# MTH 432M: Introduction to Sampling Theory
# Assignment No. 2: Horvitz-Thompson Estimator, SRSWR($n$) and SRSWOR($n$)

1. Let $\mathcal{Y} = ((X_1, Y_1), \ldots, (X_N, Y_N))$, $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $C_{\underline{X},\underline{Y}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$. Under the $SRSWR(n)$ design, find

   (i) the correlation coefficient between $\hat{X}_{HT}$ and $\hat{Y}_{HT}$, the Horvitz-Thompson estimators of $X = N\overline{X}$ and $Y = N\overline{Y}$.

   (ii) an unbiased estimator of $C_{\underline{X},\underline{Y}}$.

2. Let $\mathcal{U} = (B_1, B_2, \ldots, B_6)$ and $\mathcal{Y} = (Y_1, Y_2, \ldots, Y_6) = (10, 21, 16, 6, 7, 11)$, where, for $i = 1, 2, \ldots, 6$, $B_i$ is the $i^{th}$ bank in a country and $Y_i$ is the number non-performing loans given by $i^{th}$ bank. Consider the sampling design defined by $\mathcal{S}' = \{\underline{s} = (s_1, \ldots, s_{n(\underline{s})}) : s_i \in \{B_1, \ldots, B_6\}, i = 1, \ldots, n(\underline{s}), s_1 \neq s_2 \neq \cdots \neq s_{n(\underline{s})}\}$ and

$$P(\underline{s}) = \begin{cases} \frac{1}{60}, & \text{if } n(\underline{s}) = 2 \\ \frac{1}{240}, & \text{if } n(\underline{s}) = 3 \\ 0, & \text{otherwise} \end{cases}.$$

   Under the above design, suppose that the selected sample is $(B_2, B_5, B_6)$. Using Horvitz-Thompson estimator and the observed sample, estimate the total number of non-performing accounts in the population. Compute an unbiased estimate of the population variance.

3. Do the Problem 2, under the design $\mathcal{S}' = \{\underline{s} = (s_1, \ldots, s_{n(\underline{s})}) : s_i \in \{B_1, \ldots, B_6\}, i = 1, \ldots, n(\underline{s})\}$,

$$P(\underline{s}) = \begin{cases} \frac{1}{72}, & \text{if } n(\underline{s}) = 2 \\ \frac{1}{432}, & \text{if } n(\underline{s}) = 3 \\ 0, & \text{otherwise} \end{cases}.$$

   and if the observed sample is $(B_2, B_3, B_2)$.

4. Let $\mathcal{U} = (U_1, \ldots, U_N)$ and $\mathcal{Y} = (Y_1, \ldots, Y_N)$. A SRSWOR sample of size $n$ is drawn from $\mathcal{U}$ and subsequently a SRSWOR subsample of size $n^*$ is drawn from this sample. Let $\hat{\overline{Y}}^{(2)}$ denote the sample mean based on the combined sample of size $n+n^*$ and $\hat{\overline{Y}}$ is the sample mean based on all $n$ units in original sample. Show that

   (i) Show that $\hat{\overline{Y}}^{(2)}$ is an unbiased estimator of the population mean.

   (ii) Let $V_1$ and $V_2$, respectively, be the variances of $\hat{\overline{Y}}$ and $\hat{\overline{Y}}^{(2)}$. Show that

$$\frac{V_2}{V_1} \approx \frac{1 + 3\frac{n^*}{n}}{(1 + \frac{n^*}{n})^2}.$$

5. Let $\mathcal{Y} = ((X_1, Y_1), \ldots, (X_N, Y_N))$, $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $C_{\underline{X},\underline{Y}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$. Under the $SRSWR(n)$ design, let $\hat{\overline{X}}^*$ and $\hat{\overline{Y}}^*$ denote the sample means of $X$ and $Y$ variables, respectively, based on whole sample. Let $\hat{C}_{\underline{X},\underline{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{(i)} - \hat{\overline{X}}^*)(Y_{(i)} - \hat{\overline{Y}}^*)$, where $(X_{(i)}, Y_{(i)})$ is the $i^{th}$ sample unit. Find

   (i) the correlation coefficient between $\hat{\overline{X}}^*$ and $\hat{\overline{Y}}^*$.

   (ii) $E(\hat{C}_{\underline{X},\underline{Y}})$.

6. Let $\mathcal{U} = (U_1, \ldots, U_N)$ and $\mathcal{Y} = (Y_1, \ldots, Y_N)$. Under *SRSWR(n)*, suppose that the coefficient of variance $C = \frac{S}{\overline{Y}}$ is known. Find an estimator of the type $\hat{\overline{Y}}_k = k\hat{\overline{Y}}^*$ for the population mean that has the smaller mean squared error than $\hat{\overline{Y}}^*$; here $\hat{\overline{Y}}^*$ denotes the sample mean based on full sample. Find the relative efficiency (ratio of mean squared errors) of this estimator relative to $\hat{\overline{Y}}^*$.

7. Let $\mathcal{U} = (U_1, \ldots, U_n)$ and $\mathcal{Y} = (Y_1, \ldots, Y_N)$. Suppose that we want to estimate the population proportion $P$ of units having a given attribute.

   (i) Find unbiased estimators of $P$ under the *SRSWR(n)* and *SRSWOR(n)*.

   (ii) Compute variances of estimators derived in (i).

   (iii) Construct a 95% confidence interval for $P$ under *SRSWOR(n)* and *SRSWOR(n)*.

8. Let $\mathcal{U} = (U_1, \ldots, U_n)$ and $\mathcal{Y} = (Y_1, \ldots, Y_N)$. Suppose that it is desired to estimate the population proportion $P$ of units having a rare attribute. Consider a *SRSWOR* scheme that is continued until $m$ units possessing the rare attribute have been found. Let $M$ be the total sample size of the sample. If *fpc* is ignored, show that

$$\Pr(M = n) = \binom{n-1}{m-1} P^m (1-P)^{n-m}, \quad n = m, m+1, \ldots.$$

   Find $E[M]$ and show that $\frac{m-1}{M-1}$ is an unbiased estimator of $P$.

9. Under *SRSWR(n)* note that sample mean based on the whole sample is given by

$$\hat{\overline{Y}}^{*} = \frac{1}{n}\sum_{i=1}^{N} A_i Y_i,$$

where $A_i$ = number of times $i$th unit appears in the sample. Using the above representation. Find the mean and the variance of $\hat{\overline{Y}}^{*}$.

10. Consider $\mathcal{U} = (U_1, U_2, U_3, U_4)$ and $\mathcal{Y} = (Y_1, Y_2, Y_3, Y_4) = (1, 6, 6, 11)$.

    (i) Find the probability distribution of the sample mean under *SRSWR(2)*. Find its mean and variance.

    (ii) Repeat (i) under *SRSWOR(2)*.

    (iii) Find the probability distribution of Horvitz-Thompson Estimator under *SRSWR(2)*. Find its mean and variance.

    (iv) Compare performances of different estimators described above.

11. Under *SRSWR(n)*, let $D$ denote the number of distinct unit in the sample. Let $\hat{\overline{Y}}^{*}$ denote the sample mean based on the whole sample and $\hat{\overline{Y}}^{**}$ be the sample mean based on distinct units.

    (i) Find the probability mass function of $D$.

    (ii) Find $E(D)$ and $Var(D)$.

    (iii) Show that $E\left(\frac{1}{D}\right) \geqslant \frac{1}{N\left(1-\left(1-\frac{1}{N}\right)^n\right)}$.

    (iv) Show that $\hat{\overline{Y}}^{*}$ and $\hat{\overline{Y}}^{**}$ are unbiased estimator of $\overline{Y}$. Compare their precisions.