Gawron_Homework7

Nicholas Gawron

3/9/2021

```
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.

Problems

Question 1

Refer to the scenario discussed in question 2 of homework 6. Suppose the average test score for the class is to be estimated again at the end of the school year. Find the Neyman allocation of a sample of size 50, using the previously collected data to estimate the variances.

```
ScoreData <- (readxl::read_excel('data/HW6Data.XLS'));

#reads in data based of track / stores number of sampling units per
Track1 <- (na.omit(ScoreData$`Track I`)); T1L <- length(Track1)
Track2 <-  (na.omit(ScoreData$TrackII)); T2L <- length(Track2)
Track3 <-  (na.omit(ScoreData$TrackIII)); T3L <- length(Track3)
N<- 55+80+65 # total population of 6th grade
N1<- 55
N2<- 80
N3<- 65

#computes the wieghted average with the stratifed formula
Xst <-(1/(N))*((N1*mean(Track1))+(N2*mean(Track2))+(N3*mean(Track3)));

#FPC For each Strata
fpc1 <- (1-(T1L/N1))  # Track 1 FPC
fpc2 <- (1-(T2L/N2))  # Track 1 FPC
fpc3 <- (1-(T3L/N3)) # Track 3 FPC

#variance
var1Est <- sd(Track1)
var2Est <- sd(Track2)
var3Est <- sd(Track3)
NeymanDenominator <- N1*var1Est+N2*var2Est+N3*var3Est
a1 <- (N1*var1Est)/ NeymanDenominator
a2 <- (N2*var2Est)/ NeymanDenominator
a3 <- (N3*var3Est)/ NeymanDenominator
```

We need not compute the expected value in this case. The Neyman allocation is given by:

$$a_i = \frac{N_i \sigma_i}{\sum_{k=1}^{L} N_k \sigma_k}$$

Note that we approximate $\sigma_i \approx s_i$ These variance approximations are stored in the variable `var1Est` for the first strata. We will now compute the denominator of the Neyman allocation, $\sum_{k=1}^{L} N_k \sigma_k$. Recall that N_i is the strta's population size stores as `N#` for `#` strata. The R code line `NeymanDenominator <- N1*var1Est+N2*var2Est+N3*var3Est` computes the denominator of the allocation term. We get the value 2456.9673969 as our denominator.

- The first strata's allocation term a_1 is equal to **0.2295373**. This was computed by the R code `(N1*var1Est)/ NeymanDenominator`.
- The first strata's allocation term a_2 is equal to **0.4095342**. This was computed by the R code `(N2*var1Est)/ NeymanDenominator`.
- The first strata's allocation term a_3 is equal to **0.3609285**. This was computed by the R code `(N3*var3Est)/ NeymanDenominator`.

We will now compute the sample sized for each strata through the calculation $n_i = na_i$. In which $n = 50$.

- We consider the formula in R code `50*a1` (11.4768627). The sample size for the first strata is: 11
- We consider the formula in R code `50*a2`(20.4767121). The sample size for the second strata is: 20
- We consider the formula in R code `50*a3`(18.0464251). The sample size for the third strata is: 18

Question 2

What are the three main factors that we must consider when allocating a sample to the L strata? Give a detailed description of why they are important and how they should be considered in the allocation process.

In terms of our objective, the best allocation scheme is affected by three factors:

1. The total number of elements in each stratum.

- The number of elements in each stratum affects the quantity of information in the sample. A sample size 30 from a population of 300 elements should contain more information than a sample of 30 from 30,000 elements. Thus, large sample sizes should be assigned to strata containing large numbers of elements.

2. The variability of observations within each stratum.

- Variability must be considered because a larger sample is needed to obtain a good estimate of a population parameter when the observations are less homogeneous.

3. The cost of obtaining an observation from each stratum.

- If the cost of obtaining an observation varies from each stratum, we take small samples from strata with high costs. We want to do this because our overall goal is to keep the cost of sampling at a minimum. So we want to wiegh strata with high costs the least.

Question 3

```
Xst <- .5*7.63+.1*7.74+.4*6.55
Sn1<- .15^2/1347 #Variance given from a standard SRS for first strata
Sn2<- .35^2/163  #Variance given from a standard SRS for second strata
Sn3 <-  .11^2/1095  #Variance given from a standard SRS for third strata
Bana<- 2*sqrt(.25*Sn1+.16*Sn2+.01*Sn3)
```

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

$$\bar{x}_{st} = \frac{1}{N}(N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3)$$

is the sample mean.

- **N** is the population size of all anesthesiologists
- N_i is the population of all anesthesiologists in job i
- i corresponds to the row (or strata) on the given table: 1 corresponds to Anesthesiologist, and 2 corresponds to Anesthesiologist Resident and so on
- $\bar{X}_i$ is the mean hours of all anesthesiologists in respective sample from strata i

We know from the problem statement that since "anesthesiologists (composing approximately 50% of the population),anesthesiology residents (composing approximately 10% of the population), and nurse anesthetists (composing approximately 40% of the population)", we can create a formula for N, the total population. Note, $N_1 = .5N$, $N_2 = .4N$,$N_3 = .1N$. We will now modify our expected value formula accordingly:

$$\bar{x}_{st} = \frac{1}{N}(.5N\bar{X}_1 + .4N\bar{X}_2 + .1N\bar{X}_3) = .5\bar{X}_1 + .4\bar{X}_2 + .1\bar{X}_3$$

This is computed in the first line of executable code in the chuck above. The expected value is: **7.209**.

We will now compute the bound on the error of estimation. Consider the formula where s_i is the i-th strata's sample standard deviation:

$$B = 2\sqrt{\frac{1}{N^2}\sum_{1}^{3} N_i^2\Big(1 - \frac{n_i}{N_i}\Big)\frac{s_i^2}{n_i}}$$

Since we cannot conclude an exact relationship between n_i and N_i we must ignore the FPC: $\Big(1 - \frac{n_i}{N_i}\Big)$. We will also modify the formula by considering our relationship from the above. Since $N_1 = .5N$, $N_2 = .4N$,$N_3 = .1N$, it follows that $N_1^2 = .25N^2$, $N_2^2 = .16N^2$,$N_3^2 = .01N^2$. After some substitution and moving variables around we result in:

$$B = 2\sqrt{\frac{1}{N^2}\sum_{1}^{3} N_i^2\frac{s_i^2}{n_i}} \tag{1}$$

$$= 2\sqrt{\frac{N^2}{N^2}\Big(.25\frac{s_1^2}{n_1} + .16\frac{s_2^2}{n_2} + .01\frac{s_3^2}{n_3}\Big)} \tag{2}$$

$$= 2\sqrt{.25\frac{s_1^2}{n_1} + .16\frac{s_2^2}{n_2} + .01\frac{s_3^2}{n_3}} \tag{3}$$

$$= 2\sqrt{.25(Sn_1) + .16(Sn_2) + .01(Sn_3)} \tag{4}$$

This will be computed in the final line of executable code in the chuck above. Intermediate lines store variables in the code. The bound on our error of estimation is $\mathbf{B = 0.0223188}$.

```
EDiff12<- 7.74 - 7.63
VarDiff12<- Sn1+ Sn2
BDiff12 <- 2*sqrt(VarDiff12)

EDiff23 <- 7.74 - 6.55
VarDiff23 <- Sn2 +Sn3
BDiff23<- 2*sqrt(VarDiff23)
```

b. We will construct two different confidence intervals to inspect the difference between the populations in hours doing work without a break across the strata of anesthesiologists. Note the strata all have the same numbering as the above problem.

We will first compute the difference between Anesthesiologist resident and Anesthesiologists. This will be done by first getting the expected value: $E(\bar{X}_2 - \bar{X}_1) = \bar{x}_2 - \bar{x}_1$. This is done through line one and stored as `EDiff12`. This value is: 0.11. We will now compute a bound on the error of estimation, note that $Var(X_2 - X_1) = Var(X_1) + Var(X_2)$ this is under the assumption that the strata are independent from one another, so we omit the covarriance term.

We will consider the calculations made in part A, `Sn1` corresponds to the variance of the sample of anesthesiologists (found from an SRS), whereas `Sn2` is the same for anesthesiologists residents. This variance of differences is stored as: `VarDiff12`.

We will now compute the bound on the error of estimation, $B = 2\sqrt{Var} \implies 2\sqrt{Var(X_1) + Var(X_2)}$ the code used to compute this will be `2*sqrt(VarDiff12)`, this will be stored as `BDiff12`. The value of the bound is: 0.0554342

We will now create a confidence interval of the form $((\bar{x_2} - \bar{x_1}) \pm B)$. We are 95% confident that the interval captures the true difference in the average of hours worked between anesthesiologists and anesthesiologists residents: (0.0545658, 0.1654342). Since $0 \notin$(0.0545658, 0.1654342), we have statistical evidence that there exists a difference between the average hours worked without a break of anesthesiologist and anesthesiologist residents. In particular, since all values of the interval are positive we can conclude that residents have a greater number of hours than general anesthesiologists. This is because we concluded $X_2 - X_1 > 0 \implies X_2 > X_1$ Note the code used to compute the interval is: `(EDiff12- BDiff12,EDiff12+BDiff12)`.

We will now reproduce the results above with nurse anesthetists and residents. We will first compute the difference between Anesthesiologist resident and Anesthesiologists nurses. This will be done by first getting the expected value: $E(\bar{X}_2 - \bar{X}_3) = \bar{x}_2 - \bar{x}_3$. This is done through line one and stored as `EDiff23`. This value is: 1.19. We will now compute a bound on the error of estimation, note that $Var(X_2 - X_3) = Var(X_2) + Var(X_3)$ this is under the assumption that the strata are independent from one another, so we omit the covarriance term.

We will consider the calculations made in part A, `Sn3` corresponds to the variance of the sample of anesthesiologists nurses (found from an SRS). This variance of differences is stored as: `VarDiff23`.

We will now compute the bound on the error of estimation, $B = 2\sqrt{Var} \implies 2\sqrt{Var(X_2) + Var(X_3)}$ the code used to compute this will be `2*sqrt(VarDiff23)`, this will be stored as `BDiff23`. The value of the bound is: 0.0552298.

The interval created is given by the code: `(EDiff23- BDiff23, EDiff23+BDiff23)`‘. We are 95% confident that the interval (1.1347702, 1.2452298) captures the true difference in the average hours worked. Since $0 \notin$(1.1347702, 1.2452298), we have statistical evidence that the there exist a difference between the average hours of anesthesiologist and anesthesiologist nurses. In particular, since all values of the interval are positive we can conclude that residents have a greater number of hours than nurse anesthesiologists. This is because $X_2 - X_3 > 0 \implies X_2 > X_3$

Question 4

a. One possible reason for the outlier is that Superman may have had this fight in a cheaper area of Metropolis. Since this fight was in a cheaper area of the town- the cost of damages would have been less than damages in the more expensive areas of Metropolis.
A second possible reason for this outlier could be that this fight on the randomly sampled Day 10 could have been against a weak opponent. This would mean the Superman could have beat the opponent very quickly and therefore not cause as much damage that day. This could contribute to the lower cost of damages - since there were less damages.

```
SupermanCosts <- c(89.3,72.1, 133.1, 126.3, 90.2, 81.8, 115, 119.5, 99.6, 12)
Xbar <- mean(SupermanCosts)
SupSd<-sd(SupermanCosts)
n<-10 #samples size
N<-477 #pop size
```

```
fpc <- 1-(n/N)
B <- 2*sqrt(fpc*(SupSd^2)/n)
tau <- N*Xbar
BTot <- N*B
```

b. This is a SRS in which a sample of 10 flights were taken of the total 477. Based off our sample, it follows that `Xbar <- mean(SupermanCosts)` computes our sample mean $\bar{x}$. Based of the sample, the mean cost of a fight comes out to be **93.89** in the thousands of dollars.

We will now compute the bound on our error of estimation. Consider the SRS formula for the bound: $B = 2\sqrt{(1-\frac{n}{N})\frac{s^2}{n}}$. We will compute the samples standard deviation with the code: `sd(SupermanCosts)` this will be stored in the variable `SupSd`. Note that the sample standard deviation $s =$ `SupSd`. Now we compute our bound: $2\sqrt{(1-\frac{10}{477})\frac{s^2}{10}}$ Our R code that is used to calculate this bound is: `B <- 2*sqrt(fpc*(SupSd^2)/n)`. It follows that our bound on the error of estimation is: **21.9607823** in the thousands of dollars.

c. We will now estimate for the total amount of damage caused through the formula, $\tau = N\bar{x}$. We have computed $\bar{x}$ in the previous part, and we know our population size is $N = 477$. It follows that we can estimate our total damage to be: $\tau =$ **44785.53** in the thousands of dollar. We will now place a bound on this total. Our bound is computed as: $B_\tau = 2N\sqrt{(1-\frac{n}{N})\frac{s^2}{n}} = NB$. This is computed with the code: `BTot<-N*B`. Our bound on the total cost of damages is: `10475.29`.

d. First off we have an issue with our sample size, namely 10, is not adequate based off the central limit theorem. This sample size is to small for us to invoke the central limit theorem since it is less than 30. As such any created confidence interval with the values from the previous parts b-c are not valid. Secondly, outliers in our sample may hinder the reliability of our data.

Question 5

```
SupermanCorrected <- c(89.3,72.1, 133.1, 126.3, 90.2, 81.8, 115, 119.5, 99.6, 120)
XbCor <- mean(SupermanCorrected)
SupSdCor<-sd(SupermanCorrected)
n<-10 #samples size
N<-477 #pop size
```

a. Since we found an error, our tenth fights cost would be **120** as an entry in the table above. This correction will increase the value of the mean. Why is this? - Well let us calculate the new sample mean. We will correct our data vector and then recalculated the mean with the command: `XbCor <- mean(SupermanCorrected)`. The corrected sample mean vaule is: **104.69**. We note that this value is greater than our previous sample mean: 93.89.

Note the math, consider n_{10} as our tenth data entry in the vector and n_c is the corrected term. Note that $n_c > n_{10}$

$$
\begin{aligned}
n_{10} &< n_c \\
\sum_{i=1}^{9} n_i + n_{10} &< \sum_{i=1}^{9} n_i + n_c \\
\sum_{i=1}^{10} n_i &< \sum_{i=1}^{9} n_i + n_c \\
\frac{\sum_{i=1}^{10} n_i}{10} &< \frac{\sum_{i=1}^{9} n_i + n_c}{10} \\
\bar{x} &< \bar{x}_{corrected}
\end{aligned}
$$

```
fpc <- 1-(n/N)
BCor <- 2*sqrt(fpc*(SupSdCor^2)/n);
```

b. Our bound on the error of estimation

$$B_{cor} = 2\sqrt{(1-\frac{n}{N})\frac{s_{cor}^2}{n}} \tag{5}$$

$$= 2\sqrt{(1-\frac{10}{477})\frac{s_{cor}^2}{10}} \tag{6}$$

This is stored in the R code, `BCor <- 2*sqrt(fpc*(SupSdCor^2)/n)`. Our corrected bound on the error of estimation for the mean cost per flight is: **13.0150618**. This bound on the error of estimation is less than our uncorrected bound of error of estimation, **21.9607823**. In other words $B_{cor} < B$. Note this is because:

$$s_{cor} < s \tag{7}$$

$$s_{cor}^2 < s^2 \tag{8}$$

$$\frac{s_{cor}^2}{n} < \frac{s^2}{n} \tag{9}$$

$$(1-\frac{10}{477})\frac{s_{cor}^2}{n} < (1-\frac{10}{477})\frac{s^2}{n} \tag{10}$$

$$\sqrt{(1-\frac{10}{477})\frac{s_{cor}^2}{n}} < \sqrt{(1-\frac{10}{477})\frac{s^2}{n}} \tag{11}$$

$$2\sqrt{(1-\frac{10}{477})\frac{s_{cor}^2}{n}} < 2\sqrt{(1-\frac{10}{477})\frac{s^2}{n}} \tag{12}$$

$$B_{cor} < B \tag{13}$$

```
tauCor <- N*XbCor # corrected estimate for total damages
BTotCor <- N*BCor # corrected bound of total damages
```

c. The fact we correct out outlier in part (a) increase the estimate for the total cost of repairs? Note that $\tau = N\bar{x}$ and $\tau_{cor} = N\bar{x}_{cor}$. Since we know from part (a) of this problem that $\bar{x} < \bar{x}_{\text{corrected}}$ it follows that:

$$\bar{x} < \bar{x}_{corrected} \tag{14}$$

$$N\bar{x} < N\bar{x}_{corrected} \tag{15}$$

$$\tau < \tau_{cor} \tag{16}$$

Our corrected estimate for the total estimate of damages is: `49937.13`.

We will now compute the corrected bound of the error of estimation with the bound. Recall from Question 4 that $NB_{\bar{x}} = B_{\tau}$. In addition from Part B of question 5 that $B_{\bar{x}_{cor}} < B_{\bar{x}}$. So it follows:

$$B_{\bar{x}_{cor}} < B_{\bar{x}}$$

$$NB_{\bar{x}_{cor}} < NB_{\bar{x}}$$

$$B_{\tau_{cor}} < B_{\tau}$$

We have mathematically found that our corrected bound for the total is less than the bound for the total computed in question 4. Our corrected bound on the error of estimation for the total damages is; **6208.1844586**. This value is less than our original bound from the incorrect data set: `10475.29`.

d. Our realization that we have an outlier in our data does not help that much in creating a confidence interval. A confidence interval in this class relies on the central limit theorem. We still have a sample size of 10 which does not help satisfy the central limit theorem for our large population of 477. As such any created confidence interval with the values from the previous parts b-c are not valid.