

# 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	$49\frac{1}{3}$	160.44	5.56
$A_2 = \{1, 3\}$	0.25	$57\frac{1}{3}$	-6.22	107.56
$A_3 = \{2, 3\}$	0.25	64	-164	32

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 Estimates:**

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

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

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

**HT Variances**

$$\pi_{12} = 0.5$$

$$\pi_{13} = 0.25$$

$$\pi_{23} = 0.25$$

```
y <- c(16, 21, 18)
probs <- c(0.5, 0.25, 0.25)
# first order
pi <- c(0.75, 0.75, 0.50)
# second order
pi_ij <- matrix(c(NA, 0.5, 0.25,
                  0.5, NA, 0.25,
                  0.25, 0.25, NA),
                nrow=3, byrow=TRUE)

# samples
samples <- list(c(1,2), c(1,3), c(2,3))
sample_probs <- c(0.5, 0.25, 0.25)

var_HT <- function(sample) {
  i <- sample[1]
  j <- sample[2]

  # Pairwise term ij = ji
  ij <- 2 * ((pi_ij[i,j] - pi[i] * pi[j]) / pi_ij[i,j]) * (y[i] / pi[i]) * (y[j] / pi[j])
  # Single terms
  i_sq <- ((1 - pi[i]) / pi[i]^2) * y[i]^2
  j_sq <- ((1 - pi[j]) / pi[j]^2) * y[j]^2

  return(i_sq + j_sq + ij)
```

```
}
ht_var <- sapply(samples, var_HT)
ht_var
```

```
## [1] 160.444444 -6.222222 -164.000000
```

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

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

Setup:

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

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

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

$A_1$ :

$$\text{Var}(\hat{Y}_{HT})_{A_1} = \sum_{i \in A} \left( \frac{(1 - \pi_i)}{\pi_i^2} y_i^2 \right) + \sum_{i \neq j} \left( 2 \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} \right) = \left( \frac{0.25}{0.5625} \cdot 256 + \frac{0.25}{0.5625} \cdot 441 \right) + 2 \left( \frac{-0.125}{0.5} \cdot 597.33 \right) = (113.78 + 60.71) - 238.93 = 34.56$$

$A_2$ :

$$\text{Var}(\hat{Y}_{HT})_{A_2} = \sum_i \left( \frac{(1 - \pi_i)}{\pi_i^2} y_i^2 \right) + \sum_{i \neq j} \left( 2 \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} \right) = \left( \frac{0.25}{0.5625} \cdot 256 + \frac{0.50}{0.25} \cdot 324 \right) + 2 \left( \frac{-0.125}{0.25} \cdot 768 \right) = (113.78 + 648) - 1280 = -518.22$$

$A_3$ :

$$\text{Var}(\hat{Y}_{HT})_{A_3} = \sum_i \left( \frac{(1 - \pi_i)}{\pi_i^2} y_i^2 \right) + \sum_{i \neq j} \left( 2 \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{y_i y_j}{\pi_i \pi_j} \right) = \left( \frac{0.25}{0.5625} \cdot 441 + \frac{0.50}{0.25} \cdot 324 \right) + 2 \left( \frac{-0.125}{0.25} \cdot 1008 \right) = (196.00 + 648) - 1280 = -436$$

## SYG Variance

```
# setup done already
# no changes to inclusion probabilities
var_SYG <- function(sample) {
  i <- sample[1]
  j <- sample[2]

  # ij = ji
  # counting twice, so only need negative adjustment
```

```

  est <- (-1) * ((pi_ij[i,j] - pi[i] * pi[j]) / pi_ij[i,j]) * ((y[i] / pi[i]) - (y[j] / pi[j]))^2
  return(est)
}

syg_var <- sapply(samples, var_SYG)
syg_var

```

```
## [1] 5.555556 107.555556 32.000000
```

For thoroughness, checking the calculations explicitly:

By definition:

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

For  $A_1$ :

$$\hat{V}_{SYG}(A_1) = -1 \sum_{i \in A} \sum_{j \in A} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = -1 \left( \frac{0.5 - 0.5625}{0.5} \right) (21.33 - 28)^2 = -1(-0.125)(44.49) = 5.56$$

For  $A_2$ :

$$\hat{V}_{SYG}(A_2) = -1 \sum_{i \in A} \sum_{j \in A} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = -1 \left( \frac{0.25 - 0.375}{0.25} \right) (21.33 - 36)^2 = -1(-0.5)(215.49) = 107.56$$

For  $A_3$ :

$$\hat{V}_{SYG}(A_3) = -1 \sum_{i \in A} \sum_{j \in A} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 = -1 \left( \frac{0.25 - 0.375}{0.25} \right) (28 - 36)^2 = -1(-0.5)(64) = 32$$

## Unbiasedness

```

# Setup
probs <- c(0.5, 0.25, 0.25)

# expected variance
est_var_HT <- sum(probs * ht_var)
est_var_SYG <- sum(probs * syg_var)

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

```

```

E_Y <- sum(probs * ht_estimators)
true_variance <- sum(probs * ((ht_estimators - E_Y)^2))

# bias
bias_HT <- est_var_HT - true_variance
bias_SYG <- est_var_SYG - true_variance

bias_HT

```

```
## [1] 0
```

```
bias_SYG
```

```
## [1] -1.421085e-14
```

Both estimators are unbiased in this example, particularly noting the SYG estimator being basically zero via rounding, likely a precision error of some kind.

Detail in-depth:

The definition of Bias is:

$$\text{Bias}(\hat{V}) = E[\hat{V}] - V_{HT}$$

where  $E[\hat{V}_{HT}]$  and  $E[\hat{V}_{SYG}]$  are the expected values of the variance estimators and  $V_{HT}$  is the true variance of the HT estimator.

Taking our inclusion probabilities as defined previously (first and second order), as well as the observed values for each unit, also note the estimations calculated previously. Specifically, we have:

HT variance estimates:

$$\hat{V}_{HT}(A_1) = 160.44, \quad \hat{V}_{HT}(A_2) = -6.22, \quad \hat{V}_{HT}(A_3) = -164$$

and SYG variance estimates:

$$\hat{V}_{SYG}(A_1) = 5.56, \quad \hat{V}_{SYG}(A_2) = 107.56, \quad \hat{V}_{SYG}(A_3) = 32$$

The expected value of the HT estimator is:

$$E[Y_{HT}] = \sum_A P(A)Y_{HT}(A) = (0.5 \cdot 49.33) + (0.25 \cdot 57.33) + (0.25 \cdot 64) = 24.67 + 14.33 + 16 = 55$$

The true variance is:

$$V_{HT} = \sum_A P(A)(V_{HT}(A) - E[V_{HT}])^2 = (0.5 \cdot (49.33 - 55)^2) + (0.25 \cdot (57.33 - 55)^2) + (0.25 \cdot (64 - 55)^2) = (0.5 \cdot 32.14) + (0.25 \cdot 5.43) + (0.25 \cdot 81) = 37.68$$

The expected value of the HT variance estimator is:

$$E[\hat{V}_{HT}] = \sum_A P(A)\hat{V}_{HT}(A) = (0.5(160.44)) + (0.25(-6.22)) + (0.25(-164)) = 37.68$$

Similarly, for the expectation for the SYG variance estimator is:

$$E[\hat{V}_{SYG}] = \sum_A P(A) \hat{V}_{SYG}(A) = (0.5(5.56)) + (0.25(107.56)) + (0.25(32)) = 37.68$$

The bias of the estimators is then, noting the initial definition given:

For HT:

$$\text{Bias}(\hat{V}_{HT}) = E[\hat{V}_{HT}] - V(T_{HT}) = 37.68 - 37.68 = 0$$

For SYG:

$$\text{Bias}(\hat{V}_{SYG}) = E[\hat{V}_{SYG}] - V(T_{HT}) = 37.68 - 37.68 = 0$$

Both estimators are unbiased, as bias values are both equal to zero (or are nearly zero in the case of SVG).

## 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?

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

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

Under the special case,  $y_k = \pi_k$ , we may simplify:

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

For a fixed-size with size  $n = 2$ :

$$\text{Var}(\hat{Y}_{HT}) = \sum_{i=1}^N \sum_{j=1}^N \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) = 0$$

(As the numerator of the expression given sums to zero via construction/definition.)

Another way to think of this special case:

$$y_k = \pi_k \rightarrow \hat{Y}_{HT}(A_1) = \hat{Y}_{HT}(A_2) = \hat{Y}_{HT}(A_3) = \frac{y_i}{\pi_i} + \frac{y_j}{\pi_j} = 2, \forall i, j \in A$$

So there is no variability in the HT estimates.

So, for this special case the variance of the Horvitz-Thompson estimator is zero:

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

The more descriptive interpretation is in this special case there is no variability in the HT estimator as it always takes the same value regardless of the sample in this event.

### 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?

By definition, the formulae are as follows:

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

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

I've just modified the prior R code to adjust for this special case.

```
var_HT_sc <- function(sample) {
  i <- sample[1]
  j <- sample[2]

  # Pairwise term ij = ji
  ij <- 2 * ((pi_ij[i,j] - pi[i] * pi[j]) / pi_ij[i,j]) * (pi[i] / pi[i]) * (pi[j] / pi[j])
  # Single terms
  i_sq <- ((1 - pi[i]) / pi[i]^2) * pi[i]^2
  j_sq <- ((1 - pi[j]) / pi[j]^2) * pi[j]^2

  return(i_sq + j_sq + ij)
}

ht_var_sc <- sapply(samples, var_HT_sc)
ht_var_sc
```

```
## [1] 0.25 -0.25 -0.25
```

```
var_SYG_sc <- function(sample) {
  i <- sample[1]
  j <- sample[2]

  # ij = ji
  # counting twice, so only need negative adjustment
  est <- (-1) * ((pi_ij[i,j] - pi[i] * pi[j]) / pi_ij[i,j]) * ((pi[i] / pi[i]) - (pi[j] / pi[j]))^2
  return(est)
}

syg_var_sc <- sapply(samples, var_SYG_sc)
syg_var_sc
```

```
## [1] 0 0 0
```

Based on the above R output, we have the following:

HT Variance Estimators:

- $A_1 = \{1, 2\}$ :  $\hat{Y}_{HT}(A_1) = 0.25$
- $A_2 = \{1, 3\}$ :  $\hat{Y}_{HT}(A_2) = -0.25$
- $A_3 = \{2, 3\}$ :  $\hat{Y}_{HT}(A_3) = -0.25$

SYG Variance Estimators:

- $A_1 = \{1, 2\}$ :  $\hat{V}_{SYG}(A_1) = 0$
- $A_2 = \{1, 3\}$ :  $\hat{V}_{SYG}(A_2) = 0$
- $A_3 = \{2, 3\}$ :  $\hat{V}_{SYG}(A_3) = 0$

The two variance estimators are different. Generally, as we want greater precision, we want the smaller variance. In this case, under this criteria, we would prefer the SYG Variance Estimator as it is zero. However, there are concerns about the validity of such an estimator, as one would/should expect some amount of variation, and possibly make a case for using the HT variance estimator instead!



## 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:

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

By definition:

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

And, a relevant property of combinatorics:

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

We may rewrite  $P(A)$  as:

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

Interpretation:

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

So the probability of a sample being selected is the same regardless of whether we select  $A_1$  first and then  $A_2$ , or select all  $n$  units at once.

Also, the sampling distribution of  $A$  is uniform over all subsets of size  $n$  in the (finite) population  $U$ :

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

So a two-stage sampling design is functionally equivalent to SRS without replacement.

## 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  $\alpha$ .

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:

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

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:

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

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

Leaving us with the  $\alpha^*$  that maximizes our initial equation as:

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

Evaluating this expression explicitly:

By definition:

$$V(\bar{y}_1) = \frac{S^2}{n_1} \left(1 - \frac{n_1}{N}\right)$$

$$V(\bar{y}_2) = \frac{S^2}{n_2} \left(1 - \frac{n_2}{N}\right)$$

where:

$$S^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2$$

Also, we have the covariance (under SRS without replacement):

$$\text{Cov}(\bar{y}_1, \bar{y}_2) = -\frac{S^2}{N} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} = -\frac{S^2}{N}$$

Returning then to the maximization quantity for  $\alpha^*$ :

$$\alpha^* = \frac{V(\bar{y}_2) - \text{Cov}(\bar{y}_1, \bar{y}_2)}{V(\bar{y}_1) + V(\bar{y}_2) - 2\text{Cov}(\bar{y}_1, \bar{y}_2)} = \frac{\frac{S^2}{n_2} \left( 1 - \frac{n_2}{N} \right) + \frac{S^2}{N}}{\frac{S^2}{n_1} \left( 1 - \frac{n_1}{N} \right) + \frac{S^2}{n_2} \left( 1 - \frac{n_2}{N} \right) + 2\frac{S^2}{N}}$$

After much simplifications (and approximating for large N to aid in simplification):

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

### 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.