SSC Annual Meeting, July 1995
Proceedings of the Survey Methods Section

# ADJUSTMENT OF THE INCLUSION PROBABILITIES IN CASE OF NONRESPONSE

Y.P. Chaubey and A.N. Crisalli[1]

## ABSTRACT

We use the Horvitz-Thompson estimator $t_{HT} = \frac{1}{N}\sum_{k\epsilon S}\frac{y_k}{\pi_k}$ when only $r$ out of $n$ units $(r<n)$ respond. Five adjustments for the inclusion probabilities $\pi_k$ are considered. These adjusted probabilities reduce the nonresponse bias when nonresponse is not at random. In this paper we also compare these biases under a simple random sampling design. We also impose a superpopulation model on the response variable so as to find the expected value and variance of biases. The consequences of these adjustments are further studied by Monte Carlo simulations with artificial data sets.

## RÉSUMÉ

Nous utilisons l'estimateur de Horvitz-Thompson $t_{HT} = \frac{1}{N}\sum_{k\epsilon S}\frac{y_k}{\pi_k}$ lorsque $r$ unités seulement répondent parmi $n\,(r<n)$. Cinq ajustements des probabilités d'inclusions $\pi_k$ sont considérés. Ces probabilités ajustées réduisent le biais de non-réponses lorsque la non-réponse n'est pas aléatoire. Dans cette communication, nous comparons les biais sous un plan d'échantillonnage aléatoire simple. Nous imposons un modèle de superpopulation sur la variable de réponse afin de trouver la valeur attendue et la variance des biais. Les conséquences de ces ajustements sont ensuite étudiées en utilisant des jeux de données artificielles obtenus par simulations de Monte Carlo.

## 1. INTRODUCTION

The problem of *nonresponse* is a common one in surveys. The missing data due to nonresponse may be broadly classified into two main categories - (1) unit nonresponse and (2) item nonresponse. Unit nonresponse occurs when a sampled unit provides no information with respect to the survey variables under study while item nonresponse occurs when a sampled unit provides partial information to the survey variables under study. Two general techniques have been proposed in the literature to alleviate this problem: (1) Imputation Technique (see Little and Rubin (1987) and Rancourt, Lee and Särndal (1994)) and (2) Adjustment of the inclusion probabilities (Amahia, Chaubey and Rao (1989)). Imputation procedures replace the missing values in order to modify the estimators while adjustment procedures adjust the inclusion probabilities for the group of respondents so as to compensate for the nonrespondents.

In this paper we will focus only on adjusting the inclusion probabilities for the purpose of estimation of the population mean (or total). The estimator to be investigated here is the traditional Horvitz-Thompson (H-T) estimator $t_{HT} = \frac{1}{N}\sum_{k\epsilon S}\frac{y_k}{\pi_k}$. Section 2 presents five methods for adjusting the inclusion probabilities in the presence of unit nonresponse. Section 3 gives detailed analysis of the bias inherent in the adjusted estimators along with some theoretical comparisons. The expressions for the average and variance of these biases under a superpopulation model are also provided here for all the adjustments. A Monte Carlo study of the adjustments is carried out in section 4 in order to numerically investigate the bias and MSE properties of the proposed estimators. The final section 5 contains some general conclusions.

## 2. PROPOSED ADJUSTMENTS OF THE INCLUSION PROBABILITIES

We consider the case of a general sampling design where the H-T estimator $t_{HT} = \frac{1}{N}\sum_{k\epsilon S}\frac{y_k}{\pi_k}$ would be used if the complete sample was available, where $S$ denotes the complete sample according to some sampling design with inclusion probabilities $\pi_1, \pi_2, ..., \pi_N$. However, in the case of nonresponse we assume that $r$ out of $n$ units $(r<n)$ respond, so that a complete sample is not

[1] Y.P. Chaubey and A.N. Crisalli, Concordia University, Montréal, Québec, Canada, H3C 1M8.

available. Our proposal is to adjust the inclusion probabilities and use the same form of the estimator as given above replacing $S$ by $S_R$ where $S_R$ denotes the group of respondents. These adjustments are expected to compensate for the non-respondents and are described below.

### 2.1 Adjustment I

Here we set $\pi_k^{(1)} = (\pi_k)^{r/n}$ for those units that respond, *i.e.,* $k \epsilon \{1,2,...,r\}$. If $r=n$ then $\pi_k^{(1)} = \pi_k$, *i.e.,* no adjustment is necessary because we have full response. On the other hand, if $r=0$ then $\pi_k^{(1)} = 1, \forall k$. However, in this case we have no units in the sample and therefore adjustment is of no consequence.

### 2.2 Adjustment II

Set $\pi_k^{(2)} = (\pi_k)(\frac{r}{n})$ for those units that respond, *i.e.,* $k \in \{1, 2, ..., r\}$. Again if $r=n$ then $\pi_k^{(2)} = \pi_k$, *i.e.,* no adjustment is necessary because we have full response. However, now if $r=0$ then $\pi_k^{(2)} = 0$ and this is again inconsequential as in the previous case.

**Remark 1.** *This adjustment has an interesting property* $\sum_1^N \pi_k^{(2)} = r$, *parallel to the case of full response when* $\sum_1^N \pi_k = n$ *(see Cassel, Särndal, and Wretman (1977)).*

### 2.3 Adjustment III

The following adjustment derives from the fact that $\pi_k = Pr[I_k = 1 \mid k \epsilon S]$ where

$$I_k = \begin{cases} 1 \text{ if } k \in S \\ 0 \text{ if } k \notin S \end{cases}$$

Let, therefore, $U_R$ represent the set of units in a population that would respond to a survey question and $U_{R'}$ its complement and define the adjusted inclusion probability as:

$$\begin{aligned} \tilde{\pi}_k^{(3)} &= \Pr(I_k = 1 \mid k \in U_R) \\ &= \pi_k \frac{\Pr(k \in U_R \mid I_k = 1)}{\Pr(k \in U_R)} \end{aligned} \tag{2.1}$$

This adjustment, however, is practically not feasible because we seldom have any knowledge of $Pr(k \epsilon U_R)$ and $Pr(k \epsilon U_R \mid I_k = 1)$ in a survey situation. Hence, we try to estimate these probabilities with a rejective sampling scheme. The following theorem from Hajek (1981) is beneficial in our quest:

**Theorem 2.1** Consider rejective sampling where $\sum_1^N p_i = n$. Then the probabilities of inclusion $\pi_i$ are equal to

$$\pi_i = p_i \left[ 1 - \frac{(\bar{p} - p_i)(1-p_i)}{d} + o(d^{-1}) \right] \forall i \in \{1,2,...,N\}$$

where

$$d = \sum_1^N p_i(1-p_i)$$

and

$$\bar{p} = \frac{\sum_1^N p_i^2(1-p_i)}{d}$$

*The term* $o(d^{-1}) \to 0$ *as* $d \to \infty$, *uniformly in* $i, \forall i \epsilon \{1,2,...,N\}$.

Using the above theorem our revised adjustments to the inclusion probabilities are:

$$\begin{aligned} \pi_k^{(3)} &= \frac{r}{n} \tilde{\pi}_k^{(3)} \\ &= \frac{r}{n} \pi_k \left[ 1 - (\tilde{\pi}_k - \pi_k)(1 - \pi_k \frac{)}{d} \right] \end{aligned} \tag{2.2}$$

where

$$d = \sum_1^N \pi_k(1-\pi_k)$$

and

$$\tilde{\pi}_k = \frac{\sum_1^N \pi_k^2(1-\pi_k)}{d}.$$

Now the adjusted probabilities have similar properties as in the previous case when $r=n$ and $r=0$.

**Remark 2.** *Since* $n > r$ *we regard the sample* $S$ *sampled as a population in which the response rate removed* $n-r$ *units by rejective sampling. Hence for the remaining* $r$ *units that respond we adjust their inclusion probabilities so as to insure that* $\sum_1^N \pi_k^{(3)} = r$.

### 2.4 Adjustment IV

The following adjustment is motivated from Amahia, Chaubey and Rao(1989). This adjustment involves some knowledge of correlation between the main characteristic and an auxiliary characteristic. Letting $\hat{\rho}$ be the correlation coefficient of $y_k$ and a covariate $x_k$, the adjustment will be as follows;

$$\pi_k^{(4)} = \frac{r}{n} \left[ (1-\hat{\rho}) \frac{n}{N} + \hat{\rho} \pi_k \right] \tag{2.3}$$

**Remark 3.** *We have an interesting property regarding the sum of* $\pi_k^{(4)}$.

$$\sum_1^N \pi_k^{(4)} = r.$$

### 2.5 Adjustment V

The concept of covariates used in the previous adjustment can also be helpful in providing another adjustment by combining adjustment I and II:

$$\pi_k^{(5)} = (1-\hat{\rho})(\pi_k)^{\frac{r}{n}} + \hat{\rho} \frac{r}{n} (\pi_k) \tag{2.4}$$

where as before $\hat{\rho}$ is the correlation between $y_k$ and some covariate $x$.

**Remark 4.** *Various adjustments proposed here may be combined when response rates are correlated with some covariate differently. Furthermore, other adjustments using imputation techniques may also be incorporated in modifying the estimators. These are being pursued under a separate study.*

## 3. ANALYSIS OF BIAS OF THE ADJUSTED ESTIMATORS

Let $\pi_k' = \pi_k(r,n)$ be an adjustment of the inclusion probability $\pi_k$ when $r$ out of $n$ sampled units respond. Also let $E_P(\,)$ be the expected value for a given response rate with respect to the sampling plan $P$ that produced these inclusion probabilities $\pi_k$. An expression for the bias of the estimator resulting from the above adjustment is provided in the following theorem.

### 3.1 General Model

**Theorem 3.1.** *The conditional bias of the Horvitz-Thompson estimator given a fixed response rate* $(r/n)$ *for the adjusted estimator* $\bar{t}_{HT} = \frac{1}{N}\sum_{k\epsilon S}\frac{y_k}{\pi_k}$ *is given by:*

$$B_P(\bar{t}_{HT}) = \frac{1}{N}\sum_{1}^{N}\left[\frac{\pi_k}{\pi_k'} - 1\right] y_k$$

**Remark 5.** $B_p(\bar{t}_{HT}) = 0$ *when* $\pi_k' = \pi_k$ *that is when* $r = n$.

We now compare the conditional biases of the above adjustments under a *Simple Random Sampling* design. Thus, replacing $\pi_k$ by $\frac{n}{N}$ in each of the above adjustments we obtain the following results:

$$\begin{aligned}
B_P^{(1)}(\bar{t}_{HT}) &= \left[\left(\frac{n}{N}\right)^{\frac{n-r}{n}} - 1\right]\bar{Y}\\
B_P^{(2)}(\bar{t}_{HT}) &= \frac{n-r}{r}\,\bar{Y}\\
B_P^{(3)}(\bar{t}_{HT}) &= \frac{n-r}{r}\,\bar{Y}\\
B_P^{(4)}(\bar{t}_{HT}) &= \left[\frac{n-r}{r+\hat{\rho}(n-r)}\right]\bar{Y}\\
B_P^{(5)}(\bar{t}_{HT}) &= \left[\frac{(1-\hat{\rho}\frac{r}{n})\frac{n}{N} - (1-\hat{\rho})\left(\frac{n}{N}\right)^{\frac{r}{n}}}{(1-\hat{\rho})\left(\frac{n}{N}\right)^{\frac{r}{n}} + \hat{\rho}\frac{r}{N}}\right]\bar{Y}
\end{aligned} \tag{3.1}$$

### 3.2 Superpopulation Model

A more general exploration of adjustments I to V can be attained by imposing a model on the $Y_k$. For this purpose we consider the regression model $\xi$, such that $\{Y_k:\ k\epsilon(1,2,\ldots,N)\}$ are independent and

$$\begin{aligned}
E_\xi(Y_k) &= \beta x_k\\
V_\xi(Y_k) &= \sigma^2 x_k
\end{aligned} \tag{3.2}$$

where $\beta$ and $\sigma^2$ are unknown constants and $x_k$ are known $\forall k\epsilon\{1,2,\ldots,N\}$. Using this model, under the Simple Random Sampling Design, we are in a position to find the expected value and variance of the bias of adjustments I to V.

The expected biases under the above model for different adjustments are given by:

$$\begin{aligned}
E_\xi(B_P^{(1)}(t_{HT})) &= \beta\left[\left[\frac{n}{N}\right]^{\frac{n-r}{n}} - 1\right]\bar{X}\\
E_\xi(B_P^{(2)}(t_{HT})) &= \beta\frac{n-r}{r}\bar{X}\\
E_\xi(B_P^{(3)}(t_{HT})) &= \beta\frac{n-r}{r}\bar{X}\\
E_\xi(B_P^{(4)}(t_{HT})) &= \beta\left[\frac{n-r}{r+\hat{\rho}(n-r)}\right]\bar{X}\\
E_\xi(B_P^{(5)}(t_{HT})) &= \beta\left[\frac{(1-\hat{\rho}\frac{r}{n})\frac{n}{N} - (1-\hat{\rho})(\frac{n}{N})^{\frac{r}{n}}}{(1-\hat{\rho})(\frac{n}{N})^{\frac{r}{n}} + \hat{\rho}\frac{r}{n}}\right]\bar{X}
\end{aligned} \tag{3.3}$$

The variances of the adjustments' bias are given by:

$$\begin{aligned}
V_\xi(B_P^{(1)}(\bar{t}_{HT})) &= \left[\left[\frac{n}{N}\right]^{\frac{n-r}{n}} - 1\right]^2\frac{\bar{X}\sigma^2}{n}\\
V_\xi(B_P^{(2)}(\bar{t}_{HT})) &= \left[\frac{n-r}{r}\right]^2\frac{\bar{X}\sigma^2}{N}\\
V_\xi(B_P^{(3)}(\bar{t}_{HT})) &= \left[\frac{n-r}{r}\right]^2\frac{\bar{X}\sigma^2}{N}\\
V_\xi(B_P^{(4)}(\bar{t}_{HT})) &= \left[\frac{n-r}{r+\hat{\rho}(n-r)}\right]^2\frac{\bar{X}\sigma^2}{N}\\
V_\xi(B_P^{(5)}(\bar{t}_{HT})) &= \left[\frac{(1-\hat{\rho}\frac{r}{n})\frac{n}{N} - (1-\hat{\rho})(\frac{n}{N})^{\frac{r}{n}}}{(1-\hat{\rho})(\frac{n}{N})^{\frac{r}{n}} + \hat{\rho}\frac{r}{n}}\right]^2\bar{X}\frac{\sigma^2}{N}.
\end{aligned} \tag{3.4}$$

Therefore for finite populations generated by the model $\xi$, the bias and its variability fluctuates around $\bar{X}$ by a constant multiple. Since $x_k$ is known for all $k$'s, we can therefore estimate $E_\xi(B)$ by $\bar{e}(B)$ after we get estimates for $\beta$ and $\sigma^2$ by usual statistical procedures.

Hence we can adjust the usual Horvitz-Thompson estimator so as to obtain a model unbiased estimator of $\bar{Y}$ by considering the following estimator:

$$\hat{t}_{HT} = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k'} + \bar{e}(B) \qquad (3.5)$$

where $\pi_k'$ is one of the adjustments discussed in this paper.

## 4. THE SIMULATION PROCEDURE AND THE RESULTS

We will now consider five adjusted Horvitz-Thompson estimators corresponding to the five adjustments given in section 2. A simulation study was undertaken so as to verify if the estimators succeeded in correcting $t_{HT}$ for five different response mechanisms using four different response rates. We basically followed the same strategy and data of Rancourt, Lee and Särndal (1994). Our objective was as theirs "*to examine the corrected estimators when the finite population follows the regression model* $\xi$" given in section 3.

For the simulation study, we generated 3 different populations, each of size $N=100$. The first 100 $x_k$ values were generated by a $\Gamma$-distribution with $\alpha=3$ and $\beta=16$. Conditional on $x_k$, a covariate $y_k$ was generated so that $y_k$ is $\Gamma$-distributed with $\alpha' = \frac{\beta^2 x_k^2}{\sigma^2}$ and $\beta' = \frac{\sigma^2}{\beta x_k}$.

We considered five nonresponse mechanisms each defined by independent Bernoulli $(\theta_k)$ trials such that the probability of nonresponse $\theta_k$ for unit $k$ is given by:

**Mechanism I:** $\theta_k$ is constant and independent for all units in the population.

**Mechanism II:** $\theta_k = \exp(-\delta x_k)$ in this case $\theta_k$ is a decreasing function of $x_k$.

**Mechanism III:** $\theta_k = 1 - \exp(-\delta x_k)$ in this case $\theta_k$ is an increasing function of $x_k$.

**Mechanism IV:** $\theta_k = \exp(-\delta y_k)$ in this case $\theta_k$ is a decreasing function of $y_k$.

**Mechanism V:** $\theta_k = 1 - \exp(-\delta y_k)$ in this case $\theta_k$ is an increasing function of $y_k$.

For Mechanisms II to V, the constant $\delta$ was determined so that $\hat{\theta} = \frac{1}{N}\sum_{k=1}^{N}\theta_k$ for the response rates of 90%, 80%, 70% and 60%. Therefore, for each population (3), there were 20 different combinations of response mechanism and response rates, for a total of 3 x 20 = 60 different experiments. Now for each experiment 400 samples of size $n=30$ were drawn and for each sample 25 response sets were generated. Therefore 400 X 25 = 10,000 simulation sets were obtained for analysis.

The performance of the estimators is judged by the magnitudes of the relative bias (RB) and the relative root mean square error (RRMSE). The RB and RRMSE of the Horvitz-Thompson estimator $t_{HT}$ are defined respectively by

$$RB(\hat{t}_{HT}) = \frac{E_P E_M(\hat{t}_{HT}-\bar{Y})}{\bar{Y}} \times 100 \qquad (4.1)$$

$$RRMSE(\hat{t}_{HT}) = \frac{\sqrt{E_P E_M(\hat{t}_{HT}-\bar{Y})^2}}{\bar{Y}} \times 100 \qquad (4.2)$$

where $E_P()$ is the expectation under the sampling plan and $E_M()$ the expectation of the response mechanism. These averages were found by Monte Carlo simulations using the 600,000 simulation sets. In the following tables, ARB denotes the absolute relative bias or $|RB(\hat{t}_{HT})|$.

We have grouped our results for $|RB(\hat{t}_{HT})|$ and RRMSE$(\hat{t}_{HT})$. Under a *Simple Random Sampling Design* and *the regression model* $\xi$ Adjustment II and Adjustment III have the same properties. Therefore we only considered the adjusted Horvitz-Thompson estimator $\hat{t}_{HT}$, for the four adjustments I, II, IV and V. To be more specific we used the following estimate of $\hat{t}_{HT}$

$$\hat{t}_{HT} = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k'} + \bar{e}(B)$$

where $\pi_k'$ is one of the four adjustments discussed in this paper and $\bar{e}(B)$ an estimate of $E_\xi(B)$ the expected bias under one of the four adjustments.

**Table 1:** Average ARB and RRMSE for all Response Mechanisms for the three Populations and all Response Rates

| Mechanism | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| AARB | | | | | |
| Adj1 | 25.8 | 17.2 | 9.6 | 37.6 | 41.8 |
| Adj2 | 26.3 | 17.2 | 9.5 | 37.6 | 41.7 |
| Adj4 | 27.4 | 19.3 | 11.3 | 39.9 | 44.1 |
| Adj5 | 12.5 | 5.7 | 8.4 | 22.7 | 26.5 |
| ARRMSE | | | | | |
| Adj1 | 6.1 | 5.4 | 4.6 | 7.2 | 7.5 |
| Adj2 | 4.9 | 5.5 | 4.6 | 7.3 | 7.5 |
| Adj4 | 6.3 | 5.7 | 5.0 | 7.9 | 7.7 |
| Adj5 | 4.9 | 4.4 | 4.4 | 5.9 | 6.2 |

Based on the simulation study we may draw the following conclusions for the adjustments considered in this paper.

1. When **Mechanism I** holds then no adjustment is necessary because the response rate is constant for all units in the population. From Table 1 one finds *Adj5* to be the best amongst the adjustments considered.

2. When **Mechanism II** or **III** holds, which implies that the response probability is dependent on the covariate $x_k$, Table 1 shows that *Adj5* is the best adjustment. The AARB and ARRMSE in using this adjustment are smaller than those of the other three adjustments in the majority of the experiments considered using both mechanisms. When we compare our Table 1 with Table 2 of Rancourt, Lee and Särndal (1994) we make the following observations about adjustment 5:

   (i) **Mechanism II**: They showed that when the population was a ratio type, $k_2$ was the best adjustment. The AARB($k_2$) = 2.4 while this paper has demonstrated that AARB *(Adj5)* = 5.7. Also the ARRMSE ($k_2$) = 12.0 while ARRMSE (*Adj5*) = 4.4. Therefore, the average absolute relative bias of $k_2$ is smaller than that of *Adj5* but the average relative root mean square error $k_2$ is more than twice that of *Adj5*.

   (ii) **Mechanism III**: The best adjustment for the ratio population was $k_4$. Now the AARB($k_4$) = 7.3 while AARB(*Adj5*) = 8.4 and the ARRMSE ($k_4$) = 17.7 while ARRMSE(*Adj5*) = 4.4. Now the AARB of $k_4$ is slightly smaller than that of *Adj5* but its ARRMSE is almost four times as great.

   Therefore if **Mechanism II** or **III** affects the response rate then adjustment 5 is a better choice than $k_2$ or $k_4$ in terms of the AARMSE criterion.

3. When **Mechanism IV** or **V** hold, which implies that the response probability is a function of the variable investigated Table 1 shows that *Adj5* is the best amongst the adjustments considered here. For **Mechanism IV** or **V** the AARB and ARRMSE for this adjustment is much greater than that of **Mechanism II** or **III**. However, comparing Table 1 with Table 2 of Rancourt, Lee and Särndal (1994) we find that *Adj5* does not perform as well as the $c$ or $k$ correctors.

It has long been recognized that *nonresponse* in surveys causes bias in the estimates of the finite population parameters. Many solutions to this problem have been proposed. This paper has examined the adjustment of the inclusion probabilities. The *Adj5* was favourable in the majority of cases considered. The other adjustments did not produce results that were as favourable.

## REFERENCES


Amahia, G.N., Chaubey, Y.P., and Rao, T.J. (1989). "Efficiency of a new estimator in PPS sampling for multiple characteristics", *Journal of Statistical Planning and Inference*, 21, 75-84.

Cassel, C.G., Särndal, C.-E., and Wretman, J.H. (1977). *Foundations of inference in survey sampling*, New York: Wiley.

Hajek, J. (1981). *Sampling from a finite population*, New York: Marcel Dekker.

Little, R.J.A., and Rubin, D.B. (1987). *Statistical analysis with missing data*, New York: Wiley.

Rancourt, E., Lee, H., and Särndal, C.-E. (1994). "Bias corrections for survey estimates from data with ratio imputed values for confounded nonresponse", *Survey Methodology*, 20, 137-147.

Rubin, D.B. (1977). "Formalizing subjective notions about the effect of nonrespondents in sample surveys", *Journal of the American Statistical Association*, 72, 538-543.