Final Project Step One

Sara Morell

2/6/2021

Ever Smoked One Cigarette

The expected proportion of the population that have smoked at least one cigarette is 40% or .4.

We have a desired coefficient of variation of .05, which means that we can compute the standard error by multiplying $.4 * .05$, which is .02. This means that $\hat{V^2}$ is $.02^2$ or .0004.

The element variance for each variable is $p(1-p)$ or $.4 * .6$, which equals .24.

This means the minimum n needed to achieve the desired level of precision is $\frac{.24}{.02}$ or 600 respondents.

```
#expected proportion for smoked one cigarette is .40

#cv_1 = se_1/.4
se_1 <- .4 * .05
v2_1 <- se_1^2
v2_1
```

```
## [1] 4e-04
```

```
#p(1-p)
s2_1 <- .4*.6

s2_1/v2_1
```

```
## [1] 600
```

Ever Smoked Marijuana

The expected proportion of the population that has smoked marijuana at least once is 15% or .15.

We have a desired coefficient of variation of .05, which means that we can compute the standard error by multiplying $.15 * .05$, which is .0075. This means that $\hat{V^2}$ is $.0075^2$ or .0000563.

The element variance for each variable is $p(1-p)$ or $.15 * .85$, which equals .1275.

This means the minimum n needed to achieve the desired level of precision is $\frac{.1275}{.0000563}$ or 2266.67 respondents. Since respondents can't be fractions, this means we would want to sample 2267 individuals.

```
#expected proportion for smoked marijuana .15.

#cv_1 = se_2/.15
se_2 <- .15 * .05
v2_2 <- se_2^2
v2_2
```

```
## [1] 5.625e-05
```

```
#p(1-p)
s2_2 <- .15*.85

s2_2/v2_2
```

```
## [1] 2266.667
```

Average Age First Smoked

The expected average age when one the respondents either smoked a cigarette or marijuana is 13 and the expected standard deviation is 3.

Since our desired coefficient of variation is .05, this means that the standard error is $13 * .05$, which equals .65. This means that $\hat{V}^2$ is $.65^2$ or .4225.

The element variance for each variable is the standard deviation squared, or 3^2, which equals 9.

The formula for calculating our desired n value is $\frac{S^2}{V^2}$, since we're ignoring the finite population correction. So the minimum desired sample size is $\frac{9}{.4225}$ or 21.3 respondents. Since this is the minimum we want to achieve, it means we would need to sample at least 22 respondents.

```
#expected average age started smoking is 13. expected standard deviation is 3.

#cv_1 = se_3/13
se_3 <- 13 * .05
v2_3 <- se_3^2
v2_3
```

```
## [1] 0.4225
```

```
#standard deviation
s2_3 <- 3

s2_3^2/v2_3
```

```
## [1] 21.30178
```

Calculating Synthetic Roh

From a previous study, we have obtained estimates of the following design effects for each of these three estimates:

- Proportion ever smoked one cigarette: 2.5
- Proportion ever smoked marijuana: 2.0
- Mean age when first asked to smoke: 1.7

```
deff_1 <-2.5
deff_2 <-2.0
deff_3 <-1.7
```

This previous study featured a sample of size n = 7,500 students between the ages of 13 and 19, selected from a total of a = 150 clusters. Therefore, b = 50 elements per cluster by substitution. To calculate roh, substitute b and the corresponding design effect values to the formula $roh = \frac{deff-1}{b-1}$

```
b <- 50
```

- Proportion ever smoked one cigarette

  ```
  roh_1<-(deff_1-1)/(b-1)
  roh_1
  ```

  ```
  ## [1] 0.03061224
  ```

- Proportion ever smoked marijuana

  ```
  roh_2<-(deff_2-1)/(b-1)
  roh_2
  ```

  ```
  ## [1] 0.02040816
  ```

- Mean age when first asked to smoke

  ```
  roh_3<-(deff_3-1)/(b-1)
  roh_3
  ```

  ```
  ## [1] 0.01428571
  ```

Optimum Subsample Size

```
#Total budget; According to client, $500k is the budget after accounting for fixed cost so C-Co = 50000
budget <- 500000
#Cost per primary stage cluster
ca <- 3000
#Cost per element within cluster
cb <- 50
```

$b_{opt} = \sqrt{\frac{c_a}{c_b} * \frac{1-roh}{roh}}$

```
bopt_1 <- sqrt((ca/cb)*((1-roh_1)/roh_1))
bopt_2 <- sqrt((ca/cb)*((1-roh_2)/roh_2))
bopt_3 <- sqrt((ca/cb)*((1-roh_3)/roh_3))

bopt_1
```

```
## [1] 43.58899
```

```
bopt_2
```

```
## [1] 53.66563
```

```
bopt_3
```

```
## [1] 64.34283
```

$a_{opt} = \frac{c-c_o}{c_a+b_{opt}c_b}$

```r
aopt_1 <- (budget)/(ca + bopt_1*cb)
aopt_2 <- (budget)/(ca + bopt_2*cb)
aopt_3 <- (budget)/(ca + bopt_3*cb)

#a_opt for proportion ever smoked cigarette
aopt_1
```

```
## [1] 96.53536
```

```r
#a_opt for proportion ever smoked marijuana
aopt_2
```

```
## [1] 87.97734
```

```r
#a_opt for mean age when first asked to smoke
aopt_3
```

```
## [1] 80.42281
```