ON THE USE OF AUXILIARY DATA IN POISSON SAMPLING

M.A. Hidiroglou, and John G. Slanta[1]

Keywords: Poisson Sampling, Bernoulli Sampling, Horvitz-Thompson Estimator, Hájek Estimator, Brewer Estimator, GREG Estimator.

1. INTRODUCTION

Sampling units independently at each draw from a frame can be achieved using Bernoulli or Poisson sampling. The use of Bernoulli sampling implies that the probability of selection (π_k) is constant for each draw, whereas Poisson sampling implies that the probability of selection differs at every draw, normally being proportional to the size measure of the unit being selected. Estimators and associated variances can be of a Horvitz-Thompson type. However, auxiliary data in the form of population counts or ***x***-variables can be used advantageously to improve the efficiency of the estimator. The well-known Hájek estimator is such an example.

In the case of Bernoulli sampling, it will always have a variance that is lower than the one corresponding to the Horvitz-Thompson estimator. However, this is not always the case with Poisson sampling. Thompson and Sigman (2000) presented cases from a simulation study where the Hájek estimated variance exceeded the Horvitz-Thompson given a high correlation between probability of selection (π_k) and the variable of interest (y_k). Other possible estimators include the Brewer estimator and the optimal regression estimator

In this paper, we study the properties of the Hájek, Brewer, and optimal regression estimators. We provide conditions that determine as to which estimator has the smaller population variance. These estimators and their associated variances are numerically compared using data from the Industrial Research and Development Survey.

2. COMPARING POPULATION VARIANCES BETWEEN THE HORVITZ-THOMPSON ESTIMATOR AND THE HÁJEK ESTIMATOR

Let *U* be the population of interest, and *s* the resulting sample. For Poisson sampling, Hájek (1964), the sample s is selected by carrying out *N* binomial trials to determine whether each unit in the population is to be included in the sample or not. The Horvitz-Thompson estimator of total is $\hat{Y}_{HT} = \sum_s (y_k / \pi_k)$. Auxiliary information such as population counts can be used to improve the efficiency of $\hat{Y}_{HT}$, resulting in a number of estimators other than $\hat{Y}_{HT}$. One such estimator is the Hájek estimator. It is given by $\hat{Y}_{HAJ} = (N / \hat{N})\hat{Y}_{HT}$, where $\hat{N} = \sum_s (1/\pi_k)$. The form of this estimator is the well-known ratio-type. The associated population variances with the Hájek and Horvitz-Thompson estimators are respectively:

$$V(\hat{Y}_{HAJ}) \approx \sum_U \phi_k (y_k - \bar{y}_U)^2$$

and

$$V(\hat{Y}_{HT}) = \sum_U \phi_k \, y_k^2$$

where $\phi_k = (1/\pi_k) - 1$ and $\bar{y}_U = (\sum_U y_k) / N$.

Result 1: Assume that $\bar{y}_U > 0$. Then, the population variance of $\hat{Y}_{HAJ}$ and $\hat{Y}_{HT}$ respect the following conditions:

(i) $V(\hat{Y}_{HAJ}) < V(\hat{Y}_{HT})$, if $\bar{y}_U < \dfrac{2\sum_U \phi_k y_k}{\sum_U \phi_k}$

(ii) $V(\hat{Y}_{HAJ}) \geq V(\hat{Y}_{HT})$ otherwise.

Note that if $\bar{y}_U < 0$, then $V(\hat{Y}_{HAJ}) < V(\hat{Y}_{HT})$ if $\bar{y}_U > \dfrac{2\sum_U \phi_k y_k}{\sum_U \phi_k}$.

Proof: If $V(\hat{Y}_{HAJ}) < V(\hat{Y}_{HT})$, then we have that

$$\sum_U \phi_k \left(y_k^2 - 2\, y_k\, \bar{y}_U + \bar{y}_U^2\right) < \sum_U \phi_k \, y_k^2$$

or $\sum_U \phi_k (\bar{y}_U - 2\, y_k) < 0$. The result follows, that is, $\bar{y}_U \sum_U \phi_k < 2 \sum_U \phi_k \, y_k$.

[1] M.A. Hidiroglou, Statistics Canada and John G. Slanta, U.S. Bureau of the Census

We assume from hereon that $\bar{y}_U > 0$.

In the case of Bernoulli sampling $V\left(\hat{Y}_{HAJ}\right) < V\left(\hat{Y}_{HT}\right)$. This follows from result 1. Since $\phi_k = (N-n)/n$, we have that $\frac{2\sum_U \phi_k y_k}{\sum_U \phi_k} = \frac{2\sum_U [(N-n)/n] y_k}{\sum_U (N-n)/n} = 2\bar{y}_U > \bar{y}_U$. This implies that $V\left(\hat{Y}_{HAJ}\right) < V\left(\hat{Y}_{HT}\right)$. We next construct two examples to illustrate result 1.

Example 1: Let $\pi_k = n\frac{a}{N}$ for k = 1, 2, …, N-1. We have that $\pi_N = n - \frac{na(N-1)}{N} = n\left(1 - a + \frac{a}{N}\right)$ because $\sum_U \pi_k = n$. The bounds of a that guarantee $0 < \pi_k < 1$ are $\frac{N(n-1)}{n(N-1)} < a < \frac{N}{N-1}$ for $1 < n < N$.

Define $D_1 = \left(\frac{N}{na} - 1\right)$ and $D_2 = \left(\frac{N}{n(N - Na + a)} - 1\right)$.

Then

$$\sum_U \phi_k = (N-1)D_1 + D_2 \qquad (1)$$

and

$$\sum_U \phi_k y_k = D_1(N\bar{y}_U - y_N) + D_2 y_N \qquad (2)$$

If $V\left(\hat{Y}_{HAJ}\right) \geq V\left(\hat{Y}_{HT}\right)$, then

$$\bar{y}_U \geq \frac{2\sum_U \phi_k y_k}{\sum_U \phi_k} \qquad (3)$$

Expression (3) is an inequality, we can solve for "a" by expressing it as an equality. That is, substituting (1) and (2) into (3), we obtain

$$\bar{y}_U = 2\frac{D_1(N\bar{y}_U - y_N) + D_2 y_N}{D_1(N-1) + D_2} \qquad (4)$$

We solve for "a" using:

$$2[D_1(N\bar{y}_U - y_N) + D_2 y_N] - \bar{y}_U[D_1(N-1) + D_2] = 0 \qquad (5)$$

Equation (5) is a quadratic in "a", and its two solutions are $a_1 = \frac{-B - \sqrt{B^2 - 4AC}}{2A}$ and $a_2 = \frac{-B + \sqrt{B^2 - 4AC}}{2A}$, where $A = n(N-1)\bar{y}_U$, $B = N[2y_N - (N+n)\bar{y}_U]$, and $C = N[(N+1)\bar{y}_U - 2y_N]$. If "$a$" is contained in the closed interval, $[a_1, a_2]$, then $\bar{y}_U \geq \frac{2\sum_U \phi_k y_k}{\sum_U \phi_k}$, implying that $V\left(\hat{Y}_{HAJ}\right) \leq V\left(\hat{Y}_{HT}\right)$. If "$a$" is not contained in the interval $[a_1, a_2]$, then $\bar{y}_U < \frac{2\sum_U \phi_k y_k}{\sum_U \phi_k}$ implying that $V\left(\hat{Y}_{HAJ}\right) > V\left(\hat{Y}_{HT}\right)$.

Example 2 The variable of interest y_k and the inclusion probability π_k are normally linked via a simple regression model of the form $y_k = b_1\pi_k + b_0 + e_k$. We have that $V\left(\hat{Y}_{HAJ}\right) < V\left(\hat{Y}_{HT}\right)$ if the following conditions hold:
(i) $b_1 < 0$ and $b_0 > 0$; (ii) $E_U(e_k) = 0$; and
(iii) $Cov_U(\phi_k, e_k) = 0$.

Proof: Using the above conditions (ii) and (iii), the slope b_1 and intercept b_0 are respectively:
$b_1 = \frac{Cov_U(y_k, \phi_k)}{Cov_U(\pi_k, \phi_k)}$ and $b_0 = \bar{y}_U - b_1\bar{\pi}_U$, where $\bar{\pi}_U = n/N$. Since $E_U(e_k) = 0$ and $Cov_U(\phi_k, e_k) = 0$, we have that $\bar{y}_U = \frac{b_1 n}{N} + b_0$, and $\frac{2\sum_U \phi_k y_k}{\sum_U \phi_k} = \frac{2b_1(N-n)}{\sum_U \phi_k} + 2b_0$.

Recall the Cauchy -Schwarz inequality is $\left(\sum_U x_k y_k\right)^2 \leq \left[\left(\sum_U x_k^2\right)\left(\sum_U y_k^2\right)\right]$. Using this inequality we have that $\left(\sum_U 1\right)^2 \leq \left[\left(\sum_U \frac{1}{\pi_k}\right)\left(\sum_U \pi_k\right)\right]$. This implies that $\frac{N(N-n)}{n} \leq \sum_U \phi_k = \left(\sum_U \frac{1}{\pi_k}\right) - N$. Hence $\frac{2(N-n)}{\sum_U \phi_k} \leq \frac{2n}{N}$.

Since $b_1 < 0$ and $b_0 > 0$, $2\bar{y}_U \leq \frac{2(N-n)b_1}{\sum_U \phi_k} + 2b_0$, and from Result 1 it follows that $V\left(\hat{Y}_{HAJ}\right) < V\left(\hat{Y}_{HT}\right)$.

We next provide a numerical illustration of result 1. Let the population size be N=100, and the sample size n=10, $\bar{y}_U = 15$ and y_N = 964. The two solutions to equation 5 are $a_1 \approx -2.84843$ and $a_2 \approx 0.97638$. Note that a is bounded by $(0.\overline{90}, 1.\overline{01})$ to guarantee all π_k are between zero and one. For $V\left(\hat{Y}_{HAJ}\right) \leq V\left(\hat{Y}_{HT}\right)$, (when N =100,

n =10, $\bar{y}_U = 15$, and y_N =964), choose a in the range $(0.\overline{90}, 0.97638]$. When a is $0.\overline{90}$, there is one unit selected with probability one (π_N = 1) and the remaining units is a Bernoulli sample. The reason that $V(\hat{Y}_{HAJ}) \geq V(\hat{Y}_{HT})$ is because y_N is used for the computation of $\bar{y}_U$ in the function *g(a)*.

The plot of $g(a) = 2\dfrac{D_1(N\bar{y}_U - y_N) + D_2 y_N}{D_1(N-1) + D_2}$ versus "*a*" is given in Figure 1.

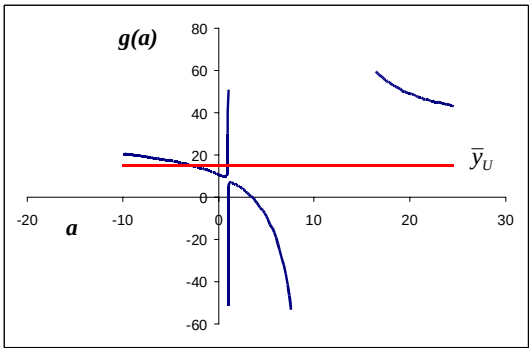

Figure 1: Plot of *g(a)* versus *a*

Figure 1 covers a wider range of *a* than constraints allow only so that the reader can see what the shape of the function *g(a)* looks like. There are two vertical asymptotes and the horizontal asymptote is at $g(a) = 2\bar{y}_U$.

3. COMPARING POPULATION VARIANCES BETWEEN THE HORVITZ-THOMPSON ESTIMATOR AND THE BREWER'S ESTIMATOR

Denote as a_k as an indicator variable taking the value one if the *k-th* unit is selected into the sample, and zero otherwise. The realized sample is then $m = \sum_U a_k$, where $m \leq N$. The expected sample size is $n = \sum_U \pi_k$. An alternative estimator to the Hájek estimator is $\hat{Y}_{BREW} = (n/m)\hat{Y}_{HT}$, given in Brewer and Hanif (1983). Its population variance is

$$V(\hat{Y}_{BREW}) \approx \sum_U \pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k} - \frac{N}{n}\bar{y}_U\right)^2$$

We next provide conditions when this variance is greater or smaller than the one associated with the Horvitz-Thompson estimator.

Result 2: The population variance of $\hat{Y}_{BREW}$ and $\hat{Y}_{HT}$ respect the following conditions:

(i) $V(\hat{Y}_{BREW}) < V(\hat{Y}_{HT})$ if $\bar{y}_U < \dfrac{2n\sum_U (1-\pi_k)y_k}{N(n-\sum_U \pi_k^2)}$.

(ii) $V(\hat{Y}_{BREW}) \geq V(\hat{Y}_{HT})$ if $\bar{y}_U \geq \dfrac{2n\sum_U (1-\pi_k)y_k}{N(n-\sum_U \pi_k^2)}$.

Proof: If $V(\hat{Y}_{BREW}) < V(\hat{Y}_{HT})$, then we have that

$$\sum_U \pi_k(1-\pi_k)\left(\frac{y_k}{\pi_k} - \frac{N}{n}\bar{y}_U\right)^2 < \sum_U \phi_k y_k^2$$

$$= \sum_U \left[\frac{N^2}{n^2}\bar{y}_U^2 \pi_k(1-\pi_k) - \frac{2N}{n}\bar{y}_U(1-\pi_k)y_k\right] < 0$$

$$= \sum_U \left[\frac{N^2}{n^2}\bar{y}_U \pi_k(1-\pi_k) - \frac{2N}{n}(1-\pi_k)y_k\right] \bar{y}_U < 0$$

Therefore, $V(\hat{Y}_{BREW}) < V(\hat{Y}_{HT})$ implies that $\dfrac{N^2}{n^2}\bar{y}_U \sum_U \pi_k(1-\pi_k) < \dfrac{2N}{n}\sum_U (1-\pi_k)y_k$. Hence we have that $\bar{y}_U < \dfrac{2n\sum_U (1-\pi_k)y_k}{N(n-\sum_U \pi_k^2)}$.

Once more for Bernoulli sampling we have that $V(\hat{Y}_{BREW}) < V(\hat{Y}_{HT})$, because $\pi_k = n/N$ and

$$\frac{2n\sum_U (1-\pi_k)y_k}{N(n-\sum_U \pi_k^2)} = 2\bar{y}_U.$$

We next construct two examples to illustrate result 2.

Example 3: Using the same pattern of π_k values given in example 1, the population variances of the Horvitz-Thompson and Brewer's estimator are equal when:

$$\bar{y}_U = 2\frac{n}{N}\frac{(N-na)\bar{y}_U - n(1-a)y_N}{n-(N-1)\left(\frac{na}{N}\right)^2 - \left[n\left(1-a+\frac{a}{N}\right)\right]^2}$$

Once more, this is a quadratic equation in the *a*'s, where $A = -n^2N(N-1)\bar{y}_U$, $B = 2n^2N(N\bar{y}_U - y_N)$, and $C = -nN[N(n+1)\bar{y}_U - 2ny_N]$. Let *N*=100, *n*=10, $\bar{y}_U = 15$ and y_N = 964. The two solutions to $\bar{y}_U = g(a)$ are $a_1 \approx -0.2025$ and $a_2 \approx 0.9244$. For $V(\hat{Y}_{BREW}) > V(\hat{Y}_{HT})$, (when *N*=100, *n*=10, $\bar{y}_U = 15$ and y_N = 964) choose *a* in the range $(0.\overline{90}, 0.9244]$

Plotting

$$g(a) = 2\frac{n}{N}\frac{(N-na)\bar{y}_U - n(1-a)y_N}{n-(N-1)\left(\frac{na}{N}\right)^2 - \left[n\left(1-a+\frac{a}{N}\right)\right]^2}$$

versus "*a*" is given in Figure 2.

Figure 2 also covers a wider range of *a* than constraints allow only so that the reader can see what the shape of the function *g(a)* looks like. There are two vertical asymptotes and the horizontal asymptote is at *g(a)* = 0.

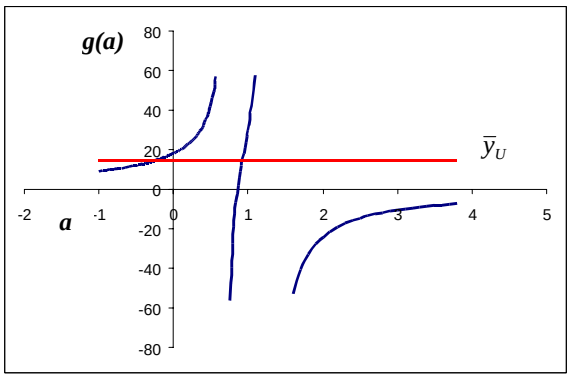

Figure 2: Plot of *g(a)* versus *a*

Note that if *a* = 1, i.e. Bernoulli Sampling, then the Hájek and Brewer estimators always have lower variances and 1 is always within the constraints of *a*. One can choose a set of (*N*, *n*, $\bar{y}_U$, and y_N) such that the Horvitz-Thompson estimator will always have higher variance, i.e. the solutions of *a* for $\bar{y}_U = g(a)$ fall outside the constraints of *a*. Example: *N* = 100, *n* = 10, $\bar{y}_U = 20$ and $y_N = 964$.

Example 4: Assume that the model linking the variable of interest y_k and the inclusion probability π_k is $y_k = b_1\pi_k + b_0 + e_k$. Then $V\left(\hat{Y}_{BREW}\right) < V\left(\hat{Y}_{HT}\right)$ if the following conditions hold: (i) $b_0 > 0$; and (ii) $E_U(e_k) = 0$; and (iii) $Cov_U(\pi_k, e_k) = 0$.

Note that this is a weaker condition that the one given in example 2 for the Hájek procedure.

Proof: Using the above conditions (ii) and (iii) the slope b_1 and intercept b_0 are

$$b_1 = \frac{Cov_U(y_k,\pi_k)}{\sigma_U^2(\pi_k)} \text{ and } b_0 = \bar{y}_U - b_1\bar{\pi}_U .$$

Conditions (ii) and (iii) respectively imply that

$$\bar{y}_U = \frac{b_1 n}{N} + b_0, \text{ and, that}$$

$$\frac{2n\sum_U (1-\pi_k)y_k}{N(n-\sum_U \pi_k^2)} = 2\frac{b_1 n}{N} + 2b_0\frac{Nn-n^2}{Nn-N\sum_U \pi_k^2}$$

Using the Cauchy-Schwarz inequality

$$n^2 = \left(\sum_U 1\cdot\pi_k\right)^2 < \left(\sum_U 1^2\right)\left(\sum_U \pi_k^2\right) = N\left(\sum_U \pi_k^2\right)$$

Hence it follows that $\frac{Nn-n^2}{Nn-N\sum_U \pi_k^2} > 1$. Since $b_0 > 0$, then $2\bar{y}_U < \frac{2n\sum_U (1-\pi_k)y_k}{N(n-\sum_U \pi_k^2)}$.

4. COMPARING POPULATION VARIANCES BETWEEN THE HORVITZ-THOMPSON ESTIMATOR AND THE OPTIMUM REGRESSION ESTIMATOR

So far we have only used counts as auxiliary data. Assume that the data $\{(\boldsymbol{x}_k, y_k), k \in s\}$ are observed, where $\boldsymbol{x}_k$ is the value for unit *k*. The population total $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ of that variable is assumed to be known from a reliable source. The GREG estimator of *Y* is $\hat{Y}_{GREG} = \sum_s w_k\, g_k\, y_k$ where

$$g_k = 1 + \left(\boldsymbol{X} - \hat{\boldsymbol{X}}_{HT}\right)' \left(\sum_s \frac{w_k\, \boldsymbol{x}_k\, \boldsymbol{x}_k'}{c_k}\right)^{-1} \frac{\boldsymbol{x}_k}{c_k}$$, and

$w_k = 1/\pi_k$. The statistician specifies the choice of c_k. Särndal (1996) discussed various possible values of c_k at length. Deville and Särndal (1992) justified the estimator $\hat{Y}_{GREG}$ as follows. New weights $\tilde{w}_k$ are generated as close as possible to the basic sampling weights w_k, subject to the calibration constraint $\sum_s \tilde{w}_k\, \boldsymbol{x}_k = \boldsymbol{X}$. The new weights (calibration weights) are given as $\tilde{w}_k = w_k\, g_k$. when the minimized distance is given by $\sum_s c_k (w_k - \tilde{w}_k)/w_k$. The form of GREG for this distance measure can alternatively be expressed as:

$$\hat{Y}_{GREG} = \boldsymbol{X}'\hat{\boldsymbol{B}} + \sum_s w_k \left(y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}}\right) \qquad (7)$$

where $\hat{\boldsymbol{B}} = \left(\sum_s w_k\, \boldsymbol{x}_k\, \boldsymbol{x}_k'/c_k\right)^{-1} \sum_s w_k\, \boldsymbol{x}_k\, y_k/c_k$.

For Poisson sampling, the population variance of $\hat{Y}_{GREG}$ is given by $V\left(\hat{Y}_{GREG}\right) \approx \sum_U \phi_k\, e_k^2$ where $e_k = y_k - \boldsymbol{x}_k'\boldsymbol{B}$

and $\boldsymbol{B} = \left(\sum_U \frac{\boldsymbol{x}_k \boldsymbol{x}_k'}{c_k} \right)^{-1} \sum_U \frac{\boldsymbol{x}_k y_k}{c_k}$ is the census regression vector. Thompson and Sigman (2000) presented results from their simulation study that showed that $\hat{V}\left(\hat{Y}_{GREG}\right) = \sum_S w_k \phi_k \hat{e}_k^2$, with $\hat{e}_k = y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}$ is very strongly negatively biased.

Another alternative solution is to use the GREG to obtain the post-stratified version of the Hájek estimator. To this end, assume that N_p is known for sub-populations U_p of the universe *U*, and that $\bigcup_{p=1}^{P} U_p = U$, where $U_p \cap U_{p'} = \phi$ if $p \neq p'$. The post-stratified Hájek estimator is given by:

$$\hat{Y}_{HAJ,POST} = \sum_{p=1}^{P} \frac{N_p}{\hat{N}_p} \sum_{s_p} y_k / \pi_k$$

where $\hat{N}_p = \sum_{s_p} 1/\pi_k$, and $s_p = s \cap U_p$. The population variance of $\hat{Y}_{HAJ,POST}$ is given by:

$$V\left(\hat{Y}_{HAJ,POST}\right) = \sum_{p=1}^{P} \sum_{U_p} \phi_k \left(y_k - \bar{y}_{U_p}\right)^2 .$$

Note that $e_k = y_k - \bar{y}_{U_p}$, where $\bar{y}_{U_p} = \sum_{U_p} y_k / N_p$. We can guarantee that the population variance associated with $\hat{Y}_{HAJ,POST}$ will be smaller than the one associated with $\hat{Y}_{HT}$ if

$$\sum_{p=1}^{P} \sum_{U_p} \phi_k \left(y_k - \bar{y}_{U_p}\right)^2 < \sum_{p=1}^{P} \sum_{U_p} \phi_k y_k^2$$

implying that $\sum_{p=1}^{P} \sum_{U_p} \phi_k \bar{y}_{U_p}^2 < 2 \sum_{p=1}^{P} \sum_{U_p} \phi_k y_k \bar{y}_{U_p}$ for *p=1, …, P.*

Särndal (1996) proposed the design optimal version of $\hat{Y}_{GREG}$. For this case, it can be shown that minimising the variance of $\hat{Y}_{GREG}$ yields the optimal estimator $\hat{Y}_{OPT}$. The form of this estimator is

$$\hat{Y}_{OPT} = \boldsymbol{x}' \hat{\boldsymbol{B}}_{OPT} + \sum_s w_k \left(y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}}_{OPT}\right)$$

where $\hat{\boldsymbol{B}}_{OPT} = \left(\sum_s w_k \phi_k \boldsymbol{x}_k \boldsymbol{x}_k'\right)^{-1} \sum_s w_k \phi_k \boldsymbol{x}_k y_k$. Suppose that $\pi_k = n z_k / \sum_U z_k$ where z_k is a size measure specified for each $k \in U$. Särndal (1996) showed that:

$$V\left(\hat{Y}_{OPT}\right) = V\left(\hat{Y}_{HT}\right) - \mathbf{B}_{OPT}' \left(\sum_s \phi_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{B}_{OPT} \leq V\left(\hat{Y}_{HT}\right)$$

where $\boldsymbol{B}_{OPT} = \left(\sum_U \phi_k \boldsymbol{x}_k \boldsymbol{x}_k'\right)^{-1} \sum_U \phi_k \boldsymbol{x}_k y_k$.

The optimal case that corresponds to the count is obtained by setting $\boldsymbol{x}_k = 1$ for $k \in U$. This yields

$$\begin{aligned} \hat{Y}_{OPT} &= N \hat{B}_{OPT} + \sum_s w_k \left(y_k - \hat{B}_{OPT}\right) \\ &= \sum_s w_k y_k + \left(N - \sum_s w_k\right) \hat{B}_{OPT} \end{aligned}$$

where $\hat{B}_{OPT} = \dfrac{\sum_U w_k \phi_k y_k}{\sum_U w_k \phi_k}$.

Once more, if post-stratified counts are available, then the optimal count estimator can be suitably modified to account for them.

5. EMPIRICAL INVESTIGATIONS USING DATA FROM THE INDUSTRIAL RESEARCH AND DEVELOPMENT SURVEY

The Industrial Research and Development (R&D) survey is a company survey conducted annually and collects, among other things, information on R&D expenditures by types of R&D, by industry groupings, and by state. The frame is split-up into some certainty strata, some random sampling strata, and some non-certainty unequal probability strata. We focused our concentration on the latter. Each stratum was an industry grouping where measures of size x_k were assigned to each company, (sampling unit). The measures of size were calculated by using regression models. Prior year total R&D was the dependent variable and payroll was the independent variable. Values of π_k were originally assigned proportional to x_k, but some values of π_k were changed to meet the minimum probability requirement. We also only looked at one variable of interest, total R&D.

If one were to look at the population and substitute x_k for the y_k then almost all the time the Hájek estimator would have higher variance than the H-T estimator and almost all the time the Brewer estimator would have lower variance than the H-T estimator. What we then decided to do was to look at the sample variances of these estimators given real data from that survey. Again we remind the readers that we focused only on the non-certainty unequal probability strata.

There were 45 industry estimates of total R&D where the sample variance was greater than zero. Figure 3 shows a histogram where the horizontal axis is the percent change, rounded to nearest 10%, in the sample variance of the Hájek estimator to the H-T estimator and the vertical axis is the number of estimates that fall within the labeled

range. The range goes from a 50% reduction in variance to a 138% increase in variance.

Figure 3: Percent change in Hájek Estimator

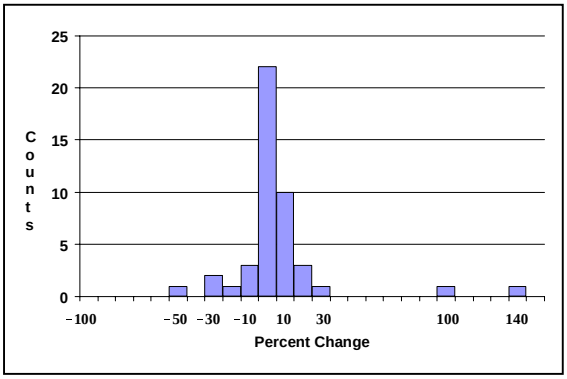

Figure 4 is the corresponding histogram for the Brewer estimator. The range goes from a 50% reduction in variance to a 24% increase in variance.

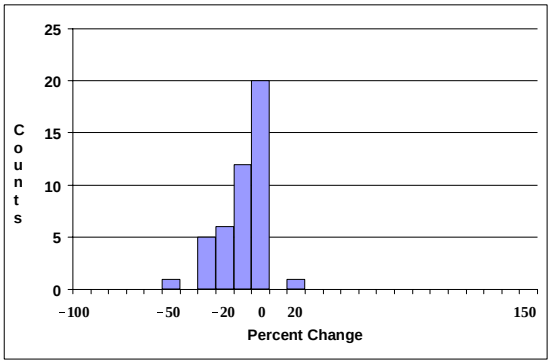

Figure 4: Percent change in Brewer Estimator

All but two sample variances of the Brewer estimator are lower than the sample variances of the H-T estimator.

Figure 5 is the corresponding histogram for the GREG estimator. The range goes from a 99% reduction in variance to a zero difference. The value of $\hat{\boldsymbol{B}}$ that was chosen, for calculation purposes, was $\hat{\boldsymbol{B}}_{OPT}$.

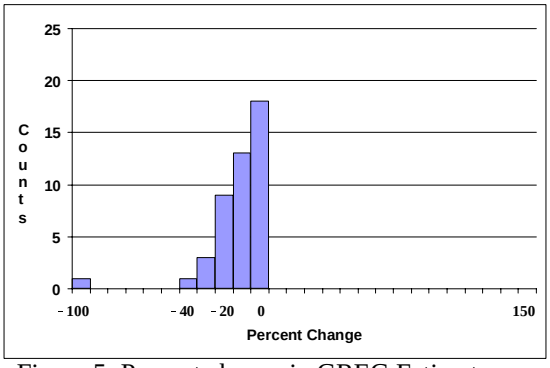

Figure 5: Percent change in GREG Estimator

As a final note, there were:

3 estimates where $\hat{V}\left(\hat{Y}_{HAJ}\right) < \hat{V}\left(\hat{Y}_{BREW}\right)$,
9 estimates where $\hat{V}\left(\hat{Y}_{HAJ}\right) < \hat{V}\left(\hat{Y}_{OPT}\right)$, and
13 estimates where $\hat{V}\left(\hat{Y}_{BREW}\right) < \hat{V}\left(\hat{Y}_{OPT}\right)$.

6. CONCLUSION

The choice of estimator between Horvitz-Thompson, Hájek, and Brewer depend on the distribution of population of y_k and π_k. If y_k and π_k are negatively correlated, then $\hat{Y}_{HAJ}$ is better than $\hat{Y}_{HT}$, while if $b_0 > 0$ then $\hat{Y}_{BREW}$ is better than $\hat{Y}_{HT}$. The best estimator, however, in terms of variance reduction, is the optimal regression estimator $\hat{Y}_{OPT}$.

References

Brewer, K.R.W. and Hanif, M. (1983). "*Sampling with unequal probabilities*", New York: Springer-Verlag.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35, 1491-1523.

Hájek, J. (1964). "*Sampling from a Finite Population*". Marcel-Decker Inc., New York.

Särndal, C.-E. (1996). Efficient Estimator with simple variance in unequal probability sampling", *Journal of the American Statistical Association*, 1289-1300.

Thompson, K.J. and Sigman, R.S. (2000). "*Calculating Variances For Poisson Samples In The Steps System: An Investigation Into Alternative Estimators*", Economic Statistical Methods Report Series ESM-0002.