Stats 332 Assignment 1

Junqiao Lin

January 16, 2019

Question 1.

Target population: Homeless people within U.S.

Frame: Homeless clinics across the country

Sampling Unit: A clinic

Observation unit: a homeless person visiting the clinic

Sample population: Any homeless person who visited a Homeless clinics

Possible Source of selection bias: Since drug use is illegal within the U.S. The target population might not answer the question truthfully when confronted during the servey.

Question 2.

a) Target population: Computer user within the world Frame: People respond to the online survey from a tech magazine.

b) i) The population frame is center around people reading a tech magazine, who presume have more experience with the computer and might have a different definition for significant malfunction then an average computer user.

ii) Since the incentive is a lottary for a new Macbook Pro, most people are incentivize to quickly fill out the survey instead of reading the 123 questions carefully.

iii) Even though the frame is from people reading tech magazine, people might not know what all 123 computer problem is specifically (Since it's so detailed). This might cause a miss understanding and confuse a minor error with a significant malfunction.

Question 3.

a)

Let I_i be the ith indicator variable for y_i note:

$$\begin{aligned} E(\hat{\tau}) &= 2\Sigma_{i=1}^{N} y_i E(I_i) \\ &= 2\Sigma_{i=1}^{N} y_i \frac{1}{2} \\ &= \Sigma_{i=1}^{N} y_i \\ &= \tau \end{aligned}$$

Therefore $\hat{\tau}$ is an unbiased estimator for τ.

b) Note for any $1 \leq i < j \leq N$,I_i and I_j are independent, thus $Cov(I_i, I_j) = 0$ for $i \neq j$, Thus:

$$\begin{aligned} Var(\hat{\tau}) &= Var(2\Sigma_{i=1}^{N} I_i y_i) \\ &= 2^2 \Sigma_{i=1}^{N} y_i^2 Var(I_i) + \Sigma_{i=1}^{N} \Sigma_{j=1, j\neq i}^{N} y_i y_j Cov(I_i, I_j) \\ &= 4\Sigma_{i=1}^{N} y_i^2 (0.5)^2 + \Sigma_{i=1}^{N} \Sigma_{j=1, j\neq i}^{N} y_i y_j 0 \\ &= \Sigma_{i=1}^{N} y_i^2 \end{aligned}$$

c) The unbiased estimator of $Var(\hat{\tau})$ is $\hat{Var}(\hat{\tau}) = 2\Sigma_{i\in S} y_i^2$. Since

$$\begin{aligned} E[2\Sigma_{i\in S} y_i^2] &= 2\Sigma_{i=1}^{N} y_i^2 E[I_i] \\ &= \Sigma_{i=1}^{N} y_i^2 \end{aligned}$$

Since $E[\hat{Var}(\hat{\tau})] = Var(\hat{\tau})$, therefore the estimator is unbiased

Question 4.

```
library(survey)
```

```
Q4.dat<- read.csv("bookssn.csv")

#initalizing the design for the variable
Q4.srs <- svydesign(ids = ~1, data = Q4.dat, fpc = ~N)
Q4.N <- Q4.dat$N[1]
Q4.N
```

```
## [1] 19552860
```

a) 95 percent CI for average number of books read by a Florida resident in the past year

```
confint(svymean(~books,Q4.srs))
```

```
##         2.5 %   97.5 %
## books 5.83667 11.10839
```

Thus we have the estimate for average being 8.47253, the SE of the estimate being 2.63586 and the 95 perent CI being (5.83667,11.10839)

b) 95 percent CI for the proportion of Florida residents who use Facebook/Twitter daily

```
confint(svymean(~sn,Q4.srs))
```

```
##        2.5 %    97.5 %
## sn 0.1726973 0.3547753
```

Thus we have the estimate for the proportion being 0.2637363, the SE of the estimate being 0.091039 and the 95 perent CI being (0.1726973, 0.3547753)

c) Note as shown in class, in order to ensure that the total width of the 95% CI for the population mean of books to be less than $c = 0.1$ we need:

$$\frac{1}{n} < \frac{1}{\sigma^2}(\frac{c}{2(1.96)})^2 + \frac{1}{N}$$

Obtain an estimate of the variance as per following:

```
Q4.varbook <- var(Q4.dat$books)
Q4.varbook
```

```
## [1] 164.5853
```

Thus we have

$$\frac{1}{n} < \frac{1}{164.5853}(\frac{0.1}{2(1.96)})^2 + \frac{1}{19552860}$$
$$n > 249678.86$$

Thus, in order to obtain a 95 percent interval wth the total width being less than 0.1, you need at lease 249679 response from the population.