complicated, primarily because we cannot assume, analogous to what we did in (3), that $E\left(\hat{Y}\,\middle|\, n_{hr}, n_{hrd}, m_h, m_{hd}\right)$ is constant, so the second term in the variance does not become 0. Instead we consider the approximate linearization variance:

$$Var\left(\hat{Y}_d\right) = Var\left(\hat{Y}_{d,strat}\right) + \left(\frac{Y_d}{N_d}\right)^2 Var\left(\hat{N}_{d,strat}\right) - 2\frac{Y_d}{N_d} Cov\left(\hat{Y}_{d,strat}, \hat{N}_{d,strat}\right) .$$

This can also be written as $Var\left(\hat{X}_{d,strat}\right)$ where $x_{id} = y_i - Y_d / N_d$. Note that x_{id} is not observable, since Y_d is unknown, but it may be convenient to write it in this form for calculations. $Var\left(\hat{Y}_d\right)$ may then be written explicitly as

$$\begin{aligned} Var\left(\hat{Y}_d\right) &= Var\left(\hat{X}_{d,strat}\right) \\ &= E\left\{Var\left(\hat{X}_{d,strat}\,\middle|\, n_{rh}, n_{rhd}, m_h, m_{hd}\right)\right\} + Var\left\{E\left(\hat{X}_{d,strat}\,\middle|\, n_{rh}, n_{rhd}, m_h, m_{hd}\right)\right\} \end{aligned}$$

As mentioned above the second term here does not disappear. Instead we have

$$E\left(\hat{X}_{d,strat}\,\middle|\, n_{rh}, n_{rhd}, m_h, m_{hd}\right) = \sum_h \frac{N_h}{n_h}\left(n_{rhd} + \frac{n_h - n_{rh}}{m_h} m_{hd}\right)\bar{X}_{hd}$$

and this will vary as n_{rhd} and m_{hd} vary. Now since

$$\hat{X}_{d,strat} = \sum_h \frac{N_h}{n_h}\left\{\sum_{i\in s_{rhd}} x_{id} + \frac{n_h - n_{rh}}{m_h}\sum_{i\in s_{2rhd}} x_{id}\right\}$$

we have

$$Var\left(\hat{X}_{d,strat}\right) = E\left\{Var\left(\hat{X}_{d,strat}\,\middle|\, n_{hr}, n_{hrd}, m_h, m_{hd}\right)\right\} + Var\left\{E\left(\hat{X}_{d,strat}\,\middle|\, n_{hr}, n_{hrd}, m_h, m_{hd}\right)\right\} \tag{7}$$

For the first component in this expression we have

$$Var\left[\sum_h \frac{N_h}{n_h}\left\{\sum_{i\in s_{hrd}} x_{id} + \frac{n_h - n_{hr}}{m_h}\sum_{i\in s_{2hrd}} x_{id}\right\}\middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right]$$

$$= E\left[Var\left[\sum_h \frac{N_h}{n_h}\left\{\sum_{i\in s_{hrd}} x_{id} + \frac{n_h - n_{hr}}{m_h}\sum_{i\in s_{2hrd}} x_{id}\right\}\middle| s_{hd}, s_{hrd}, m_h, m_{hd}\right]\middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right]$$
$$+ Var\left[E\left[\sum_h \frac{N_h}{n_h}\left\{\sum_{i\in s_{hrd}} x_{id} + \frac{n_h - n_{hr}}{m_h}\sum_{i\in s_{2hrd}} x_{id}\right\}\middle| s_{hd}, s_{hrd}, m_h, m_{hd}\right]\middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right]$$

$$= E\left[\sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{hr}}{m_h}\right)^2 m_{hd}^2\left(\frac{1}{m_{hd}} - \frac{1}{n_{hd} - n_{hrd}}\right)s_{x_d}^2\left(s_{hd} - s_{hrd}\right)\middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right]$$
$$+ Var\left[\sum_h \frac{N_h}{n_h}\left\{\sum_{i\in s_{hrd}} x_{id} + \frac{n_h - n_{hr}}{m_h}\frac{m_{hd}}{n_{hd} - n_{hrd}}\sum_{i\in s_{hd} - s_{hrd}} x_{id}\right\}\middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right]$$

$$= \sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{hr}}{m_h}\right)^2 m_{hd}^2\left(\frac{1}{m_{hd}} - \frac{1}{n_{hd} - n_{hrd}}\right)\sigma_{x_{hd}}^2$$
$$+ Var\left[\sum_h \frac{N_h}{n_h}\left\{\sum_{i\in s_{hd}} x_{id} + \frac{n_h - n_{hr}}{m_h}\frac{m_{hd}}{n_{hd} - n_{hrd}}\sum_{i\in s_{hd} - s_{hrd}} x_{id}\right\}\middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right]$$

$$= \sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{hr}}{m_h}\right)^2 m_{hd}^2\left(\frac{1}{m_{hd}} - \frac{1}{n_{hd} - n_{hrd}}\right)\sigma_{x_{hd}}^2$$
$$+ \sum_h \frac{N_h^2}{n_h^2} n_{hd}^2\left(\frac{1}{n_{hd}} - \frac{1}{N_{hd}}\right)\sigma_{x_{hd}}^2$$
$$+ \sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{hr}}{m_h}\frac{m_{hd}}{n_{hd} - n_{hrd}} - 1\right)^2\left(n_{hd} - n_{hrd}\right)^2\left(\frac{1}{n_{hd} - n_{hrd}} - \frac{1}{N_{hd}}\right)\sigma_{x_{hd}}^2$$
$$+ 2\sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{hr}}{m_h}\frac{m_{hd}}{n_{hd} - n_{hrd}} - 1\right)Cov\left(\sum_{i\in s_{hd}} x_{id}, \sum_{i\in s_{hd} - s_{hrd}} x_{id}\middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right)$$

$$
\begin{aligned}
&= \sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{rh}}{m_h}\right)^2 m_{hd}^2 \left(\frac{1}{m_{hd}} - \frac{1}{n_{hd} - n_{hrd}}\right)\sigma_{x_{hd}}^2 \\
&\quad + \sum_h \frac{N_h^2}{n_h^2} n_{hd}^2 \left(\frac{1}{n_{hd}} - \frac{1}{N_{hd}}\right)\sigma_{x_{hd}}^2 \\
&\quad + \sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{rh}}{m_h}\frac{m_{hd}}{n_{hd} - n_{hrd}} - 1\right)^2 \left(n_{hd} - n_{hrd}\right)^2 \left(\frac{1}{n_{hd} - n_{hrd}} - \frac{1}{N_{hd}}\right)\sigma_{x_{hd}}^2 \qquad (8) \\
&\quad + 2\sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{rh}}{m_h}\frac{m_{hd}}{n_{hd} - n_{hrd}} - 1\right)\frac{n_{hd} - n_{hrd}}{n_{hd}} n_{hd}^2 \left(\frac{1}{n_{hd}} - \frac{1}{N_{hd}}\right)\sigma_{x_{hd}}^2
\end{aligned}
$$

The second component of this sum can be interpreted approximately as what the variance would be with no non-response at the first phase. The first component represents variance due to sub-sampling of non-respondents. The third and fourth components are due to variability in the realized domain sub-sample sizes, m_{hd}, at the second phase. The fourth term has expected value zero, and the expected value of the third term can be calculated as

$$
\begin{aligned}
&\sum_h \frac{N_h^2}{n_h^2} Var\left(\frac{m_{hd}}{m_h}\right)\left(n_h - n_{hr}\right)^2 \left(\frac{1}{n_{hd} - n_{hrd}} - \frac{1}{N_{hd}}\right)\sigma_{x_{hd}}^2 \\
&= \sum_h \frac{N_h^2}{n_h^2}\left(\frac{1}{m_h} - \frac{1}{n_h - n_{hr}}\right)\frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\left(1 - \frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\right)\left(n_h - n_{hr}\right)^2 \left(\frac{1}{n_{hd} - n_{hrd}} - \frac{1}{N_{hd}}\right)\sigma_{x_{hd}}^2
\end{aligned}
$$

For the second component of (7) we have

$$
\begin{aligned}
&Var\left\{E\left(\hat{X}_{d,strat} \middle| n_{hr}, n_{hrd}, m_h, m_{hd}\right)\right\} = Var\left\{\sum_h \frac{N_h}{n_h}\left(n_{hrd} + \frac{n_h - n_{hr}}{m_h} m_{hd}\right)\bar{X}_{hd}\right\} \\
&= \sum_h \frac{N_h^2}{n_h^2}\bar{X}_{hd}^2 Var\left\{n_{hrd} + \frac{n_h - n_{hr}}{m_h} m_{hd}\right\} \\
&= \sum_h \frac{N_h^2}{n_h^2}\bar{X}_{hd}^2 E\left[Var\left\{n_{hrd} + \frac{n_h - n_{hr}}{m_h} m_{hd} \middle| n_{hr}, n_{hrd}, m_h\right\}\right] \\
&= \sum_h \frac{N_h^2}{n_h^2}\bar{X}_{hd}^2 E\left[\left(n_h - n_{hr}\right)^2 \left(\frac{1}{m_h} - \frac{1}{n_h - n_{hr}}\right)\frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\left(1 - \frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\right)\right]
\end{aligned}
$$

For allocation at the second phase we can substitute the conditional expected values of m_{hd}, $E\left[m_{hd} \middle| n_{hr}, n_{hrd}, m_h\right] = m_h \frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}$ in the first two components of (8), take the expected value of the last two, and work with the simplified expression

$$Var\left(\hat{X}_{d,strat}\right) = E\left\{Var\left(\hat{X}_{d,strat}\,|n_{hr}, n_{hrd}, m_h, m_{hd}\right)\right\} + Var\left\{E\left(\hat{X}_{d,strat}\,|n_{hr}, n_{hrd}, m_h, m_{hd}\right)\right\}$$

$$\doteq \sum_h \frac{N_h^2}{n_h^2}\left(\frac{n_h - n_{hr}}{m_h}\right)^2 \frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\left(\frac{1}{m_h} - \frac{1}{n_h - n_{hr}}\right)\sigma^2_{x_{hd}}$$

$$+ \sum_h \frac{N_h^2}{n_h^2} n_{hd}^2 \left(\frac{1}{n_{hd}} - \frac{1}{N_{hd}}\right)\sigma^2_{x_{hd}}$$

$$+ \sum_h \frac{N_h^2}{n_h^2} \bar{X}^2_{hd}\left(n_h - n_{hr}\right)^2 \left(\frac{1}{m_h} - \frac{1}{n_h - n_{hr}}\right)\frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\left(1 - \frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\right)$$

$$+ \sum_h \frac{N_h^2}{n_h^2}\left(\frac{1}{m_h} - \frac{1}{n_h - n_{hr}}\right)\frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\left(1 - \frac{n_{hd} - n_{hrd}}{n_h - n_{hr}}\right)\left(n_h - n_{hr}\right)^2 \left(\frac{1}{n_{hd} - n_{hrd}} - \frac{1}{N_{hd}}\right)\sigma^2_{x_{hd}}$$

For given values of $cv^2_{hx}(d) = \sigma^2_{x_{hd}} \big/ \bar{X}^2_{hd}$, we can search for a set of m_h that simultaneously satisfies the CV criterion for all domains of interest and minimizes the overall follow-up sample size.

4. Stratified two-stage sampling design at the first phase

Suppose that at the first phase we have a stratified two-stage design in which PSUs are selected within strata at the first stage and a SRSWOR of units is then selected from each of the sampled PSUs. Let s_{1h} denote the first-stage sample of PSUs in stratum *h* and s_1 the combined sample of PSUs from all strata. A SRSWOR s_{1j} of size n_j is drawn from PSU *j* at the first phase, n_{rj} of these respond at the first phase, and a subsample of m_j non-respondents is selected for follow-up. Note that the follow-up sample sizes m_j are now determined within PSUs.

The full sample estimator has the form $\hat{Y}_d = \hat{Y}_{d,strat} N_d \big/ \hat{N}_{d,strat}$, where $\hat{Y}_{d,strat} = \sum_h \sum_{j \in s_{1h}} \frac{1}{\pi_j}\frac{N_j}{n_j}\left\{\sum_{i \in s_{rjd}} y_i + \frac{n_j - n_{rj}}{m_j}\sum_{i \in s_{2rjd}} y_i\right\}$ and $\hat{N}_{d,strat}$ is defined similarly. Here the subscript *j* denotes the PSU, N_j and n_j are the population and sample size for PSU *j*, *etc*.

As in Section 3 we use the linearization approximation to the variance of $\hat{Y}_d$, *i.e.* $Var\left(\hat{Y}_d\right) = Var\left(\hat{X}_{d,strat}\right)$ where $x_{id} = y_i - Y_d / N_d = y_i - \bar{Y}_d$ and

$$\hat{X}_{d,strat} = \sum_h \sum_{j \in s_{1h}} \frac{1}{\pi_j}\frac{N_j}{n_j}\left\{\sum_{i \in s_{rjd}} x_{id} + \frac{n_j - n_{rj}}{m_j}\sum_{i \in s_{2rjd}} x_{id}\right\}$$

Now approximate expression for the variance can be derived as before. We skip the details of the derivation for space considerations (available on request from the first author).

After some simplifying assumptions as before, we have that $Var\left(\hat{X}_{d,strat}\right)$ can be estimated by

$$
\begin{aligned}
& Var\left(\hat{X}_{d,strat}\right) \\
\doteq\ & \sum_h \sum_{j\in s_{1h}} \frac{1}{\pi_j^2}\frac{N_j^2}{n_j^2} n_{jd}^2 \left(\frac{1}{n_{jd}} - \frac{1}{N_{jd}}\right)\sigma_{x_d}^2(h) \\
& + \sum_h \sum_{j\in s_{1h}} \frac{1}{\pi_j^2}\frac{N_j^2}{n_j^2}\left[n_{rjd}^2\left(\frac{1}{n_{rjd}} - \frac{1}{n_{jd}}\right) + \left(n_{jd} - n_{rjd}\right)^2 \left(\frac{1}{m_j}\frac{n_j - n_{rj}}{n_{jd} - n_{rjd}} - \frac{1}{n_{jd}}\right)\right]\sigma_{x_d}^2(h) \\
& - 2\sum_h \sum_{j\in s_{1h}} \frac{1}{\pi_j^2}\frac{N_j^2}{n_j^2}\left[n_{rjd}^2\left(\frac{1}{n_{rjd}} - \frac{1}{n_{jd}}\right)\right]\sigma_{x_d}^2(h) \\
& + Var\left\{ \sum_h \sum_{j\in s_{1h}} \frac{1}{\pi_j}\frac{N_j}{n_j} n_{jd}\bar{X}_{jd}\right\}
\end{aligned}
$$

Note that the last component of this sum, which corresponds to $Var\left\{E\left(\hat{X}_{d,strat}\middle| s_1, n_{jd}, n_{rjd}, m_j, m_{jd}\right)\right\}$, depends only on the first phase sampling. Nevertheless, if we want to allocate follow-up sample in order to meet some CV criteria this component is important. It could be estimated, for example, by substituting

$$
\hat{\bar{X}}_{jd} = \frac{1}{n_{rjd}}\sum_{i\in s_{rjd}} y_i - \left(\sum_h \sum_{j\in s_{1h}} \frac{1}{\pi_j}\frac{N_j}{n_j}\sum_{i\in s_{rjd}} y_i\right)\Bigg/\left(\sum_h \sum_{j\in s_{1h}} \frac{1}{\pi_j}\frac{N_j}{n_j} n_{rjd}\right)
$$

for $\bar{X}_{jd}$ and then using a Rao-Wu bootstrap to estimate this component of the variance. We could then search for a set of PSU follow-up sample sizes, m_j, to satisfy the CV criterion.

5. General stratified sampling at the first phase

Suppose that for the first phase we have a general, unspecified sampling plan within strata. We sample n_h units from the N_h units in stratum *h*, with inclusion probabilities π_i for $i \in h$. Within stratum *h* we have n_{rh} respondents, and we subsample m_h of the $N_h - n_h$ non-respondents for follow-up using SRSWOR. Now the expansion estimator for Y_d is

$$
\hat{Y}_{d,\exp} = \sum_h \left\{ \sum_{i\in s_{hrd}} \frac{y_i}{\pi_i} + \frac{n_h - n_{hr}}{m_h}\sum_{i\in s_{2hrd}} \frac{y_i}{\pi_i}\right\} = \sum_h \hat{Y}_{hd,\exp}\,. \qquad (9)
$$

As before, we will consider the combined ratio estimator $\hat{Y}_d = \hat{Y}_{d,\exp} N_d \big/ \hat{N}_{d,\exp}$ with variance approximated by

$$Var\left(\hat{Y}_d\right) = Var\left(\hat{X}_{d,\exp}\right) = \sum_h Var\left\{\sum_{i\in s_{rhd}} \frac{x_{id}}{\pi_i} + \frac{n_h - n_{rh}}{m_h}\sum_{i\in s_{2rhd}} \frac{x_{id}}{\pi_i}\right\} = \sum_h Var\left\{\hat{X}_{hd,\exp}\right\}$$

where, as before, $x_{id} = y_i - Y_d / N_d$.

Suppose that each unit in the population either will or will not respond to the first phase of the survey, *i.e.* that response to first phase is a fixed characteristic of each unit. Now we can write

$$Var\left(\hat{X}_{d,\exp}\right) = \sum_h \left\{Var\left[E\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right)\right] + E\left[Var\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right)\right]\right\}. \qquad (10)$$

For the first component of this sum we have

$$Var\left[E\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right)\right] = Var\left(\sum_{i\in s_{hd}} \frac{x_{id}}{\pi_i}\right) = \sum_{i\in U_{hd}}\sum_{j\in U_{hd}} \Delta_{ij}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j}$$

where $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$. An unbiased estimator of this variance is then given by

$$\hat{Var}\left[E\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right)\right] = \sum_{i\in s_{rhd}\cup s_{2rhd}}\sum_{j\in s_{rhd}\cup s_{2rhd}} \frac{\Delta_{ij}}{\pi^*_{ij}}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j} \qquad (11)$$

where

$$\pi^*_{ij} = \begin{cases} \pi_{ij} & \text{if both } i \text{ and } j \text{ respond to the first phase} \\ \pi_{ij}\dfrac{m_h}{n_h - n_{rh}} & \text{if only one of } i \text{ or } j \text{ responds to the first phase} \\ \pi_{ij}\dfrac{m_h}{(n_h - n_{rh})}\dfrac{m_h - 1}{(n_h - n_{rh} - 1)} & \text{otherwise} \end{cases}$$

The second component of the variance (10), $Var\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right)$, is just the variance due to sub-sampling of the non-respondents in stratum *h*. This can be written explicitly as

$$Var\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right) = \left(\frac{n_h - n_{rh}}{m_h}\right)^2 \sum_{i\in s_{hd}-s_{rhd}}\sum_{j\in s_{hd}-s_{rhd}} \Delta_{2ij}\frac{x_{id}}{\pi_{2i}}\frac{x_{jd}}{\pi_{2j}}$$

where $\Delta_{2ij} = \pi_{2ij} - \pi_{2i}\pi_{2j}$, and

$$\pi_{2i} = \frac{m_h}{n_h - n_{rh}}, \quad \pi_{2ij} = \frac{m_h}{(n_h - n_{rh})}\frac{m_h - 1}{(n_h - n_{rh} - 1)}. \qquad (12)$$

An unbiased estimator of this variance is given by

$$\hat{Var}\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right) = \left(\frac{n_h - n_{rh}}{m_h}\right)^2 \sum_{i\in s_{2rhd}}\sum_{j\in s_{2rhd}} \frac{\Delta_{2ij}}{\pi_{2ij}}\frac{x_{id}}{\pi_{2i}}\frac{x_{jd}}{\pi_{2j}}. \qquad (13)$$

Now the problem is to choose stratum follow-up sample sizes to make the resultant variances satisfy given CV criteria. One problem is that neither of the estimators (11) or (13) are calculable before having the data from the follow-up sample. One way to get around this is to suppose that the distribution of values of $\frac{\Delta_{2ij}}{\pi_{2ij}}\frac{x_{id}}{\pi_{2i}}\frac{x_{jd}}{\pi_{2j}}$ for pairs of units that respond to the first phase is the same as that over the entire sample. The sum in (11) may then be written as

$$\hat{V}ar\left[E\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right)\right] = \sum_{i\in s_{rhd}\cup s_{2rhd}}\sum_{j\in s_{rhd}\cup s_{2rhd}}\frac{\Delta_{ij}}{\pi^{*}_{ij}}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j}$$

$$= \sum_{i\in s_{rhd}}\sum_{j\in s_{rhd}}\frac{\Delta_{ij}}{\pi_{ij}}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j}$$

$$+ 2\frac{n_h-n_{rh}}{m_h}\sum_{i\in s_{rhd}}\sum_{j\in s_{2rhd}}\frac{\Delta_{ij}}{\pi_{ij}}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j}$$

$$+ \frac{n_h-n_{rh}}{m_h}\frac{n_h-n_{rh}-1}{m_h-1}\sum_{i\in s_{2rhd}}\sum_{j\in s_{2rhd}}\frac{\Delta_{ij}}{\pi_{ij}}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j}$$

$$\doteq \sum_{i\in s_{rhd}}\sum_{j\in s_{rhd}}\frac{\Delta_{ij}}{\pi_{ij}}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j}\left(1+2\frac{n_h-n_{rh}}{m_h}\frac{m_{hd}}{n_{rhd}}+\frac{n_h-n_{rh}}{m_h}\frac{n_h-n_{rh}-1}{m_h-1}\frac{m_{hd}^2}{n_{rhd}^2}\right)$$

Replacing $\frac{m_{hd}}{n_{rhd}}$ in this sum by its expected value then gives

$$\hat{V}ar\left[E\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right)\right]$$

$$\doteq \sum_{i\in s_{rhd}}\sum_{j\in s_{rhd}}\frac{\Delta_{ij}}{\pi_{ij}}\frac{x_{id}}{\pi_i}\frac{x_{jd}}{\pi_j}\left(1+2\frac{n_h-n_{rh}}{n_{rh}}+\frac{n_h-n_{rh}}{n_{rh}}\frac{n_h-n_{rh}-1}{n_{rh}-1}\frac{m}{n_{rh}}\right)$$

Similarly the expression (13) might be estimated by

$$\hat{V}ar\left(\hat{X}_{hd,\exp}\middle|s_{hd}\right) = \left(\frac{n_h-n_{rh}}{m_h}\right)^2\sum_{i\in s_{2rhd}}\sum_{j\in s_{2rhd}}\frac{\Delta_{2ij}}{\pi_{2ij}}\frac{x_{id}}{\pi_{2i}}\frac{x_{jd}}{\pi_{2j}}$$

$$\doteq \frac{m_h^2}{n_{rh}^2}\left(\frac{n_h-n_{rh}}{m_h}\right)^2\sum_{i\in s_{rhd}}\sum_{j\in s_{rhd}}\frac{\Delta_{2ij}}{\pi_{2ij}}\frac{x_{id}}{\pi_{2i}}\frac{x_{jd}}{\pi_{2j}}$$

The problem is then to find follow-up sample sizes m_h so that the estimated variances based on the first-phase data satisfy the CV criteria.

6. Relaxing the assumption of full response to follow-up

Up to this point we have assumed that we would be able to obtain full response to the follow-up sub-sample. In practice this is unlikely to be the case, since resources for follow-up would be limited, and in any case some sample members may simply refuse to respond with no prospect of being converted to respondents. In such a situation the estimators discussed above would inevitably need to be adjusted for non-response.

One way to adjust the estimators for non-response is to form non-response adjustment classes, and then adjust the weights of respondents within adjustment classes to sum to the total weight of both respondents and non-respondents.

A second approach to non-response adjustment is to model the propensity to respond at the individual unit level, based on models using covariates available for both respondents and non-respondents, and to then adjust the individual weights of respondents by the inverse of the estimated propensities to respond. This adjustment should be followed by a final calibration of the weights to known population totals. Of course, it is also possible to base non-response adjustment classes on these model-based response propensities, and this is often the preferred approach since it leads to more stable weight adjustments.

In the present context, with a first phase of sampling and collection followed by a second phase of non-response follow-up, we can consider non-response adjustment for the first phase or the second phase. Let p_i denote the propensity for unit *i* to respond to the first phase and q_i the propensity for unit *i* to respond to the second phase. A statistical model relating p_i and q_i to covariates available for both the respondents and non-respondents would be used to obtain estimates of these propensities.

Taking the general stratified sampling of Section 5 as an example, we may consider alternatives to the expansion estimator (9). Using estimates $\hat{p}_i$ of p_i, and $\hat{q}_i$ of q_i, we have the estimator

$$\hat{Y}^*_{d,\exp} = \sum_h \left\{ \sum_{i \in s_{rhd}} \left(\alpha \frac{1}{\hat{p}_i} + 1 - \alpha \right) \frac{y_i}{\pi_i} + \frac{n_h - n_{rh}}{m_h} \sum_{i \in s_{2rhd}} (1-\alpha) \frac{1}{\hat{q}_i} \frac{y_i}{\pi_i} \right\} = \sum_h \hat{Y}^*_{hd,\exp} \ .$$

Probably some simulations would be needed to choose a suitable value of α. We could also consider domain-specific values for α, but in that case the domain estimators would not add up to the overall estimator, so it is probably better to have a single α value for all domains.

We still consider the combined ratio estimator. For example, if we take $\alpha = 0$ then our estimator would have variance approximated by

$$Var\left(\hat{Y}^*_{d,\exp}\right) = Var\left(\hat{X}^*_{d,\exp}\right) = \sum_h Var\left\{ \sum_{i \in s_{rhd}} \frac{x_{id}}{\pi_i} + \frac{n_h - n_{rh}}{m_h} \sum_{i \in s_{2rhd}} \frac{1}{\hat{q}_i} \frac{x_{id}}{\pi_i} \right\}$$

where, as before, $x_{id} = y_i - Y_d / N_d$. Now we simply note that

$$Var\left(\hat{X}^*_{d,\exp}\right) \doteq \sum_h Var\left\{\sum_{i\in s_{rhd}} \frac{x_{id}}{\pi_i} + \frac{n_h - n_{rh}}{m_h}\sum_{i\in s_{2hd}} \frac{x_{id}}{\pi_i}\right\}$$
$$+ \sum_h E\left[\left(\frac{n_h - n_{rh}}{m_h}\right)^2 \sum_{i\in s_{2hd}} \left(\frac{1}{\hat{q}_i}\frac{x_{id}}{\pi_i}\right)^2 \hat{q}_i\left(1-\hat{q}_i\right)\right]$$
$$= Var\left(\hat{X}_{d,\exp}\right) + \sum_h E\left[\left(\frac{n_h - n_{rh}}{m_h}\right)^2 \sum_{i\in s_{2rhd}} \left(\frac{1}{\hat{q}_i}\frac{x_{id}}{\pi_i}\right)^2 \hat{q}_i\left(1-\hat{q}_i\right)\right]$$

So the variance of the estimator with estimated response propensities is simply the variance of the estimator with full response to the non-response follow-up, plus an additional component due to the non-response at the second phase. Determination of follow-up allocation to satisfy CV constraints for domains is now done as before, but accounting for this extra variability. In general, $\hat{q}_i$ would not be available at the time of allocation, so some assumptions about q_i would need to be made.

Another key assumption would be that the propensity to respond does not depend on the domain. If it does then bias becomes a concern, though calibrating to the known domain sizes through the use of the combined ratio estimator.

7. Concluding Remarks

We have shown that it is possible to find analytically approximate sample allocations required for non-response follow-up to satisfy CV targets for domain totals, if such allocations exist. Such analytical solutions would depend on some necessary assumptions about the population and the non-response process. It is best to try to keep all required assumptions neutral, but to incorporate any knowledge that is available.

In some situations it may not be possible to achieve the CV targets for all domains if the domain sample size is too small. The effectiveness of the follow-up strategy would also be very much affected by what is known about the non-respondents to the first phase of the survey, in particular whether or not the domain of these non-respondents is known after the first phase. In many practical situations the domain of non-respondents would not be known, and the scope for targeting non-response follow-up would be quite limited.

Although in the main development of the paper we assumed that we could obtain 100% response to the non-response follow-up, we have also shown that allowing for some non-response to the follow-up is not too difficult. Essentially such non-response at the follow-up phase just adds another component to the variance that would need to be anticipated and allowed for in the follow-up allocation.

In our development we have assumed relatively simple estimators of totals – post-stratified or ratio type estimators. For more complex estimation procedures, in particular in the case that there are more calibration constraints, the problem may more complex, but should in principle be solvable using a linearization approach similar to that described in Section 3. This could be the focus of future work.