Gawron_Homework11

Nicholas Gawron

4/26/2021

Question 1

Part A

We consider the formula $k \leq \frac{N}{n}$. Given a population size of $N = 650$ and $k = 10$ we arrive at: $10 \leq \frac{650}{n}$. Algebra tells us $n \leq \frac{650}{10} = 65$. In this case $n = 65$.

Part B

```
N<- 650
n<- 65
phat <- 48/65
varp <- (1 - 65/650)*((phat*(1-phat))/(n-1))
B<-2*sqrt(varp)
```

We assume the population is random ordered. This allows us to use SRS formulas. $\hat{p} = 48/65$, which equals 0.7384615. We also can find our variance thru the formula: $var(\hat{p}) \approx (1 - \frac{n}{N})\frac{pq}{n-1}$. We use $B = 2\sqrt{Var}$ to compute the bound.

- Proportion of members in favor in sample:0.7384615
- Our variance of the estimate is : 0.002716
- Bound on error of estimation: 0.1042301

Part C

```
Br <- .1
q <- 1 -phat
nnewQu1 <- ((N*phat*q)/(((N-1)*Br^2)/4+phat*q));nnewQu1
```

```
## [1] 69.14296
```

Given our required bound of $B = .1$, the sample size needed is: 70. This required us to round up to the nearest whole number. This makes logical sense due to the fact our bound decreases slightly from the bound in part B. This calls for slightly more data points - to the tune of 5 more. This tells us that $k < N/70 = 650/70 = 9.285714$, this tells us that $k = 9$. Our K value decreases.

Question 2

Part A

We again note that k in this case is 50. Our population size is given as $N = 15200$. It follows from the formula: $k \leq N/n$ that $n \leq N/k = 15200/50 = 304$

Part B

It is important to note that we will be using proportions once again. Since $\hat{p}$ is the proportion of renters, $N\hat{p}$ should be an estimate of the total number of renters. Further more $Var(N\hat{p}) = N^2 Var(\hat{p})$

```
n<-304 #sample size
N <- 15200
phat2 <- 88/304
Tau2 <- phat2*N
q2 <- 1 - phat2 #q hat
#below calculates variance of p hat
varp2 <- (1 - 304/15200)*((phat2*(q2))/(n-1))
varpTot <- (N^2)*varp2
B2<-2*sqrt(varpTot); B2
```

```
## [1] 784.0792
```

So it follows that our estimate are:

- For the total number of renters: 4400
- Our variance for the estimate is 1.5369505×10^5
- Our bound on the error of this estimation is: 784.0792039. This comes from taking the above variance,sqaure rooting it and multiplying by 2.

Question 3

Discuss the relative merits of ratio, regression and difference estimation.

The three methods of estimation: regression ratio and difference estimation rely on a certain kind of data set. The estimation process for these three schemes involves using an auxiliary variable in addition to the response of interest to gain more information about the population parameters with respect to the response. We will be given information about the population in terms of an auxiliary variable. Ratio estimation is most useful when the response and auxiliary variables havea strong realationship and is linear through the origin. The main pitfall of ratio estimation is when we do not have our data go through the origin. Regression and difference estimation cover auxlillry response data that can be affine (or not go through the origin). Regression estimation requires the most computational work - but allows us to consider a general linear least squares regression model which takes into account a general slope and intercept for the affine data. As with all other regession methods listed here - we assume the aux. variable is fixed and something already observed. Difference estimation allows for us to do something similar to regression - but we set the 'slope' term to one. This would be great of we had our data follow a general trend of $y = x + C$ for some affine shift C. This makes the method a lot less computationally difficult than regression estimation;. However, again our variables need to be highly correlated and on the same scale - to allow us to use a slope of 1.

Question 4

A question that could force a response in a certain direction because of its strong wording.

"Do you agree that **common sense** gun laws should be put in place to stop **evil and heinous violence** ?"

The use of the phrase *common sense* talks down to the respondent making them believe that these laws are very basic and sensible to any person. Further charged language is used when address the impacts of gun violence as **evil and heinous**. While this is true, using this phrasing in a question may make a respondent feel they themselves are not doing the right thing is they answer no to the question. If they answer no to the question - given this phrasing - a respondent may believe that they are supporting these *evil and heinous* acts. Since people don't want to be perceived as evil this may skew results in favor of supporting legislation.

Question 5

Response rate is an important consideration because if certain elements of the population do not respond the survey could have missed an important part of the population of interest. We want the lowest possible response rate so we can get a greater representation of the population as a whole.

Some methods for reducing non-response:

- Including an incentive (usually monetary)is a great way to decrease non-response.
- We can also consider the length of the interview, especially in telephone interviews, and make them as short as possible to increase response rate. According to a study from the textbook : When the interviewer would make a mention of a 10- minute interview surveyors got a 43% compliance rate. In contrast to when no mention of time was made (implying shorter than 10 minutes) got a much higher compliance rate.

Question 6

```
 Cluster <- read.table("~/RCodeST308/ST432/Cluster.txt", quote="\"", comment.char="")

print("mean number of saws per industry")
```

```
## [1] "mean number of saws per industry"
```

```
mean(Cluster$V2)
```

```
## [1] 6.5
```

```
head(Cluster)
```

```
##   V1 V2  V3
## 1  1  3  50
## 2  2  7 110
## 3  3 11 230
## 4  4  9 140
## 5  5  2  60
## 6  6 12 280
```

```
N<-96 #total number of clusters
n<-20 # number of clusters collected in SRS
mi <- Cluster$V2 # number of saws in each cluster in the sample.
mbar <- mean(mi);mbar # average cluster size for the sample, avg number of saws
```

```
## [1] 6.5
```

```
yi <- Cluster$V3 #total observations of cost per sampled cluster
```

Part A

```
Ybar <- sum(yi)/sum(mi);Ybar
```

```
## [1] 19.73077
```

```
SrSqr <- (1/(n-1))*sum((yi-Ybar*mi)^2);SrSqr
```

```
## [1] 845.5607
```

```
VarY <- (1-n/N)*( (SrSqr)/(n*mbar^2) ) ;VarY
```

```
## [1] 0.792192
```

```
BoundY <- 2*sqrt(VarY);BoundY
```

```
## [1] 1.780103
```

It is important to note that since we do not know M, we must estimate $\bar{M}$ with $\bar{m}$, a quantity that we do know. This works out to be that $\bar{M}^2 \approx \bar{m}^2$ 42.25 We come to the following conclusions.

- Our estimate for the average repair cost per saw is estimated by $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} m_i}$ is the value 19.7307692
- Our S_r^2 value comes from `SrSqr <- (1/(n-1))*sum((yi-Ybar*mi)^2)` and is: $\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}m_i)^2$ which computes to 845.5607287
- Our variance is calculated with $\approx (1 - n/N)\frac{S_r^2}{n\bar{m}^2}$, which is:0.792192
- Our estimate for the bound on the error of estimation is: 1.7801034

Part B

We are estimating totals - the only issue is that we do not know the value of M or the number of saws used in all industries.The requires us to compute $\bar{Y}_t = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Further on the second line of code below we compute an unbiased estimator for the totals in the population $N\bar{Y}_t = \tau_t$

```
Yt <- sum(yi)/n
tauNoM <- N*Yt
StSqr <- (1/(n-1))*sum((yi - Yt)^2) #S_t^2
VarTNoM<- (N^2)*(1-(n/N))*(StSqr/n) # Variance of Tau
BNoM <- 2*sqrt(VarTNoM)
```

Our estimate follow:

- Our estimate for the total,τ_t, is: 1.2312×10^4
- Variable $S_t^2 = (\frac{1}{n-1})\sum(Y_i - \bar{Y}_t)^2$ which is 6908.6184211
- Our estimate for the Bound is: 3175.0678733

Part C

The new information in this part gives us the fact that $M = 710$ since there re 710 band saws across all N industries. We will now recompute the total. We note the estimate is : $\hat{\tau} = M\bar{Y}$ where $\bar{Y}$ is the estimate of the mean from **part (a)**. Our variance calculation follows in a similar manner: $Var(\tau) = Var(M\bar{Y}) = M^2 Var(\bar{Y})$ This leads to a nice simplification since $\bar{M}^2 = (M/N)^2$. It follows that

$$Var(\tau) = M^2\left(1 - \frac{n}{N}\right)\frac{Sr^2}{n} = M^2 Var(\bar{Y})$$

```
M<-710
TauYwithM <- M*Ybar
VarTau <- (M^2)*(VarY)
BwM <- 2*sqrt(VarTau);BwM
```

```
## [1] 1263.873
```

We come to the conclusions that

- Our new estimate for the total is: 1.4008846×10^4
- Our new estimate for the bound is: 1263.8733945
- Given that this bound is less than the bound calculated in **part b** which was 3175.0678733, we know the bound is better when we have the M at our disposal.

Part D

We want to estimate the average repair cost per saw for next month. How many clusters should he select for his sample if he wants the bound on the error of estimation to be no larger than $2?

- We note that $\sigma_r^2 \approx S_r^2$
- We note that $B = 2$ in this case and $D = \frac{B^2 \bar{M}^2}{4}$, this is computed in line 5 below
- We then compute $n = \frac{N\sigma_2^2}{ND+\sigma_2^2} \approx \frac{NS_r^2}{ND+S_r^2}$, which is computed on line 6

```
B<- 2 # 2 dollars is bound
N<-96 #given previously as num industry
M<- 710 #given from part c
MBar <- M/N
D <- .25*(B^2)*(MBar^2) # computes D
n <- (N*SrSqr)/ (N*D+SrSqr);n
```

```
## [1] 13.3146
```

- If M is known we have sample size (or g the number of clusters to sample): 13.3146013. We will round the value up to get the correct number of clusters needed for this bound: 14

What if M is not known. We then use the approximation that $\bar{m}^2 \approx \bar{M}^2$ in our computation of D.

```
DNoM <- .25*(B^2)*(mean(mi)^2) # computes D
nNoM <- (N*SrSqr)/ (N*DNoM+SrSqr);nNoM
```

```
## [1] 16.56081
```

If M is not known, we will round up the value to get the correct number of clusters needed for this bound which is: 17

This all makes sense because we are increasing our bound slightly (from 1.7801034 to 2), we will not need as much data points since we are being less accurate with our error on an estimation of the mean.