Gawron_Homework8

Nicholas Gawron

3/22/2021

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

Corresponding code for a problem's part will be **ABOVE** the solution.

Question 1

Professorship Frequency Sample Size Full 266 403 Associate 101 299 Assistant 261 598

```
Fullp <- 266/403 ; n1 <- 403
Assop<-101/299; n2 <- 299
Assip<-261/598; n3<- 598

pst<- .1*Fullp+.25*Assop+.45*Assip

Sn1 <- Fullp*(1-Fullp)/(n1-1)
Sn2 <- Assop*(1-Assop)/(n2-1)
Sn3 <- Assip*(1-Assip)/(n3-1)

BProf <- 2*sqrt( (.3^2)*Sn1+(.25^2)*Sn2+(.45^2)*Sn3)
```

a.

We will be using the data given as a stratified random sample. Note that

$$\hat{p}_{st} = \frac{1}{N}(N_1\hat{p}_1 + N_2\hat{p}_2 + N_3\hat{p}_3)$$

is the sample mean.

- **N** is the population size of all professors
- N_i is the population of all professors in job i
- i corresponds to the row (or strata) on the given table: 1 corresponds to Full Prof. , and 2 corresponds to Associate and $i = 3$ is Assistant.
- $\hat{p}_i$ is the proportion of faculty that believe they give good effoer of all professors in respective sample from strata i

We note from the problem statement that "full professors, are composing 30% of the population, associate professors, composing 25% of the population, assistant professors, composing 45% of the population", we can create a formula for N, the total population.
Note, $N_1 = .3N$, $N_2 = .25N$,$N_3 = .45N$. We will now modify our expected value formula accordingly:

$$\hat{p}_{st} = \frac{1}{N}(.3N\hat{p}_1 + .25N\bar{X}_2 + .45N\hat{p}_3) = .1\hat{p}_1 + .25\hat{p}_2 + .45\hat{p}_3$$

This is computed in the fourth line of executable code in the chuck above. The expected value of the proportion is: **0.3468578**.

We will now compute the bound of error of estimation. Note that it follows from above that: $N_1^2 = .3^2N^2$, $N_2^2 = .25^2N^2$,$N_3^2 = .45^2N^2$

$$\begin{aligned}
B &= 2\sqrt{\frac{1}{N^2}\sum_1^3 N_i^2 \frac{\hat{p}_i\hat{q}_i}{n_i-1}} && (1)\\
&= 2\sqrt{\frac{N^2}{N^2}\Big(.3^2\frac{\hat{p}_1\hat{q}_1}{n_1-1} + .25^2\frac{\hat{p}_2\hat{q}_2}{n_2-1} + .45^2\frac{\hat{p}_3\hat{q}_3}{n_3-1}\Big)} && (2)\\
&= 2\sqrt{.3^2\frac{\hat{p}_1\hat{q}_1{}^2}{n_1-1} + .25^2\frac{\hat{p}_2\hat{q}_2}{n_2-1} + .45^2\frac{\hat{p}_3\hat{q}_3}{n_3-1}} && (3)\\
&= 2\sqrt{.3^2(Sn_1) + .25^2(Sn_2) + .45^2(Sn_3)} && (4)
\end{aligned}$$

We compute the bound in the bottom of our chunk listed above. The value of the bound is calculated by the equation, `BProf <- 2*sqrt( (.3^2)*Sn1+(.25^2)*Sn2+(.45^2)*Sn3)` and stored as: `BProf`. Our bound on the error of estimation is **0.0268759**.

```
p1<-Fullp ;q1 <- 1-p1 ;  n1<- 403
p2<-Assop; q2<- 1 - p2 ; n2<- 299
p3<-Assip; q3<-1-p3 ;   n3 <-598


pDiff12<- p1-p2
VarDiff12<- (p1*q1)/(n1-1) + (p2*q2)/(n2-1)
BDiff12 <- 2*sqrt(VarDiff12)
UpdDiff12 <- pDiff12 + BDiff12
LwBdDiff12 <- pDiff12 - BDiff12
```

b. We are going to consider a SRS of the first two strata: full professors and associate professors. We will first compute the expected difference. This is taken by subtracting the proportion of professors from strata one's sample and subtracting the proportion from strata 2's sample. This value is stored in `pDiff12`. So it follows that $E(\hat{p}_{diff}) = \hat{p}_1 - \hat{p}_2 = 0.322257$.

We will now compute a bound on the error of estimation. Note that since we do not have a relationship between N_i and n_i we must ignore our FPC Correction. This is where our line **4** follows.

$$\begin{aligned}
Var(\hat{p}_1 - \hat{p}_2) &= Var(\hat{p}_1) + Var(\hat{p}_2) && (5)\\
&= (1-\frac{n_1}{N_1})\frac{\hat{p}_1\hat{q}_1}{n_1-1} + (1-\frac{n_2}{N_2})\frac{\hat{p}_2\hat{q}_2}{n_2-1} && (6)\\
&= \frac{\hat{p}_1\hat{q}_1}{n_1-1} + \frac{\hat{p}_2\hat{q}_2}{n_2-1} && (7)
\end{aligned}$$

This formula is computed with the line `(p1*q1)/(n1-1) + (p2*q2)/(n2-1)`. We will now compute our bound on the error of estimation: $B = 2\sqrt{Var(\hat{p}_1 - \hat{p}_2)}$. The bound is computed in the line `BDiff12 <- 2*sqrt{VarDiff12}`. The bound on our error of estimation is: $B_{d12} = 0.0723548$.

We will now compute a confidence interval by using: $((\hat{p}_1 - \hat{p}_2) \pm B_{d12})$. The lower bound of our interval is computed by: `LwBdDiff12 <- pDiff12 - BDiff12`, the upperbound is `UpdDiff12 <- pDiff12 + BDiff12`. We know our interval is: (0.2499022,0.3946117). Since 0 is not in our interval, we can conclude that with statistical evidence at the 95 percent confidence level that there is a difference in the proportion of profs that believe the level affected effort in teaching. In particular since all the values in the interval are positive and we considered the difference between full and associate professors, it follows that full professors think in greater proportion that level of tenure impacts teaching ability.

```r
pDiff23<- p2-p3
VarDiff23<- (p2*q2)/(n2-1) + (p3*q3)/(n3-1)
BDiff23 <- 2*sqrt(VarDiff23)
UpdDiff23 <- pDiff23 + BDiff23
LwBdDiff23 <- pDiff23 - BDiff23
```

c. We are going to consider a SRS of the first two strata: full professors and associate professors. We will first compute the expected difference. This is taken by subtracting the proportion of professors from strata one's sample and subtracting the proportion from strata 2's sample. This value is stored in `pDiff12`. So it follows that $E(\hat{p}_{diff}) = \hat{p}_1 - \hat{p}_2 =$ -0.0986622.

We will now compute a bound on the error of estimation. Note that since we do not have a relationship between N_i and n_i we must ignore our FPC Correction. This is where our line **4** follows.

$$\begin{align} Var(\hat{p}_2 - \hat{p}_3) &= Var(\hat{p}_2) + Var(\hat{p}_3) \tag{8} \\ &= (1 - \frac{n_2}{N_2})\frac{\hat{p}_2\hat{q}_2}{n_2 - 1} + (1 - \frac{n_3}{N_3})\frac{\hat{p}_3\hat{q}_3}{n_3 - 1} \tag{9} \\ &= \frac{\hat{p}_3\hat{q}_3}{n_3 - 1} + \frac{\hat{p}_2\hat{q}_2}{n_2 - 1} \tag{10} \end{align}$$

This formula is computed with the line `(p3*q3)/(n3-1) + (p2*q2)/(n2-1)`. We will now compute our bound on the error of estimation: $B = 2\sqrt{Var(\hat{p}_2 - \hat{p}_3)}$. The bound is computed in the line `BDiff23 <- 2*sqrt{VarDiff23}`. The bound on our error of estimation is: $B_{d23} = 0.0681947$.

We will now compute a confidence interval by using: $((\hat{p}_2 - \hat{p}_3) \pm B_{d23})$. The lower bound of our interval is computed by: `LwBdDiff23 <- pDiff23 - BDiff23`, the upperbound is `UpdDiff23 <- pDiff23 + BDiff23`. We know our interval is: (-0.1668569,0.3946117). Since 0 is contained in the confidence interval, we do not have statistical evidence backing the claim that there is a difference in the proportion between associate and assistant professors.

d. We can safely say that all conclusions drawn are reliable. Consider part (A) where we used a stratified sample in which the unknown population was all professors and the sample size was $n = 1300 > 30$. Since this sample size is greater than 30, we can safely conclude that the Central Limit Theorem can be applied and our sample roughly follows a normal distribution. Therefore in the stratified case, we can apply our formulas for the bound and expected value. As for part (b) and (c), we used simple random samples from each of the strata to compare the differences in the true proportions acorss each strata. Since each strata's sample size is greater than 30, it follows that the Centeral Limit Therom applies and that our results can be considered reliable.

Question 2

```r
c1<-4 ; c2<-6;c3<-10
N<-10000
N1 <- N*.3
N2<- N*.25
N3<-.45*N

AlDenominator <-((N1*sqrt(p1*q1))/(sqrt(c1))) + ((N2*sqrt(p2*q2))/(sqrt(c2))) + ((N3*sqrt(p3*q3))/(sqrt
a1 <- ((N1*sqrt(p1*q1))/(sqrt(c1)))/AlDenominator
a2<- ((N2*sqrt(p2*q2))/(sqrt(c2)))/AlDenominator
a3<- ((N3*sqrt(p3*q3))/(sqrt(c3)))/AlDenominator
B<- 0.02;

#Below is the line computing the sample numerator
SampleNum<- ((N1^2*p1*q1)/(a1)) + ((N2^2*p2*q2)/(a2)) +((N3^2*p3*q3)/(a3))
```

```
#Below calculates the left side of the denominator
LeftDenom <- (B^2*N^2)/4
#Below computes the rightward sum of denominator
RightDenom <- (N1*p1*q1)+(N2*p2*q2)+(N3*p3*q3)
SampleSize <- SampleNum/(LeftDenom+RightDenom)
```

a. Note that our allocation for strata i is:

$$a_i = \frac{\frac{N_i\sigma_i}{\sqrt{c_i}}}{\sum_{k=1}^{L} \frac{N_K\sigma_K}{\sqrt{c_K}}}$$

We will now use an estimate for $\sigma_i \approx \sqrt{\hat{p}_i\hat{q}_i}$. We calculate the first allocation term for the full prof. strata with the function: `a1 <-  (N1*sqrt(p1*q1))/(sqrt(c1)))/AlDenominator`. We compute `AlDenominator`with the sum of the three respective numerators for each strata:`AlDenominator <-((N1*sqrt(p1*q1))/(sqrt(c1))) + ((N2*sqrt(p2*q2))/(sqrt(c2))) + ((N3*sqrt(p3*q3))/(sqrt(c3)))` The variables `c#` represent the cost of sampling a professor in the correspondingly numbered strata.

We will now consider our sample size calculation from the hefty formula below:

$$n \approx \frac{\sum_{k=1}^{L} \frac{N_k^2 p_k q_k}{a_i}}{\frac{B^2N^2}{4} + \sum_{K=1}^{L} N_k p_k q_k}$$

Note that our bound $B = 0.02$, $N = 10000$ where $N_1 = 3000$,$N_2 = 2500$,$N_3 = 4500$. The numerator in our sample size calculation is computed by the line: `SampleNum<- ((N1^2*p1*q1)/(a1)) + ((N2^2*p2*q2)/(a2)) +((N3^2*p3*q3)/(a3))`. Our value for the numerator is: 2.4299192×10^7. Now we will construct the denominator. The command `LeftDenom <- (B^2*N^2)/4` computes that value $(\frac{BN}{2})^2$ Our estimated sample size is: 1969.2674506. We will round up this value for optimal sampling. So our ideal sample size is $n \approx 1970$. We will now observe our allocations, note:

$$n_1 = a_1(n) \tag{11}$$
$$n_2 = a_2(n) \tag{12}$$
$$n_3 = a_3(n) \tag{13}$$

Our allocation is:

- Our sample of full professors will be: $737.1071148 = 737$
- Our sample of associate professors will be: $500.7601554 = 501$
- Our sample of assistant professors will be: $732.1327297 = 732$

Allocations are 0.374166, 0.254193 0.371641 for the variables a_1 a_2 and a_3 respectively.

```
NoCostDenom <- (N1*sqrt(p1*q1))+ (N2*sqrt(p2*q2))+ (N3*sqrt(p3*q3))
a1NoC <-(N1*sqrt(p1*q1))/NoCostDenom
a2NoC<- (N2*sqrt(p2*q2))/NoCostDenom
a3NoC<- (N3*sqrt(p3*q3))/NoCostDenom

LeftDenom <- (B^2*N^2)/4
#Below computes the rightward sum of denominator
RightDenomNoCost <- (N1*p1*q1)+(N2*p2*q2)+(N3*p3*q3)
SampleNumNoCost<- ((N1*sqrt(p1*q1))+ (N2*sqrt(p2*q2))+ (N3*sqrt(p3*q3)))^2
SampleSizeNoCost <- SampleNumNoCost/(LeftDenom+RightDenom)
n1NC <- a1NoC*SampleSizeNoCost
n2NC <- a2NoC*SampleSizeNoCost
n3NC <- a3NoC*SampleSizeNoCost
```

It may cost more to obtain information from an assistant professor than a full professor due to their job security. Full professors have tenure and therefore job security - they are more likely to be easily found in their office at their main institution. This is in contrast to Assistant professors that are starting out in thier careers and will naturally be more busy. Some of these assistant professors may even be an adjunct at another college so they may be less likely to be around their office and therefore harder to collect a sample.

b. Allocation will change slightly in addition we will observe for the most part that the sample size decreases. This can be because we have one less consideration or constraint (cost) and therefore do not need as many units in the sample. Since variability is changing across strata this is a Neyman allocation. Observe the allocations that disregard cost :

$$b_i = \frac{N_i\sigma_i}{\sum_{k=1}^{L} N_K\sigma_K} \approx \frac{N_i\sqrt{p_iq_i}}{\sum_{k=1}^{L} N_K\sqrt{p_kq_k}}$$

Our allocations without considering cost are stored in the variables `a1NoC`,`a2NoC`,`a3NoC` where the number corresponds to the strata. We will now consider out new sample size calculations.

$$n \approx \frac{\left(\sum_{k=1}^{L} N_k\sqrt{p_kq_k}\right)^2}{\frac{B^2N^2}{4} + \sum_{K=1}^{L} N_kp_kq_k} \tag{14}$$

Our newly computed sample size is: 1894.7251429 which can be rounded up to $n = 1895$ Our allocations new are:

$$n_1 \approx n(b_1) \tag{15}$$
$$n_2 \approx n(b_2) \tag{16}$$
$$n_3 \approx n(b_3) \tag{17}$$

These computations are computed by the lines: `n1NC <- a1NoC*round(SampleSizeNoCost)`, `n2NC <- a2NoC*round(SampleSizeNoCost)` and `a3NC <- a3NoC*round(SampleSizeNoCost)`. The allocations are:

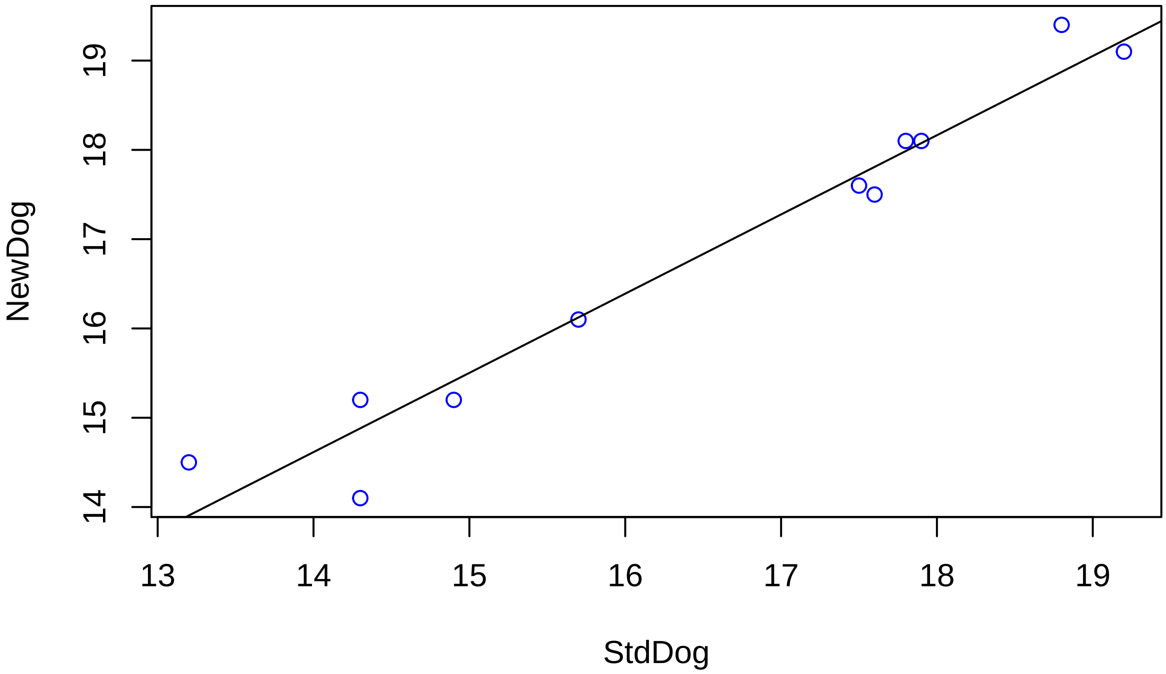
- $n_1 = 556.8610718$, which is about 557
- $n_2 = 463.331312$,which is about 463
- $n_3 = 874.5327591$,which is about 875

Allocations are 0.2939007, 0.2445375 0.4615618 for the variables b_1 b_2 and b_3 respectively. These allocations weigh assistant professors more highly. This is because we do not have to consider our cost variable where assistants were so costly! c. We now are considering allocation when we assume cost and variability is constant across strata. In other words we use proportional allocation. These proportions of the population are actually given by our problem statement. Allocations in this case are .3 .25 .45 respectively for strata 1 2 and 3. Our sample size for this case will be less than in our other cases from parts b and a. This is because we have even less variables to consider since both cost and variability are set to be constant across strata.

Question 3

```
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)

plot(StdDog,NewDog, col = "blue")
abline(lm(NewDog ~ StdDog))
```



```
lm(NewDog~StdDog)
```

```
##
## Call:
## lm(formula = NewDog ~ StdDog)
##
## Coefficients:
## (Intercept)       StdDog
##      2.1940       0.8872
```

a. Our scatterplot is printed above. We see a generally positive trend in our data with a slope slightly less than positive one. We see two clusters of data points in our plot, some in the bottom left corner and in the top right. This absence of times reordered from 15 to 16 seconds could be considered unusual - but is probably due to the smaller sample size given.

```
rhat <- mean(NewDog)/mean(StdDog) # r coeff.
Ndog<- 700
ndog <- 11
```

b. Our data here is appropriate for ratio sampling. We have a general dog population of $N = 700$ as well as a sample size $n = 11$ under each condition for new and standard training. We do have a substantial amount of information for the subsidiary variable, which is the dog's times under standard training. In particular the population average time under standard training is $\mu_x = 17$. From this information, and the data given, we can compute our r coefficient (since we can compute the sample means under each form of training). We will have to assume our sample can be applied with the central limit theorem, even though the size is 11.

c. Our $\hat{r}$ is 1.0204194, computed from the division of our sample means. In other words: $r = \frac{\hat{y}}{\hat{x}}$ where $\hat{x}$ is the sample mean time under standard training conditions. Additionally note that `ndog` is the sample size given

```
MuNew<- rhat*17
fpc <- (1-ndog/Ndog)
SrSqu<- (1/(ndog-1))*(sum((NewDog-rhat*StdDog)^2))
VarNewDog <- fpc*(SrSqu/ndog)
BoundYnew <-2*sqrt(VarNewDog)
```

c. We will use the fact that $\mu_{\hat{Y}} = r\mu_{\hat{X}}$. This is computed with the line of code `MuNew<- rhat*17`. Our expected value on the mean time for new training is: 17.3471302. Before we find a bound on the error of estimation, we will compute an important constant $S_r^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{r}x_i)^2}{n-1}$ Note that (x_i, y_i) are sample data points where i ranges from 1 to 11. This constant S_r^2 is computed by the line: `SrSqu<- (1/(ndog-1))*(sum((NewDog-rhat*StdDog)^2))`. It follows that $S_r^2 = \mathbf{0.2277407}$. We now want to compute a bound on the error of estimation of μ_y. Observe the derivation:

$$B_{\hat{\mu_y}} = 2\sqrt{Var(\hat{\mu_y})} \tag{18}$$

$$= 2\sqrt{\left(1 - (\frac{n}{N})\right)\frac{s_r^2}{n}} \tag{19}$$

$$= 2\sqrt{\left(1 - (\frac{11}{700})\right)\frac{s_r^2}{11}} \tag{20}$$

Our variance of $\hat{\mu_y}$ is computed by the line: `VarNewDog <- fpc*(SrSqu/ndog)`. The variance is 0.0203784. We then will square root this value and multiply it by 2 thru the line: `BoundYnew <-2*sqrt(VarNewDog)`.

It follows that, our bound is 0.2855055. We will be able to compute a confidence interval which can capture the true population mean of the time under New Training. It follows that $(\mu_y \pm B_{\mu_y})$. It follows that our confidence interval is **(17.0616247, 17.6326358)**. Since all values in our confidence interval are greater than $\mu_{\hat{x}} = 17$, we can say with 95% confidence that the average task time under new training is greater than 17. It follows we can say with 95% confidence that the new training actually poses an increase (albeit small) from the mean time under standard training.

```
TotalMuX <- 700*17
TotalMuY <- TotalMuX*rhat
BoundTotY <- 700*BoundYnew
```

d. We can estimate the total time required to complete the task under the new training. The general fromula is:

$$\hat{\tau_y} = \hat{r}\tau_x$$

We must know τ_x to utilize this formula. From our given information, since the populations mean time was $\mu_x = 17$ and $N = 700$ $\tau_x = 700 * 17 = 11900$. We will utilize our previous estimate for $\hat{r} = 1.0204194$ which is from part (c). It follows that our estimate for the total time under new training is $\tau_{\mu_{new}} = \hat{\tau_y} =$ `12142.99`. Note this was computed by the line: `TotalMuY <- TotalMuX*rhat`. We will now compute a bound on the error of estimation. Note our Bound formula where $N = 700$:

$$B_{\hat{\tau_y}} = 2\sqrt{Var(\hat{\tau_y})} \tag{21}$$

$$= 2\sqrt{N^2(1 - \frac{n}{N})\frac{S_r^2}{n}} \tag{22}$$

$$= N2\sqrt{(1 - \frac{n}{N})\frac{S_r^2}{n}} \tag{23}$$

$$= NB_{\hat{\mu_y}} \tag{24}$$

We will use our Bound from the previous part, the calculation is stores in the line: `BoundTotY <- 700*BoundYnew`. Our bound on the total time is: $B_{\tau_y} = 199.853871$.