

# HW1

Samuel Olson

## Problem 1: (15 pt)

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	$49\frac{1}{3}$		
$A_2 = \{1, 3\}$	0.25	$57\frac{1}{3}$		
$A_3 = \{2, 3\}$	0.25	64		

1.

Compute the HT estimators and the two variance estimators for each sample. Check the unbiasedness of the variance estimators. (May assume  $y_1 = 16$ ,  $y_2 = 21$ ,  $y_3 = 18$  here only.)

```
y <- c(16, 21, 18)
samples <- list(c(1,2), c(1,3), c(2,3))

probs <- c(0.5, 0.25, 0.25)
pi <- c(0.75, 0.75, 0.5)

pi_12 <- 0.5
pi_13 <- 0.25
pi_23 <- 0.25
pi_ij <- matrix(c(NA, pi_12, pi_13, pi_12, NA, pi_23, pi_13, pi_23, NA), nrow=3, byrow=TRUE)

ht_estimators <- sapply(X = samples,
                        FUN = function(A) sum(y[A] / pi[A])
                        )

names(ht_estimators) <- c("A_1", "A_2", "A_3")
round(ht_estimators, 2)
```

```
##   A_1   A_2   A_3
## 49.33 57.33 64.00
```

First-order inclusion probabilities:

$$\pi_1 = 0.5 + 0.25 = 0.75$$

$$\pi_2 = 0.5 + 0.25 = 0.75$$

$$\pi_3 = 0.25 + 0.25 = 0.5$$

**HT Estimate:**

$$\hat{Y}_{HT,A_1} = \sum_{i \in A} \frac{y_i}{\pi_i} = (16/0.75) + (21/0.75) = 49\frac{1}{3}$$

$$\hat{Y}_{HT,A_2} = \sum_{i \in A} \frac{y_i}{\pi_i} = (16/0.75) + (18/0.5) = 57\frac{1}{3}$$

$$\hat{Y}_{HT,A_3} = \sum_{i \in A} \frac{y_i}{\pi_i} = (21/0.75) + (18/0.5) = 64$$

**HT Variance**

$$\pi_{12} = 0.5$$

$$\pi_{13} = 0.25$$

$$\pi_{23} = 0.25$$

$$E[\hat{Y}_{HT}] = \pi_{ij} \hat{Y}_{HT} = (0.5)(49\frac{1}{3}) + (0.25)(57\frac{1}{3}) + (0.25)(64) = 55$$

$$\text{Var}(\hat{Y}_{HT}) = (0.5)(49\frac{1}{3} - 55)^2 + (0.25)(57\frac{1}{3} - 55)^2 + (0.25)(64 - 55)^2 = (0.5)(32.15) + (0.25)(5.43) + (0.25)(81) = 37\frac{2}{3}$$

```
# true mean
E_YHT <- sum(probs * ht_estimators)
E_YHT
```

```
## [1] 55
```

```
# true variance
var_YHT <- sum(probs * (ht_estimators - E_YHT)^2)
round(var_YHT,2)
```

```
## [1] 37.67
```

```

# HT variance
ht_var_estimator <- function(A) {
  sum(sapply(A, function(i) {
    sapply(A, function(j) {
      if (i != j) {
        (pi_ij[i, j] - pi[i] * pi[j]) * (y[i] * y[j]) / (pi[i] * pi[j])
      } else {
        0
      }
    })
  }))) / 2
}

ht_var_estimators <- sapply(X = samples,
                           FUN = ht_var_estimator)
names(ht_var_estimators) <- c("A_1", "A_2", "A_3")
ht_var_estimators

```

```

##           A_1           A_2           A_3
## -37.33333  -96.00000 -126.00000

```

Negative variance is very odd, so explicitly double checking my calculations:

$$\text{Var}(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j}$$

Setup:

$$y_1 = 16, y_2 = 21, y_3 = 18$$

$$\pi_1 = 0.75, \quad \pi_2 = 0.75, \quad \pi_3 = 0.5$$

$$\pi_{12} = 0.5, \quad \pi_{13} = 0.25, \quad \pi_{23} = 0.25$$

A\_1:

$$\text{Var}(\hat{Y}_{HT, A_1}) = (\pi_{12} - \pi_1 \pi_2) \frac{y_1 y_2}{\pi_1 \pi_2} = (0.5 - 0.75(0.75)) \left( \frac{16(21)}{0.75(0.75)} \right) = (-0.0625)(597.33) = -37\frac{1}{3}$$

A\_2:

$$\text{Var}(\hat{Y}_{HT, A_2}) = (\pi_{13} - \pi_1 \pi_3) \frac{y_1 y_3}{\pi_1 \pi_3} = (0.25 - 0.75(0.5)) \left( \frac{16(18)}{0.75(0.5)} \right) = (-0.125)(768) = -96$$

A\_3:

$$\text{Var}(\hat{Y}_{HT, A_3}) = (\pi_{23} - \pi_2 \pi_3) \frac{y_2 y_3}{\pi_2 \pi_3} = (0.25 - 0.75(0.5)) \left( \frac{21(18)}{0.75(0.5)} \right) = (-0.125)(1008) = -126$$

## SYG Variance

```
# SYG Variance
syg_var_estimator <- function(A) {
  sum(sapply(A, function(i) {
    sapply(A, function(j) {
      if (i != j) {
        (pi_ij[i, j] - pi[i] * pi[j]) / (pi_ij[i, j] * pi[i] * pi[j]) * (y[i] * y[j])
      } else {
        0
      }
    })
  }))
}

syg_var_estimators <- sapply(X = samples,
                             FUN = syg_var_estimator)
names(syg_var_estimators) <- c("A_1", "A_2", "A_3")
syg_var_estimators
```

```
##           A_1           A_2           A_3
## -149.3333  -768.0000 -1008.0000
```

Again, odd that we have negative variance estimates. Given this, I'm going to explicitly check/detail the calculations. Similar setup used for the HT variance estimates:

The **SYG variance estimator** is given by:

$$\hat{V}_{SYG} = \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} y_i y_j$$

A\_1:

$$\hat{V}_{SYG}(A_1) = \frac{\pi_{12} - \pi_1 \pi_2}{\pi_{12} \pi_1 \pi_2} y_1 y_2 = \frac{0.5 - (0.75)(0.75)}{0.5(0.75)(0.75)} 16(21) = \left( \frac{-0.0625}{0.28125} \right) (336) = (-0.2222)(336) = -74\frac{2}{3}$$

A\_2:

$$\hat{V}_{SYG}(A_2) = \frac{\pi_{13} - \pi_1 \pi_3}{\pi_{13} \pi_1 \pi_3} y_1 y_3 = \frac{0.25 - (0.75)(0.5)}{0.25(0.75)(0.5)} 16(18) = \left( \frac{-0.125}{0.09375} \right) (288) = (-1.3333)(288) = -384$$

A\_3:

$$\hat{V}_{SYG}(A_3) = \frac{\pi_{23} - \pi_2 \pi_3}{\pi_{23} \pi_2 \pi_3} y_2 y_3 = \frac{0.25 - (0.75)(0.5)}{0.25(0.75)(0.5)} 21(18) = \left( \frac{-0.125}{0.09375} \right) (378) = (-1.3333)(378) = -504$$

## Unbiasedness

Both variance calculations are negative, so they are certainly different from the true variance, making both estimates biased. Specifically:

From previous calculations:

$$\text{Var}(\hat{Y}_{HT}) = 37\frac{2}{3}$$

Bias of HT variance estimator:

$$E[\hat{V}_{HT}] = P(A_1)\hat{V}_{HT}(A_1) + P(A_2)\hat{V}_{HT}(A_2) + P(A_3)\hat{V}_{HT}(A_3) = (0.5)(-37\frac{1}{3}) + (0.25)(-96) + (0.25)(-126) = -18.67 - 24 - 31.5$$

$$E[\hat{V}_{HT}] = -148.33 \neq \text{Var}(\hat{Y}_{HT})$$

Making the HT variance estimator as given (with a fixed sample size of 2) biased.

Bias of SYG Variance Estimator:

$$E[\hat{V}_{SYG}] = P(A_1)\hat{V}_{SYG}(A_1) + P(A_2)\hat{V}_{SYG}(A_2) + P(A_3)\hat{V}_{SYG}(A_3) = (0.5)(-74\frac{2}{3}) + (0.25)(-384) + (0.25)(-504) = -37.33 - 96 - 126$$

$$E[\hat{V}_{SYG}] = -259.33 \neq \text{Var}(\hat{Y}_{HT})$$

Making the SYG variance estimator as given (with a fixed sample size of 2) biased.

## 2.

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

In this special case, the variance of the Horvitz-Thompson estimator is zero, i.e.:

$$\text{Var}(\hat{Y}_{HT}) = 0$$

Based on the original formula, this special case simplifies the expression to:

$$\hat{Y}_{HT} = \sum_{i \in A} \frac{\pi_i}{\pi_i} = \sum_{i \in A} 1 = 2$$

as all possible samples consist of a fixed size  $n=2$ .

This also simplifies the expression used for the variance of the HT estimator:

$$\text{Var}(\hat{Y}_{HT}) = E(\hat{Y}_{HT}^2) - \left(E[\hat{Y}_{HT}]\right)^2 = (0.5(2^2)) + (0.25(2^2)) + (0.25(2^2)) - 2^2 = (0.5 + 0.25 + 0.25)(4) - 4 = 0$$

The more descriptive interpretation is in this circumstance there is no variability in the HT estimator as it always takes the same value for its point estimate.

### 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?

General Approach:

$$\text{HT: } \text{Var}(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j=1}^N (\pi_{ij} - \pi_i \pi_j) \frac{y_i y_j}{\pi_i \pi_j} \quad \text{SYG: } \text{Var}(\hat{Y}_{SYG}) = \sum_{i \neq j} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_j} y_i y_j$$

```
# Updated Setup
pi <- c(0.75, 0.75, 0.5)
pi_12 <- 0.5
pi_13 <- 0.25
pi_23 <- 0.25
pi_ij <- matrix(c(NA, pi_12, pi_13,
                  pi_12, NA, pi_23,
                  pi_13, pi_23, NA),
                nrow=3, byrow=TRUE)

samples <- list(c(1,2), c(1,3), c(2,3))

# HT
ht_var_estimator <- function(A) {
  sum(sapply(A, function(i) {
    sapply(A, function(j) {
      if (i != j) {
        # swap all prior y[i]/y[j] with pi[i]/pi[j]
        (pi_ij[i, j] - pi[i] * pi[j]) * (pi[i] * pi[j]) / (pi[i] * pi[j])
      } else {
        0
      }
    })
  }))) / 2
}

# Compute for each sample
ht_var_estimators <- sapply(samples, ht_var_estimator)
names(ht_var_estimators) <- c("A_1", "A_2", "A_3")
ht_var_estimators

##      A_1      A_2      A_3
## -0.0625 -0.1250 -0.1250

# SYG
syg_var_estimator <- function(A) {
  sum(sapply(A, function(i) {
    sapply(A, function(j) {
      if (i != j) {
        # swap all prior y[i]/y[j] with pi[i]/pi[j]
        (pi_ij[i, j] - pi[i] * pi[j]) / (pi_ij[i, j] * pi[i] * pi[j]) * (pi[i] * pi[j])
      } else {
        0
      }
    })
  })))
}
```

```

    )))
  }

syg_var_estimators <- sapply(samples, syg_var_estimator)
names(syg_var_estimators) <- c("A_1", "A_2", "A_3")
syg_var_estimators

```

```

##   A_1   A_2   A_3
## -0.25 -1.00 -1.00

```

HT Variance Estimators:

- Sample  $A_1 = \{1, 2\}$ :  $\hat{V}_{HT}(A_1) = -0.0625$
- Sample  $A_2 = \{1, 3\}$ :  $\hat{V}_{HT}(A_2) = -0.125$
- Sample  $A_3 = \{2, 3\}$ :  $\hat{V}_{HT}(A_3) = -0.125$

SYG Variance Estimators:

- Sample  $A_1 = \{1, 2\}$ :  $\hat{V}_{SYG}(A_1) = -0.125$
- Sample  $A_2 = \{1, 3\}$ :  $\hat{V}_{SYG}(A_2) = -0.5$
- Sample  $A_3 = \{2, 3\}$ :  $\hat{V}_{SYG}(A_3) = -0.5$

Some Thoughts:

Both variance estimates across all samples are still negative, which is problematic and implies that neither is that great for the data provided. That point notwithstanding, the HT Variance Estimator comparatively has greater precision, i.e. smaller variance, across all samples, so is comparatively better than SYV in that sense.

## Problem 2: (15 pt)

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?

This is a two-stage sampling design. Analyzing them stage-by-stage will, I believe, illustrate the point of this question.

First Stage: Select  $A_1$  using simple random sampling without replacement (SRSWOR) of fixed size  $n_1$  from the population  $U$  of size  $N$ . Each subset  $A_1$  of size  $n_1$  has an equal probability of being selected. The probability of selecting a specific sample,  $A_1$  is:

$$P(A_1) = \frac{1}{\binom{N}{n_1}}.$$

Second Stage: Select  $A_2$  from the remaining units  $U \setminus A_1$  using SRSWOR of fixed size  $n_2$ . Given  $A_1$ , each subset  $A_2$  of size  $n_2$  from the remaining  $N - n_1$  units has an equal probability of being selected. The probability of selecting a specific  $A_2$ , given  $A_1$ , is:

$$P(A_2|A_1) = \frac{1}{\binom{N-n_1}{n_2}}.$$

Taken together, the final sample is:

$$A = A_1 \cup A_2$$

Thus,  $A$  consists of exactly  $n = n_1 + n_2$  elements.

Using the law of conditional probability, the probability of selecting a specific final sample  $A = A_1 \cup A_2$  is:

$$P(A) = P(A_1)P(A_2|A_1) = \frac{1}{\binom{N}{n_1}} \left( \frac{1}{\binom{N-n_1}{n_2}} \right)$$

Given:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

And:

$$\binom{N}{n} = \binom{N}{n_1} \binom{N-n_1}{n_2}$$

We may rewrite  $P(A)$  as:



$$P(A) = \frac{1}{\binom{N}{n_1} \binom{N-n_1}{n_2}} = \frac{1}{\binom{N}{n}}$$

The probability of selecting any specific final sample  $A$  does not depend on the intermediate selection of  $A_1$  and  $A_2$ . That is, the two-stage sampling process yields the same probability distribution as direct simple random sampling without replacement of size  $n$ .

This result shows that two-stage sequential SRSWOR is equivalent to single-stage SRSWOR. This is a fundamental property of simple random sampling: Whether we select  $A_1$  first and then  $A_2$ , or select all  $n$  units at once, each sample of size  $n$  has the same probability of being selected.

This property is useful because it allows for a sequential selection procedure without altering the randomness of final sample selection.

The sampling distribution of  $A$  is uniform over all subsets of size  $n$  in  $U$ :

$$P(A) = \frac{1}{\binom{N}{n}}$$

The interesting result is that this two-stage sampling design is equivalent to simple random sampling without replacement of size  $n$ , meaning that the order in which units are selected does not affect the final probability distribution of the sample.

## 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  $\alpha$ .

Our goal is to show:

$$E[\bar{y}_\alpha] = \bar{Y}$$

A Few Key Definitions to note not explicitly included in the beginning of the problem:

(Finite) Population Mean:

$$\bar{Y} = \frac{1}{N} \sum_{i \in U} y_i$$

Sample Means:

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i \in A_1} y_i, \quad \bar{y}_2 = \frac{1}{n_2} \sum_{i \in A_2} y_i$$

Estimator:

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

Via linearity of expectation:

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

Since both  $A_1$  and  $A_2$  are selected using simple random sampling without replacement, their expected sample means are unbiased estimators of the population mean:

$$E[\bar{y}_1] = \bar{Y}, \quad E[\bar{y}_2] = \bar{Y}$$

Substituting these into the expectation equation:

$$E[\bar{y}_\alpha] = \alpha\bar{Y} + (1 - \alpha)\bar{Y} = (\alpha + 1 - \alpha)\bar{Y} = \bar{Y}$$

And hence  $\bar{y}_\alpha$  is unbiased for  $\bar{Y}$  for any  $\alpha$ .

### 3.

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

Hints for (3): Since

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

it is minimized at

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

To find the optimal  $\alpha^*$  that minimizes variance, we use our typical calculus technique, i.e. take the derivative of  $V(\bar{y}_\alpha)$  with respect to  $\alpha$  and set it to zero.

Taking the derivative:

$$\frac{d}{d\alpha} V(\bar{y}_\alpha) = 2\alpha V(\bar{y}_1) - 2(1 - \alpha)V(\bar{y}_2) + 2(1 - 2\alpha)\text{Cov}(\bar{y}_1, \bar{y}_2) = 0$$

Simplifying the expression:

$$\alpha V(\bar{y}_1) - (1 - \alpha)V(\bar{y}_2) + (1 - 2\alpha)\text{Cov}(\bar{y}_1, \bar{y}_2) = 0$$

Further simplifying:

$$\alpha V(\bar{y}_1) + \alpha \text{Cov}(\bar{y}_1, \bar{y}_2) = V(\bar{y}_2) - \text{Cov}(\bar{y}_1, \bar{y}_2)$$

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

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

### Problem 3: (10 pt)

A community in the San Francisco Bay area consists of approximately 100,000 persons. 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. That is, we wish to achieve

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

where  $P$  is the true proportion satisfying  $0.1 \leq P \leq 0.2$ . Assuming simple random sampling, how large a sample is needed?

Given the premise/setup of this problem, we need have:

$$Pr\left(\left|\hat{P} - P\right| \leq 0.15P\right) = 0.95 \rightarrow 1.96(\text{SE}(\hat{P})) \leq 0.15P$$

Noting that 1.96 is the critical value from the standard normal distribution, its selection is based on the standard normal CDF. This follows from the normal approximation, which is justified via the CLT.

Under SRS design, the standard error of  $\hat{P}$  is:

$$\text{SE}(\hat{P}) = \sqrt{\frac{P(1-P)}{n}}$$

Thus we may simplify our prior expression:

$$1.96\sqrt{\frac{P(1-P)}{n}} \leq 0.15P \rightarrow (1.96)^2 \frac{P(1-P)}{n} \leq (0.15P)^2$$

Our goal is to find a suitable  $n$  in this equation. To that end:

$$n \geq \frac{(1.96)^2 P(1-P)}{(0.15P)^2} \rightarrow n \geq \frac{3.8416(1-P)}{0.0225P}$$

Since the true proportion  $P$  is believed to be between 0.1 and 0.2, we compute  $n$  for both extremes and take the largest  $n$  to provide a conservative estimate of the sample size required.

```
# functional form
sample_size <- function(P) {
  (3.8416 * (1 - P)) / (0.0225 * P)
}

# setup
P_values <- seq(0.1, 0.2, length.out = 1000)
sample_sizes <- sapply(P_values, sample_size)

# optimize, maximize
max_n <- max(sample_sizes)
optimal_P <- P_values[which.max(sample_sizes)]
max_n
```

```
## [1] 1536.64
```

```
optimal_P
```

```
## [1] 0.1
```

**P = 0.1:**

$$n \geq \frac{3.8416(1 - 0.1)}{0.0225(0.1)} = \frac{3.8416(0.9)}{0.00225} = \frac{3.4574}{0.00225} = 1536.64 \rightarrow 1537$$

**P = 0.2:**

$$n \geq \frac{3.8416(1 - 0.2)}{0.0225(0.2)} = \frac{3.8416(0.8)}{0.0045} = \frac{3.0733}{0.0045} = 683.95 \rightarrow 684$$

To ensure the margin of error requirement holds for all values of  $P$  in  $[0.1, 0.2]$ , we choose the largest required sample size, 1537 sample size required, rounding up.