# STA 522 HW6 (Multi-Stage Sampling)

*Daniel Truver*

*2/21/2018*

### (1) Lohr, Chapter 6 Problem 4

We begin with showing $\hat{t}_\psi$ is unbiased.

$$E(\hat{t}_\psi) = E\left(\frac{t_i}{\psi_i}\right) = \sum_i \left(\frac{t_i}{\psi_i}\right)\psi_i = \sum_i t_i = t.$$

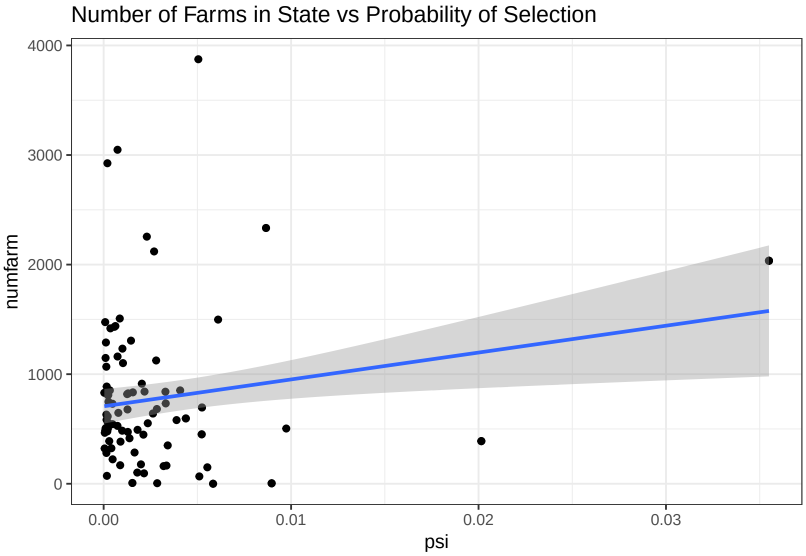
The variacne of $\hat{t}_\psi$ is given by the following.

$$\begin{aligned} Var(\hat{t}_\psi) &= E\left[(\hat{t}_\psi - t)^2\right] \\ &= \sum_i \left(\frac{t_i}{\psi_i} - t\right)^2 \psi_i \\ &= (11(16/7) - 300)^2(7/16) + (20(16/3) - 300)^2(3/16) \\ &\quad + (24(16/3) - 300)^2(3/16) + (245(16/3) - 300)^2(3/16) \\ &\approx 2.36 \times 10^5 \end{aligned}$$

This variance is much worse than the variance of the $\pi$ps estimator used in section 6.1, an estimator which is also unbiased. Therefore, we do not have a reason to use this distribution for $\psi_i$. Intuitively, it does not make sense to put so much weight on the smallest of the supermarkets.

### (2) Lohr, Chapter 6 Problem 12

### (a) Plotting Probability vs Number of Farms



There appears to be some positive correlation here, but it is not strong. Considering how much leverage Los Angeles county has, we cannot say in good conscience that this sampling design will be efficient for estimating total number of farms.

Does this make intuitive sense? Yes, the counties were sampled based on population size, but rural areas with plentiful farms tend to have lower population.

### (b) Total Number of Farms in US

The survey package seems to have a problem dealing in true with-replacement sampling.

```
svy.farms = svydesign(ids = ~countyunique, probs = statepop$psi, data = statepop)
total.farms = svytotal(~numfarm, svy.farms)
```

Table 1: Suspicious Estimate for Total Farms

| Total | Standard Error |
|---|---|
| 189630020 | 36742251 |

189 million farms in the U.S seems fairly suspicious.

If we use the Hansen-Hurwitz estimator for with-replacement sampling from Lohr (pg. 228),

$$\hat{t}_\psi = \frac{1}{n}\sum_{i=1}^{n}\frac{t_i}{\psi_i}$$

$$V\hat{a}r(\hat{t}_\psi) = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{t_i}{\psi_i} - \hat{t}_\psi\right)^2$$

we get the following results:

```
hansen.hurwitz = function(totals, psi, n){
  estimate = 1/n * sum(totals/psi)
  variance = (1/n)*1/(n-1) * sum( (totals/psi - estimate)^2 )
  SE = sqrt(variance)
  return(list("estimate" = estimate, "SE" = SE))
}
total.farms = hansen.hurwitz(statepop$numfarm, statepop$psi, n)
```

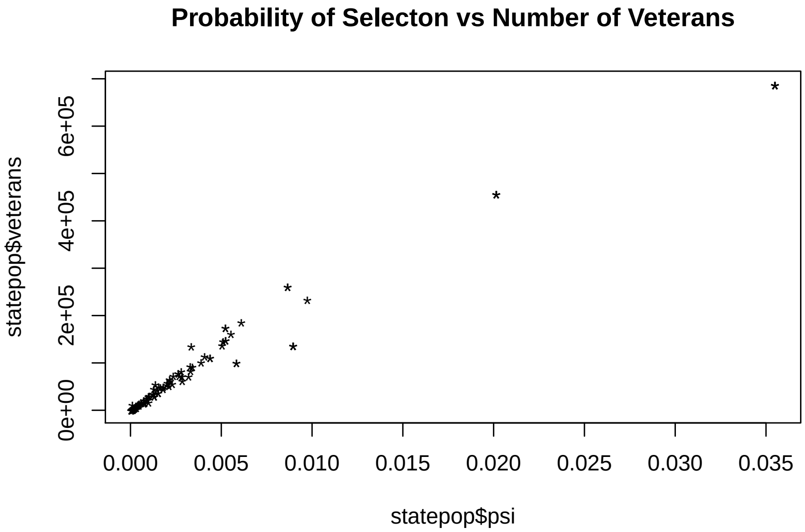
Table 2: Less Suspicious Estimate of Total Farms

| Total | SE |
|---|---|
| 1896300 | 367422.5 |

A brief google search confirms the suspicion. Until I have a moment to ask about the survey package functionality in with-replacement surveys, I will use the function constructed above.

### (3) Lohr, Chapter 6 Problem 13

### (a) Plotting Veterans vs Probability



This correlation appears much stronger than the correlation between probability and number of farms. The $\pi$ps design will be efficient here.

Does this fit with intution? The counties are sampled proportional to population. Greater population, more veterans.

### (b) Total Number of Veterans

```
vets.total = hansen.hurwitz(statepop$veterans, statepop$psi, n)
```

Table 3: Less Suspicious Estimate of Veterans in U.S.

| Total | SE |
|---|---|
| 27914180 | 1087453 |

### (c) Total Vietnam Veterans

```
statepop = statepop %>%
  mutate(vietVets = veterans * (percviet/100))
vietVets.total = hansen.hurwitz(statepop$vietVets, statepop$psi, n)
```

Table 4: Less Suspicious Estimate of Vietnam Veterans

| Total | SE |
|---|---|
| 8050477 | 327337.2 |

### (4) National Crime Victimization Survey

```
ncvs = read.csv("ncvs2000.csv") %>%
  mutate(individual = 1:nrow(.)) %>%
  mutate(ppsu = 10*pstrat + ppsu)
svy.ncvs = svydesign(ids = ~ppsu,strata = ~pstrat, weights = ~pweight,data = ncvs)
```

### (a) Total Crime in US

```
total.crime = svytotal(~numinc, svy.ncvs)
confi.crime = confint(total.crime) %>% data.frame()
confi.crime = cbind(total = total.crime[1], confi.crime)
```

Table 5: Total Crime Incidents Reported (2000)

| Total | 2.5% | 97.5% |
|---:|---:|---:|
| 15850360 | 14885814 | 16814906 |

### (b) Total Number of Crimes by Race

```
ncvs = ncvs %>%
  mutate(white.inc = 1*(race == 1) * numinc) %>%
  mutate(black.inc = 1*(race == 2) * numinc) %>%
  mutate(nativ.inc = 1*(race == 3) * numinc) %>%
  mutate(asian.inc = 1*(race == 4) * numinc)
svy.ncvs = svydesign(ids = ~ppsu,strata = ~pstrat, weights = ~pweight,data = ncvs)
total.white.inc = svytotal(~white.inc, svy.ncvs)
confi.white.inc = confint(total.white.inc) %>% cbind(total.white.inc[1], .)
total.black.inc = svytotal(~black.inc, svy.ncvs)
confi.black.inc = confint(total.black.inc) %>% cbind(total.black.inc[1], .)
total.nativ.inc = svytotal(~nativ.inc, svy.ncvs)
confi.nativ.inc = confint(total.nativ.inc) %>% cbind(total.nativ.inc[1], .)
total.asian.inc = svytotal(~asian.inc, svy.ncvs)
confi.asian.inc = confint(total.asian.inc) %>% cbind(total.asian.inc[1], .)
```

Table 6: Total Crimes Reported by Race (2000)

| | Total | 2.5% | 97.5% |
|---|---:|---:|---:|
| White | 12758528.6 | 11888204.5 | 13628852.7 |
| Black | 2531386.2 | 2235182.6 | 2827589.8 |
| Native American | 171359.5 | 34534.4 | 308184.6 |
| Asia-Pacific | 389085.9 | 299454.8 | 478717.1 |

### (c) Medical Expenses of Crime

```
total.expense = svytotal(~medexp, svy.ncvs)
confi.expense = confint(total.expense) %>% cbind(total.expense[1], .)
```

Table 7: Total Medical Expenses as a Result of Crime (2000)

| | Total | 2.5% | 97.5% |
|---|---:|---:|---:|
| medexp | 86285440 | 30711825 | 141859055 |

### (d) Average Robberies per Person

```
mean.rob = svymean(~robbery, svy.ncvs)
conf.rob = confint(mean.rob) %>% cbind(mean.rob[1], .)
```

Table 8: Average Robberies reported per Person (2000)

| Total | 2.5% | 97.5% |
|---:|---:|---:|
| 0.00164 | 0.00128 | 0.002 |

### (5) Return to NCVS with Incompetent Friend

First, we will explore what the results of our friend's method would be.

```
n = nrow(ncvs)
avg_friend = mean(ncvs$robbery)
SE_friend = sqrt(sum( (ncvs$robbery - avg_friend)^2 )/n)
noDesign = data.frame(avg_friend, SE_friend)
withDesign = data.frame(mean.rob) %>% unname()
```

| Friend's AVG | Friend's SE | HT-AVG | HT-SE |
|---:|---:|---:|---:|
| 0.0014869 | 0.0404462 | 0.00164 | 0.0001833 |

Let $\tilde{y}$ denote our friend's estimate of the average robberies per person.

We first note the the variance of $\tilde{y}$ is much higher than the variance of the Horvitz-Thompson Estimator. Immediately, this makes us suspicious. We want accuracy in our estimate, especially if we inted to craft policy or interventions based on this data.

Also, this estimate is biased.

$$E(\tilde{y}) = E\left(\frac{1}{m}\sum_{i=1}^{m} y_i\right) = \frac{1}{m}\sum_{i=1}^{M_0} y_i\psi_i$$

A biased estimator with high variance isn't of much use to use when we have an unbiased estimator with lower variance.