Gawron_Homework9

Nicholas Gawron

3/30/2021

```
knitr::opts_chunk$set(echo = TRUE, warning=FALSE, message= FALSE ,tidy=FALSE)
library(tidyverse)
library(readxl)
library(tinytex)
library(ggplot2)
```

Problems

Question 1

```
StdDog <- c(14.3, 15.7 ,17.8 ,17.5 ,13.2 ,18.8 ,17.6 ,14.3 ,14.9 ,17.9 ,19.2)
NewDog<- c(15.2, 16.1, 18.1, 17.6, 14.5, 19.4, 17.5, 14.1, 15.2, 18.1, 19.1)

rhat <- sum(NewDog)/sum(StdDog) # r coeff.
Ndog<- 700
ndogP <- 11 #n' is the sample size pilot  study

Mux<- 17

SigSq<- (1/(ndogP-1))*(sum((NewDog-rhat*StdDog)^2))

#Bound of 0.01, sample size for estimate of the ratio of averages
Br = 0.01
sample4r <- (Ndog*SigSq)/(((Ndog*Br^2*Mux^2)/4)+SigSq)
sample4r
```

```
## [1] 30.16295
```

```
#bound is .25 seconds
Bux = 0.25
sample4uy <- (Ndog*SigSq)/(((Ndog*Bux^2)/4)+SigSq)
sample4uy
```

```
## [1] 14.2781
```

```
#bound of total, consider bound for 250 seconds
Bty <- 250
sample4ty <-(Ndog*SigSq)/(((Bty^2)/(4*Ndog))+SigSq)
sample4ty
```

```
## [1] 7.069815
```

1. Question on finding sample size. Note that we will use a piloit study of a smaller sample size to estimate

$\sigma^2 \approx S_R^2 = \frac{1}{n'-1}\sum_{i=1}^{n'}(y_i - \hat{r}x_i)^2$. Note that n‘ represents the sample size of 11 used in our pilot study. Note that our approximated value of σ^2 is 0.2277407.

a. Using the data above, how many dogs must be looked at to estimate the ratio of averages to within a bound of B_r =.01? We will use $n_r = \frac{N\sigma^2}{\frac{NB^2\mu_x^2}{4}+\sigma^2}$

   This is computed by the R code: `sample4r <- (Ndog*SigSq)/(((Ndog*Br^2*Mux^2)/4)+SigSq)`. Our rounded up sample size is: 31

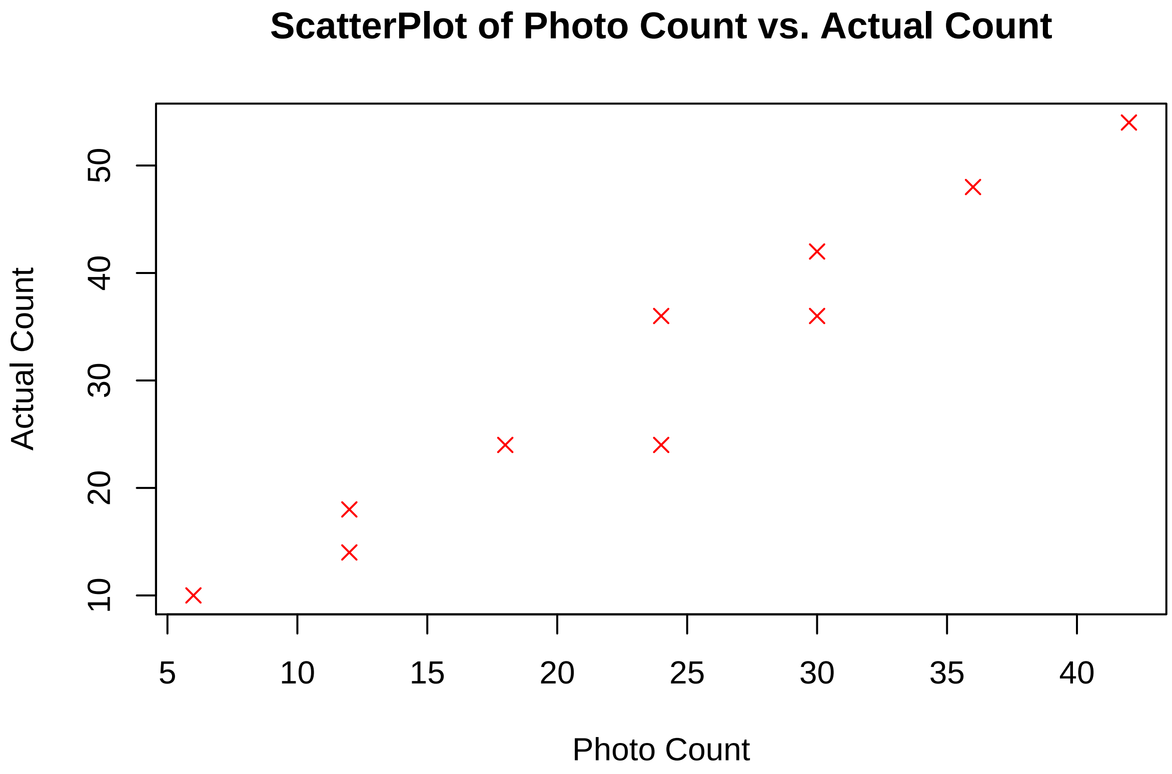
b. We will now compute hjow many dogs must be looked at to estimate the average time required under the new training to within a bound of .25 seconds. Since we are considering an *average* we will use the formula ($n_\mu = \frac{N\sigma^2}{\frac{NB^2}{4}+\sigma^2}$ This is computed by the R code: `sample4uy <- (Ndog*SigSq)/(((Ndog*Bux^2)/4)+SigSq)`. Our rounded up sample size is: 15

c. We will now determine how many dogs must be looked at to estimate the *total* time required under the new training to within a bound of 250 seconds. In this case of $B = 250$. Note that since we are determining a sample size with estimating a total we will use the formula: $n_\tau = \frac{N\sigma^2}{\frac{B^2}{4N}+\sigma^2}$

   This is computed by the R code: `sample4ty <-(Ndog*SigSq)/(((Bty^2)/(4*Ndog))+SigSq)`. Our rounded up sample size is: 8

Question 2

```r
Pct <- c(12, 30, 24, 24, 18, 30, 12, 6, 36, 42)
Act <- c(18 ,42 ,24 ,36 ,24 ,36, 14 ,10 ,48 ,54)
plot(Act~Pct, main="ScatterPlot of Photo Count vs. Actual Count",
     ylab = "Actual Count", xlab ="Photo Count", pch= 4, col = "red")
```



```
N = 200 # Number of plots
n = 10 # number of plots

rhat <- mean(Act)/mean(Pct) #r value

SrSqu<- (1/(n-1))*(sum((Act-rhat*Pct)^2))

TPct<- 4200

TAct<- rhat*TPct
fpc <- (1-n/N)
SrSqu<- (1/(n-1))*(sum((Act-rhat*Pct)^2))
VarAct <- N^2*fpc*(SrSqu/n)
BoundYnew <-2*sqrt(VarAct)
```

Note that we estimate r thru our sample by computing the sum of our data points for the actual count divided by the sum of the photo count. This is computed as `rhat <- mean(Act)/mean(Pct)`. Our $\hat{r}$ value is 1.3076923. We compute our estimated total for number of dead firs thru the R code `TAct<- rhat*TPct`. This takes advantage of the formula $\tau_{actual} = \hat{r}\tau_{Photo}$. It follows that our estimated total for the actual number of dead firs is 5492.3076923. We will now compute the bound on the error of estimation. This will first require us to compute $S_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - rx_i)^2$, where y_i corresponds to the **actual value** of dead firs in the sample. This value is written out by the code`SrSqu<- (1/(n-1))*(sum((Act-rhat*Pct)^2))`, and is computed to be: 12.0762656.

Our bound is then computed thru the formula:

$$B = 2\sqrt{N^2(1 - \frac{n}{N})\frac{S_R^2}{n}}$$

This is calculated thru the R code `VarAct <- N^2*fpc*(SrSqu/n)` and `BoundYnew <-2*sqrt(VarAct)`. Note that the `fpc` term is our finite population correction. The bound on the error of estimation is: 428.4381371. We can create a confidence interval of the form: $(\tau_{pct} \pm B_{pct})$

It follows that we have the 95% confidence interval: (5063.8695552, 5920.7458294). Our true total of dead firs is captures by this interval (with 95% confidence).

Question 3

We use ratio estimation when we cannot compute our sample mean for a certain variable. We also will have to assume we consider measurements under another condition. We must assume that the response variable (Y) and the subsidiary variable (X) are related. This relationship would come up from our constant $\hat{R}$.

The method of multiplying $N\bar{y}$ is not feasible because in some cases it can be too time-consuming and costly to determine N (i.e., to count the total number in a population). We can avoid the need to know N by noting the following two facts. First, that auxiliary and response variables, as stated above, are closely related. Furthermore, the ratios our of desired variable equal the ratio of our sample values. In particular $\frac{\mu_y}{\mu_x} = \frac{\tau_y}{\tau_x}$. This will allow us to forgo calculating a value N if it is impossible - or not cost effecitve.

Suppose we want to determine μ_y for a large value of N, we could use a sample mean method. However, if x and y are correlated, a ratio estimator that uses information from the auxiliary variable x frequently provides a more precise estimator of μ_y.

Question 4

A simple random sample between each strata is used of customers for each brand is contacted and asked to provide a potential sales figure y (in number of units) for the coming quarter. Last year's true sales figure, for the same quarter, is available for each of the sampled customers and is denoted by x.

```r
B1x<-c(204,143,82,256,275,198)
B1y<-c(210,160,75,280,300,190)
B2x<-c(137,189,119,63,103,107, 159,63,87)
B2y<-c(150,200,125,60,110,100,180,75,90)
B1 <- data.frame(x = B1x, y =B1y)
B2 <- data.frame(x = B2x, y =B2y)
Tx1 <- 24500; Tx2 <- 21200
N1 <- 120 ; N2<- 180; N<- N1+N2
```

a. We will use separated sampling:

We note that

$$\begin{aligned}
\tau_{y_{SR}} &= N\mu_{y_{SR}} \\
&= N\sum_{i=1}^{2}\frac{N_i}{N}\mu_{x,i}\hat{R}_i \\
&= \sum_{i=1}^{2}N_i\mu_{x,i}\hat{R}_i \\
&= \sum_{i=1}^{2}\frac{N_i}{N}\tau_{x,i}\hat{R}_i && \text{Since}N\mu = \tau
\end{aligned}$$

Now we compute our bound.

$$\begin{aligned}
B_{\tau_y} &= 2\sqrt{Var(\tau_y)} \tag{1}\\
&= 2\sqrt{Var(N\mu_y)} \tag{2}\\
&= 2N\sqrt{Var(\mu_y)} \tag{3}\\
&= 2N\sqrt{\frac{N_1}{N}^2(1-\frac{n_1}{N_1})\frac{S^2_{r,1}}{n_1}+\frac{N_2}{N}^2(1-\frac{n_2}{N_2})\frac{S^2_{r,2}}{n_2}} \tag{4}\\
&= NB_{\mu_y} \tag{5}
\end{aligned}$$

```r
R1hat <- mean(B1$y)/mean(B1$x)
R2hat <- mean(B2$y)/mean(B2$x)
n2 <-length(B2$x)
n1 <- length(B1$x)
n <- n1 +n2

Ty <- R1hat*(N1/N)*Tx1 + R2hat*(N2/N)*Tx2; Ty
```

```
## [1] 23782.68
```

```r
Sr1sq <- sum((B1$y-R1hat*B1$x)^2)/(n-1)

Sr2sq <- sum((B2$y-R2hat*B2$x)^2)/(n-1)

VarMuy <- (((N1/N)^2)*(1-n1/N1)*(Sr1sq/n1))+(((N2/N)^2)*(1-n2/N2)*(Sr2sq/n2))

BTauY <- 2*sqrt(VarMuy*N^2); BTauY
```

```
## [1] 986.1686
```

The estimated total is computed by the line `Ty <- R1hat*(N1/N)*Tx1 + R2hat*(N2/N)*Tx2`. We have determined that our estimated total is: 2.3782676×10^4.

Our estimated bound is computed by the lines: `BTauY <- 2*sqrt(VarMuy*N^2)`. Here we use our bound equation but must mulitpliy the variance of the mean by N^2. Note that the variance of μ_y is given by: Our bound is computed to be: 986.1685785

b. We will now show combined method of computing the total. Note that

- $R_{CR} = \frac{\bar{Y}_{st}}{\bar{X}_{st}}$
- We first compute each stratified estiamted mean i.e $\bar{X}_{st}$ from the line `Xst <- (mean(B1$x)*N1 + mean(B2$x)*N2)/N`
- We then compute R_{CR} by dividing these stratified samples: `RhatC <- Yst/Xst`
- We then compute our total for the y variable by computing: `TyCR <- RhatC*TxCr`. Note that `TxCr` is the combined total across all strata's for the subsidiary variable (x).
- We then compute the variance of the mean of our desired variable y… this is thru the line: `VarCR<- (((N1/N)^2)*(1 - (n1/N1))*(SrCRsq1/n1)) + (((N2/N)^2)*(1 - (n2/N2))*(SrCRsq2/n2))`
- We will compute the variance of the total by noting that $var(N\mu_y) = N^2 var(\mu_y)$
- We then compute the bound of the total: `BoundCR <- 2*sqrt(VarCR*N^2)`

```
Xst <- (mean(B1$x)*N1 + mean(B2$x)*N2)/N
Yst<- (mean(B1$y)*N1 + mean(B2$y)*N2)/N
RhatC <- Yst/Xst ; RhatC
```

```
## [1] 1.05492
```

```
TxCr <- 24500+21200
TyCR <- RhatC*TxCr; TyCR
```

```
## [1] 48209.84
```

```
SrCRsq1 <- (sum((B1$y-RhatC*B1$x)^2))/(n1-1);SrCRsq1
```

```
## [1] 159.2125
```

```
#SR_2^2 value
SrCRsq2 <- (sum((B2$y-RhatC*B2$x)^2))/(n2-1);SrCRsq2
```

```
## [1] 58.32331
```

```
#Variance of the total total under  Y
VarCR<- N^2*((((N1/N)^2)*(1 - (n1/N1))*(SrCRsq1/n1)) + (((N2/N)^2)*(1 - (n2/N2))*(SrCRsq2/n2)));VarCR
```

```
## [1] 562470.3
```

```
# bound on error of estimation below
BoundCR <- 2*sqrt(VarCR);BoundCR
```

```
## [1] 1499.96
```

- Our $\hat{R}_{CR}$ value is 1.0549199
- Our estimated total $\tau_{y_{CR}}$ is: 4.820984×10^4
- Our variance of the total $var(\tau_{y_{CR}})$ is: 5.0622323×10^{10}
- Var = `(180/300)^2*(1-9/180)*(58.32331/9)+(120/300)^2*(1-6/120)*(159.2125/6)` is the variance of the mean. We then multiply this value by `300^2` or N^2 to give us the variance of the total with respect to the variable y. Finally - we square root this value and multiply it by 2 to achieve our bound as follows: `2*sqrt(((180/300)^2*(1-9/180)*(58.32331/9)+(120/300)^2*(1-6/120)*(159.2125/6))*300^2)`. Through this calculation our bound is: 1499.9602931
- Our bound on the error of estimation for the total is: 1499.9603356

As from lecture - we see that the seperated ratio sampling scheme yeilds a better (or less) bound on the error of estimation.