STAT 521: Homework Assignment 1 - Solution

Problem 1:

Consider the following sampling design from a finite population $U = \{1, 2, 3\}$. Let y_i be the study item of interest in unit i in the population. We are interested in estimating the population total of y.

Sample (A)	$Pr\,(A)$	HT estimator	HT var. est.	SYG var. est.
$A_1 = \{1, 2\}$	0.5			
$A_2 = \{1, 3\}$	0.25			
$A_3 = \{2, 3\}$	0.25			

1. Compute the HT estimators and the two variance estimators for each sample. Check the unbiasedness of the variance estimators. (May assume $y_1 = 3, y_2 = 6, y_3 = 2$ here only.)

Solution: Note that

$$\pi_1 = 0.5 + 0.25 = 0.75,\ \pi_2 = 0.5 + 0.25 = 0.75,\ \pi_3 = 0.25 + 0.25 = 0.5$$

HT estimator

$$\hat{Y}_{\text{HT}} = \begin{cases} y_1/\pi_1 + y_2/\pi_2 = 3/0.75 + 6/0.75 = 12 & \text{for } A = \{1, 2\} \\ y_1/\pi_1 + y_3/\pi_3 = 3/0.75 + 2/0.5 = 8 & \text{for } A = \{1, 3\} \\ y_2/\pi_2 + y_3/\pi_3 = 6/0.75 + 2/0.5 = 12 & \text{for } A = \{2, 3\} \end{cases}$$

HT variance estimator (for $A = \{1, 2\}$):

$$\begin{aligned} \hat{V}_{\text{HT}} &= \sum_{i \in A} \frac{y_i^2}{\pi_i^2}(1 - \pi_i) + \sum_{i \neq j} \frac{y_i y_j}{\pi_i \pi_j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \\ &= \frac{y_1^2}{\pi_1^2}(1 - \pi_1) + \frac{y_2^2}{\pi_2^2}(1 - \pi_2) + 2\frac{y_1 y_2}{\pi_1 \pi_2} \frac{\pi_{12} - \pi_1 \pi_2}{\pi_{12}} \\ &= \frac{3^2}{0.75^2}(1 - 0.75) + \frac{6^2}{0.75^2}(1 - 0.75) + 2\frac{3}{0.75}\frac{6}{0.75}\left(1 - \frac{0.75 \cdot 0.75}{0.5}\right) = 12 \end{aligned}$$

SYG variance estimator (for $A = \{1, 2\}$):

$$\begin{aligned} \hat{V}_{\text{SYG}} &= \frac{1}{2} \sum_{i \in A} \sum_{j \in A} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \\ &= \left(\frac{y_1}{\pi_1} - \frac{y_2}{\pi_2}\right)^2 \frac{\pi_1 \pi_2 - \pi_{12}}{\pi_{12}} \\ &= \left(\frac{3}{0.75} - \frac{6}{0.75}\right)^2 \left(\frac{0.75 \cdot 0.75}{0.5} - 1\right) = 2 \end{aligned}$$

Solution: Therefore, we obtain

Sample (A)	$Pr\,(A)$	HT estimator	HT var. est.	SYG var. est.
$A_1 = \{1, 2\}$	0.5	12	12	2
$A_2 = \{1, 3\}$	0.25	8	-4	0
$A_3 = \{2, 3\}$	0.25	12	-8	8

Note that

$$\begin{aligned} E\left(\hat{Y}_{\mathrm{HT}}\right) &= 0.5 * 12 + 0.25 * 8 + 0.25 * 12 = 11 \\ V\left(\hat{Y}_{\mathrm{HT}}\right) &= 0.5 * (12 - 11)^2 + 0.25 * (8 - 11)^2 + 0.25 * (12 - 11)^2 \\ &= 3. \end{aligned}$$

Now, we can check

$$\begin{aligned} E\left(\hat{V}_{\mathrm{HT}}\right) &= 12 * 0.5 - 4 * 0.25 - 8 * 0.25 = 3 \\ E\left(\hat{V}_{\mathrm{SYG}}\right) &= 0.5 * 2 + 0.25 * 8 = 3. \end{aligned}$$

Thus, both variance estimators are unbiased.

2. Now, consider the special case of $y_k = \pi_k$, where π_k is the first order inclusion probability of unit k. What is the variance of the HT estimator ?

Solution: HT estimator (under $y_i = \pi_i$): $\hat{Y}_{HT} = \sum_{i \in A} y_i/\pi_i = |A| = 2$ for $A = A_1, A_2, A_3$. Thus, $V(\bar{Y}_{HT}) = 0$.

3. Also, under the case of $y_k = \pi_k$, compute HT variance estimator and SYG variance estimator for each sample. (They are not the same.) Which variance estimator do you prefer ? Why ?

Solution:

- SYG variance estimator

$$\hat{V}_{\mathrm{SYG}} = \frac{1}{2}\sum_{i\in A}\sum_{j\in A}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}} = 0,$$

  as $y_i/\pi_i = 1$ for all $i = 1, 2, 3$.

- HT variance estimator

$$\hat{V}_{\mathrm{HT}} = \sum_{i\in A}\frac{y_i^2}{\pi_i^2}(1-\pi_i) + \sum_{i\neq j}\frac{y_i y_j}{\pi_i\pi_j}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}$$

  For $A = \{1, 2\}$ with $y_i = \pi_i$, we obtain

$$\begin{aligned}\hat{V}_{\mathrm{HT}} &= (1-\pi_1) + (1-\pi_2) + 2\cdot\left(1 - \frac{\pi_1\pi_2}{\pi_{12}}\right)\\ &= 0.25 + 0.25 + 2 * (1 - 0.75 * 0.75/0.5) = 0.25.\end{aligned}$$

  For $A = \{1, 3\}$, we obtain

$$\begin{aligned}\hat{V}_{\mathrm{HT}} &= (1-\pi_1) + (1-\pi_3) + 2\cdot\left(1 - \frac{\pi_1\pi_3}{\pi_{13}}\right)\\ &= 0.25 + 0.5 + 2 * (1 - 0.75 * 0.5/0.25) = -0.25.\end{aligned}$$

  Similarly, we can obtain $\hat{V}_{HT} = -0.25$ for $A = \{2, 3\}$.

  Therefore, we also obtain $E\{\hat{V}_{\mathrm{HT}}\} = 0$, but can take a negative value. Thus, $\hat{V}_{\mathrm{SYG}}$ is preferred.

Problem 2:

Let U be a finite population of size N. We define the following sampling design: we first select a sample A_1 according to a simple random sampling (without replacement) of fixed size n_1. We then select a sample A_2 in U outside of A_1 according to a simple random sampling design without replacement of fixed size n_2. The final sample A consists of A_1 and A_2.

1. What is the sampling distribution of A? What is interesting about this result?

Solution:

Note that

$$P(A_1) = \begin{cases} \binom{N}{n_1}^{-1} & \text{if } |A_1| = n_1 \\ 0 & \text{otherwise} \end{cases}$$

and, for given A_1 with $n_1 = |A_1|$,

$$P(A_2 \mid A_1) = \begin{cases} \binom{N-n_1}{n_2}^{-1} & \text{if } |A_2| = n_2 \\ 0 & \text{otherwise.} \end{cases}$$

Now, for given n_1 and n_2,

$$\begin{aligned} P(A) &= P(A_1 \cup A_2; |A_1| = n_1, |A_2| = n_2) \\ &= \sum_{\{A_1; |A_1| = n_1\}} P(A_1) P(A_2 \mid A_1) I(|A_2| = n_2) \\ &= \frac{\binom{n_1+n_2}{n_1}}{\binom{N}{n_1}\binom{N-n_1}{n_2}} I(|A_1| = n_1) I(|A_2| = n_2) \\ &= \frac{1}{\binom{N}{n_1+n_2}} I(|A| = n_1 + n_2). \end{aligned}$$

Thus, it is a SRS of size $n = n_1 + n_2$ from N.

2. We define the estimator of $\bar{Y}$, the finite population mean of y, by

$$\bar{y}_\alpha = \alpha \bar{y}_1 + (1 - \alpha)\bar{y}_2$$

with $0 < \alpha < 1$, where $\bar{y}_1$ is the sample mean of y in A_1 and $\bar{y}_2$ is the sample mean of y in A_2. Show that $\bar{y}_\alpha$ is unbiased for $\bar{Y}$ for any α.

Solution:

Since the sampling design for A_1 is a SRS of size n_1, we have $E(\bar{y}_1) = \bar{Y}_N$. Also, the conditional sampling design for A_2 is a SRS of size n_2 from A_1^c, we have $E(\bar{y}_2 \mid A_1) = (N-n_1)^{-1} \sum_{i \in A_1^c} y_i = (N-n_1)^{-1}(N\bar{Y}_N - n_1\bar{y}_1)$. Thus,

$$E(\bar{y}_2) = E\{E(\bar{y}_2 \mid A_1)\} = (N-n_1)^{-1}\{N\bar{Y}_N - n_1 E(\bar{y}_1)\} = \bar{Y}_N.$$

Therefore, since both $\bar{y}_1$ and $\bar{y}_2$ are unbiased for $\bar{Y}_N$, $\bar{y}_\alpha = \alpha\bar{y}_1 + (1-\alpha)\bar{y}_2$ is also unbiased for $\bar{Y}_N$.

3. Find the optimal value of α that minimizes the variance of $\bar{y}_\alpha$.

Solution: Since

$$V(\bar{y}_\alpha) = \alpha^2 V(\bar{y}_1) + (1-\alpha)^2 V(\bar{y}_2) + 2\alpha(1-\alpha)Cov(\bar{y}_1, \bar{y}_2),$$

it is minimized at

$$\alpha^* = \frac{V(\bar{y}_2) - Cov(\bar{y}_1, \bar{y}_2)}{V(\bar{y}_1) + V(\bar{y}_2) - 2Cov(\bar{y}_1, \bar{y}_2)}.$$

Now, we have

$$\begin{aligned}
V(\bar{y}_1) &= \left(\frac{1}{n_1} - \frac{1}{N}\right) S^2 \\
V(\bar{y}_2) &= V\{E(\bar{y}_2 \mid A_1)\} + E\{V(\bar{y}_2 \mid A_1)\} \\
&= V\left(\bar{Y}_{1c}\right) + E\left[\{n_2^{-1} - (N-n_1)^{-1}\} S_{1c}^2\right] \\
&= \{(N-n_1)^{-1} - N^{-1}\} S^2 + \{n_2^{-1} - (N-n_1)^{-1}\} S^2 \\
&= \left(n_2^{-1} - N^{-1}\right) S^2
\end{aligned}$$

where $S_{1c}^2 = (N-n_1-1)^{-1} \sum_{i \in A_1^c} \left(y_i - \bar{Y}_{1c}\right)^2$ and $\bar{Y}_{1c} = (N-n_1)^{-1} \sum_{i \in A_1^c} y_i$. Also,

$$\begin{aligned}
Cov(\bar{y}_1, \bar{y}_2) &= Cov\{\bar{y}_1, E(\bar{y}_2 \mid A_1)\} + E\{Cov(\bar{y}_1, \bar{y}_2 \mid A_1)\} \\
&= Cov(\bar{y}_1, \bar{Y}_{1c}) \\
&= -\left(\frac{n_1}{N-n_1}\right) V(\bar{y}_1) \\
&= -\left(\frac{n_1}{N-n_1}\right)\left(\frac{1}{n_1} - \frac{1}{N}\right) S^2 = -\frac{1}{N} S^2.
\end{aligned}$$

Therefore, $\alpha^* = \frac{n_1}{n_1+n_2}$.

Problem 3:

A community in the San Francisco Bay area consists of approximately 100,000 persons, of whom approximately 40% are Caucasians, 25% are African American, 20% are Hispanic, and 15% are Asian. It is desired to estimate in this community, the proportion of persons who are not covered by some form of health insurance. One would like to be 95% certain that this estimate is within 15% of the true proportion, which is believed to lie somewhere between 10% and 20% of the total population. Assuming simple random sampling, how large a sample is needed?

Solution: We wish to achieve

$$P\left(\frac{\left|\hat{P}-P\right|}{P} \leq 0.15\right) = 0.95.$$

The 95% C.I. for P is

$$P\left(\frac{\left|\hat{P}-P\right|}{\sqrt{n^{-1}(1-n/N)P(1-P)}} \leq 1.96\right) = 0.95.$$

Thus, we can solve

$$0.15P \geq 1.96\sqrt{\frac{1}{n}(1-\frac{n}{N})\frac{N}{N-1}P(1-P)}.$$

After some algebra, we obtain

$$\frac{N-1}{N}\left(\frac{0.15P}{1.96}\right)^2 \geq \frac{1}{n}P(1-P) - \frac{1}{N}P(1-P)$$

$$\iff \frac{N-1}{N}\left(\frac{0.15P}{1.96}\right)^2 + \frac{1}{N}P(1-P) \geq \frac{1}{n}P(1-P)$$

$$\iff n \geq \frac{P(1-P)}{0.005857P^2 + 10^{-6}P(1-P)} = \frac{(1-P)}{0.005857P + 10^{-6}(1-P)}$$

We want to find the smallest n satisfying the above inequality for all $P \in (0.1, 0.2)$. The maximum of the right side of the above inequality is achieved when $P = 0.1$ and the minimum sample size is equal to $n = 1,535$.