STAT 521 - Homework #1

Sam Tyner

1/30/2018

Problem 1: Identifying sample survey terms

A survey is conducted to find the average weight of cows in a region. A list of all farms is available for the region, and 50 farms are selected at random. Then the weight of each cow at the 50 selected farms is recorded.

1. *What is the target population?*
2. *What is the element?*
3. *What is the sampling unit?*
4. *What is the frame?*
5. *List two possible sources of nonsampling errors.*

Solution:

1. The target population is all cows in the region.
2. The element is each individual cow.
3. The sampling unit is a farm.
4. The frame is the list of all farms in the region.
5. Two sources of nonsampling errors:
    a. Undercoverage: the frame might not be complete. There could be families in the region with a small amount of land and a few cows that are not included in the list of all farms in the region.
    b. Measurement error: cows might not stand completely still on the scale, leading to the recorder to have to make a best guess of true weight when the number on the scale keeps moving up and down.[1]

Problem 2

For a fixed sample size design (i.e. n is a fixed number), prove that the variance of the HT estimator can alternatively be written as

$$V(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2,$$

and its unbiased estimator is

$$\tilde{V}(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{k\in S}\sum_{l\in S}\frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_k}{\pi_k}-\frac{y_l}{\pi_l}\right)^2,$$

provided that all $\pi_{kl} > 0$.

Solution:

Let $\check{y}_k = \dfrac{y_k}{\pi_k}$ for $k = 1, \dots, n$, where n is a fixed sample size. We will show that

$$V(\hat{Y}_{HT}) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\check{y}_k\check{y}_l = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\left(\check{y}_k-\check{y}_l\right)^2.$$

[1] I'm assuming that's how cows are weighed, but I'm from a suburb so I have no idea how they are actually weighed.

Proof.

$$\begin{aligned}
V(\hat{Y}_{HT}) &= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\left(\check{y}_k - \check{y}_l\right)^2 && (1)\\
&= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\left(\check{y}_k^2 - 2\check{y}_k\check{y}_l + \check{y}_l^2\right) && (2)\\
&= -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\check{y}_k^2 + \frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}2\check{y}_k\check{y}_l - \frac{1}{2}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\check{y}_l^2 && (3)\\
&= \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\check{y}_k\check{y}_l - \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\check{y}_k^2 && (4)
\end{aligned}$$

We will now show that the second term above, $\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\check{y}_k^2$, is equal to 0, and thus the proof is complete.

$$\begin{aligned}
\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\check{y}_k^2 &= \sum_{k\in U}\check{y}_k^2\sum_{l\in U}\Delta_{kl} && (5)\\
&= \sum_{k\in U}\check{y}_k^2\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l) && (6)\\
&= \sum_{k\in U}\check{y}_k^2\left[\sum_{l\in U}\pi_{kl} - \sum_{l\in U}\pi_k\pi_l\right] && (7)\\
&= \sum_{k\in U}\check{y}_k^2\left[\sum_{l\in U}\pi_{kl} - \pi_k\sum_{l\in U}\pi_l\right] && (8)\\
&= \sum_{k\in U}\check{y}_k^2\left[\sum_{l\in U, k\neq l}\pi_{kl} + \pi_{kk} - \pi_k n\right] \quad (\textit{n fixed, definition of } E[n_s]) && (9)\\
&= \sum_{k\in U}\check{y}_k^2\left[\pi_k n - \pi_k + \pi_k - \pi_k n\right] \quad (\textit{definition } \pi_{kk}\textit{, result 2.6.2 in SSW}) && (10)\\
&= 0. && (11)
\end{aligned}$$

□

Next, let $\check{\Delta}_{kl} = \dfrac{\Delta_{kl}}{\pi_{kl}}$. To show that the unbiased estimator of $V(\hat{Y}_{HT})$, $\tilde{V}(\hat{Y}_{HT})$, can be written as

$$\tilde{V}(\hat{Y}_{HT}) = \sum_{k\in U}\sum_{l\in U}\check{\Delta}_{kl}\check{y}_k\check{y}_l = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U}\check{\Delta}_{kl}\left(\check{y}_k - \check{y}_l\right)^2,$$

simply substitute $\check{\Delta}_{kl}$ for Δ_{kl} in the previous proof.

Problem 3

Consider a population of size $N = 3$, $U = \{1, 2, 3\}$. *Suppose* $y_1 = 5, y_2 = 3$, and $y_3 = 7$. *Consider the sampling scheme given in the table below.*

Sample	A	P(A)
1	{1,2}	0.4
2	{1,3}	0.3
3	{2,3}	0.2
4	{1,2,3}	0.1

1. *Compute the first order inclusion probability, π_k, for each element, k.*
2. *Compute the second order inclusion probabilities, $\pi_{k\ell}$ for each pair, k, ℓ such that $k \neq \ell$.*
3. *What is the expected sample size?*
4. *What is the population total, t_y for this population*
5. *Compute the HT estimator of t_y for each of the four samples in the table above. Then, verify that $\hat{t}_{HT}$ is an unbiased estimator of t_y.*
6. *Compute the HT variance estimator for each of the four samples in the table above. Then, verify that $\hat{V}(\hat{t}_{HT})$ is an unbiased estimator of $V(\hat{t}_{HT})$.*
7. *Now consider estimating the proportion of the population with a value greater than or equal to 5. What is the population proportion? Using the sample in the first row of the table above, give the HT estimator of the proportion and a corresponding standard error.*
8. *Do you notice anything peculiar about the HT estimate for sample 4? Why/why not? For this design, can you define a different estimator of the total of y that is unbiased for t_y and has a smaller variance than the HT estimator.*

Solution:

1. $\pi_1 = 0.4 + 0.3 + 0.1 = 0.8$; $\pi_2 = 0.4 + 0.2 + 0.1 = 0.7$; $\pi_3 = 0.3 + 0.2 + 0.1 = 0.6$.
2. $\pi_{12} = 0.4 + 0.1 = 0.5$; $\pi_{13} = 0.3 + 0.1 = 0.4$; $\pi_{23} = 0.2 + 0.1 = 0.3$.
3. The expected sample size is: $\sum_{k \in U} \pi_k = 0.8 + 0.7 + 0.6 = 2.1$.
4. The population total is: $t_y = y_1 + y_2 + y_3 = 15$.
5. See table below. $E[\hat{t}_{TH}] = \sum_A P(A) \cdot \hat{t}_{TH} = 0.4 \cdot 10.536 + 0.3 \cdot 17.917 + 0.2 \cdot 15.952 + 0.1 \cdot 22.202 = 15 = t_y$, thus $\hat{t}_{TH}$ is an unbiased estimator for t_y.

A	$\hat{t}_{HT}$
{1,2}	$\frac{5}{.8} + \frac{3}{.7} = 10.536$
{1,3}	$\frac{5}{.8} + \frac{7}{.6} = 17.917$
{2,3}	$\frac{3}{.7} + \frac{7}{.6} = 15.952$
{1,2,3}	$\frac{5}{.8} + \frac{3}{.7} + \frac{7}{.6} = 22.202$

6. First, I compute the true value of the variance of $\hat{t}_{HT}$, $V(\hat{t}_{HT})$:

```r
library(tidyverse)
# data, inclusion probs
y <- c(5, 3, 7)
pi_y <- c(0.8, 0.7, 0.6)
# pairs, 2-way inclusion probs
pi_kl <- data_frame(k = c(1, 1, 2, 1, 2, 3, 2, 3, 3), l = c(2, 3, 3, 1, 2, 3,
    1, 1, 2), pi_kl = c(0.5, 0.4, 0.3, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3))
# Compute V(t_ht) for U = {1,2,3}
Delta <- matrix(0, nrow = 3, ncol = 3)
for (i in 1:3) {
    for (j in 1:3) {
        Delta[i, j] <- filter(pi_kl, k == i, l == j)$pi_kl - pi_y[i] * pi_y[j]
    }
}
Check_y <- (y/pi_y) %*% t((y/pi_y))
v_ht <- sum(colSums(Delta * Check_y))
v_ht
```

```
## [1] 15.89286
```

Next, I compute the value of the estimator, $\hat{V}(\hat{t}_{HT})$, for each of the four samples.

```
# samples, pr(A)s
samps <- data_frame(A = list(c(1, 2), c(1, 3), c(2, 3), c(1, 2, 3)), PrA = c(0.4,
    0.3, 0.2, 0.1))
# Compute estimator for each
v_ht_hat <- function(samp, dat = y, piy = pi_y, pikl = pi_kl) {
    n <- length(samp)
    Check_delta <- matrix(0, nrow = n, ncol = n)
    for (i in 1:n) {
        for (j in 1:n) {
            pkl <- filter(pikl, k == i, l == j)$pi_kl
            Check_delta[i, j] <- (pkl - piy[i] * piy[j])/pkl
        }
    }
    y_s <- dat[samp]
    piy_s <- piy[samp]
    check_y <- (y_s/piy_s) %*% t((y_s/piy_s))
    est_v_ht <- sum(colSums(Check_delta * check_y))
    return(est_v_ht)
}
samps2 <- samps %>% mutate(V_ht_est = map_dbl(A, v_ht_hat), EV = PrA * V_ht_est)
samps2$EV %>% sum
```

```
## [1] 17.81995
```

7. Let $U_d = \{i : i \geq 5\}$, $U_d \subset U$. Then, $p_U = \frac{1}{3}\sum_{i=1}^{3} \mathbb{I}(i \in U_d) = \frac{2}{3}$.

8. It is peculiar that we've taken a census, and thus have all possible information about t_y, but our estimate is nearly 50% greater than the true value.

Problem 4

Consider a simple random sample without replacement S of size n from a finite population U of size N, and two distinct individuals k and l.

1. *Compute $Pr(k \in S, l \notin S)$.*
2. *Suppose we select a further subsample S_1 from S by simple random sampling of size n_1. Let S_2 be the complementary sample of S_1 in S. Thus $S_1 \cup S_2 = S$ and $S_1 \cap S_2 = \emptyset$. Let k and l be any two distinct individuals belonging to the sample S. Compute $Pr(k \in S_1 and l \in S_2)$.*
3. *Let $I_{k1} = \mathbb{I}(k \in S_1)$ and $I_{k2} = \mathbb{I}(k \in S_2)$, where $\mathbb{I}(A)$ represents the indicator for event A. Compute $Cov(I_{k1}, I_{k2})$.*
4. *Let $\bar{y}_i$ be the simple mean of y in the sample S_i. Also, let $\bar{y}$ be the sample mean of y in the sample S. Thus, $n\bar{y} = n_1\bar{y}_1 + n_2\bar{y}_2$ with $n_2 = n - n_1$. Prove that $Cov(\bar{y}_1, \bar{y}_2) = -\frac{S_y^2}{N}$ where $S_y^2 = \sum_{i=1}^{N} \frac{(y_i - \bar{Y}_N)^2}{N-1}$.*
5. *Under the above setup, compute $Cov(\bar{y}, \bar{y}_1 - \bar{y})$ and the mean and variance of $\bar{y}_1$.*

Solution:

1. $Pr(k \in S, l \notin S) = \dfrac{\binom{N-2}{n-1}}{\binom{N}{n}} = \dfrac{n(N-n)}{N(N-1)}$.