Gawron_Homework6

Nicholas Gawron

3/3/2021

```
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

A marketing research firm estimates the proportion of potential customers preferring a certain brand of lipstick by "randomly" selecting 100 women who come by their booth in a shopping mall. Of the 100 sampled, 65 women stated a preference for brand A.

a. We would estimate the true proportion of women preferring brand A by first calculating the proportion of women in the sample that had preferred the brand A. This would be done by computing $\hat{p} = \frac{65}{100} = .65$. We would then attempt to find a bound on the error of estimation B by considering the formula $B = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = 2\sqrt{\frac{.65(1-.65)}{99}} = 0.0958745$. Note we ignore the finite population correction, *fpc*, term since the target population of all potential customers $= N$, which is very large, adheres to the formula $100 * 20 \leq N$.

b. The target population for the study is all potential customers for a certain lipstick.
c. This is not a simple random sample, since all population elements did not have an equal likelihood of being selected to be a part of the sample. In particular this was a convince sample, since the sample consisted only of "women who came to the booth".
d. This type of sampling method introduced bias into the results. Since the sample consists of elements that came to the booth and people that were already in a mall- these customers may be more likely to prefer certain brands advertised by this booth.

Question 2

A school desires to estimate the average score that may be obtained on a reading comprehension exam for students in the sixth grade. The school's students are grouped into three tracks, with the faster learners in track I, the slower learners in track III, and the rest in track II. The school decides to stratify on tracks because this method should reduce the variability in test scores. The sixth grade contains 55 students in track I, 80 in track II and 65 in track III. A stratified random sample of 50 students is proportionally allocated and yields simple random samples of $n_1 = 14$, $n_2 = 20$ and $n_3 = 16$ from tracks I, II and III respectively. The data is contained in the accompanying excel file.

```r
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
Varst <- (1/(N)^2)*(N1*(fpc1*sd(Track1)^2/T1L)+ N2*(fpc2*sd(Track2)^2/T2L)+N3*(fpc3*sd(Track3)^2/T3L) )
BdT <- 2*sqrt(Varst)
```

a. The average score for the 6th grade comes from stratified random sample. We consider the average score with the excepted value under the stratified sample comes from:

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

- **N** is the population size of the 6th grade
- N_i is the population of all students in Track i
- $\bar{X}_i$ is the mean score of all students in respective sample from Track i

The average score is computed with: `(1/(N))*((N1*mean(Track1))+(N2*mean(Track2))+(N3*mean(Track3)));`

Note this is the stratified formula.

- Note that the value comes out to: $\bar{x}_{st} = 59.9886161$.

Next the bound on the error of estimation B comes from the formula below, where s_i is the sample standard deviation:

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

Observe that all the FPC values are less than 0.95, so they must be included in out calculations. Note the FPC values are stored in the above chuck under `fpc#` where the pound sign corresponds to the track number of the sample.

- Track One: $0.7454545 < 0.95 \implies$ Include this value
- Track Two: $0.75 < 0.95 \implies$ Include this value
- Track Three: $0.7538462 < 0.95 \implies$ Include this value

Our variance for the stratified sample is:

```
(1/(N)^2)*(N1*(fpc1*sd(Track1)^2/T1L)+N2*(fpc2*sd(Track2)^2/T2L)+ N3*(fpc3*sd(Track3)^2/T3L))
```

Our bound on the error of estimation is: B= 0.3677677. We can construct the confidence interval: (59.6208483,60.3563838). Since 70 is not contained in the interval, we can conclude with 95% confidence that the average student did not pass the reading comprehension test.

```
TDiff <- mean(Track1) - mean(Track2)
VT1 <- (N1^2*fpc1*(sd(Track1)^2/T1L))/N^2
VT2 <- (N2^2*fpc2*(sd(Track2)^2/T2L))/N^2
VDiff <- VT1 + VT2
BdDiffT <- 2*sqrt(VDiff)
```

b. We consider the mean of difference across the random sample's means for track 1 and track 2. The variance for each sample is calculated using the general formula from the SRS section: $Var_i = N_i^2 * fpc_i * \frac{s_i^2}{n}$ The variance for the difference of the two track's is computed by $Var_1 + Var_2$. This is computed by `VDiff <- VT1 + VT2`. Lastly the bound is computed by: `BdDiffT <- 2*sqrt(VDiff)` Consider the confidence interval: (12.6211494, 17.307422). Since 0 is not contained in the interval, we can conclude that there is a significant statistical difference between the two tracks at the 95% confidence interval. We will further note that since all values of the interval are positive reals than we can also conclude that the scores of Track 1 students is higher than that of Track 2 students.

Question 3

We will now look at cavities.

```
n <-10;N<- 400;
Cavities <- c(1 ,4 ,1 ,0 ,3 ,2, 4 ,0 ,3 ,2)
Ybar <- mean(Cavities)
Sd <- sd(Cavities)
fpc<- 1-(n/N)
Bcav <- 2*sqrt(fpc*(Sd^2)/n)
```

a. The mean number of cavities for all children in the group is 2. Note that the mean is computed by the R function `mean(Cavities)`.

b. To create an approximate 95% confidence interval for the mean we must find a bound for the error B. We will consider the FPC since $1 - \frac{n}{N} = 1 - \frac{10}{400} = .975$. So the calculation of the bound comes from s as the sample standard deviation and $n = 10$:

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

So B = 0.9309493. So our 95% confidence interval is formed by $(\bar{x} \pm B) \implies (\bar{x} - B, \bar{x} + B)$. We are 95% confident that the true mean is captured inside the interval: (1.0690507, 2.9309493). Since 2.2 is in the open interval, we do not have statistical evidence that the new toothpaste prevents cavities.

c. This is some concern here, our sample size is only of size 10 - which brings up issues related to the central limit theorem. We cannot apply this theorem to get normality of X_1 which is what we need to compute a reliable confidence interval.

Question 4

```r
Abv <-  c(48 ,48.7 ,50.1 ,43.3 ,47.5 ,49.4 ,39.9 ,52 ,46.7 ,50.5,45.6 ,49.7 ,45.3 ,46.9 ,48.5)
Bel <- c(37.8, 45, 44.2, 60, 54.2, 56.4, 59.3, 44.4, 41.8, 52.9,45.7, 57, 48.1, 58.2, 42.5, 41.1)
AllTemp <- c(Abv,Bel)

#part a
meanAll <- mean(AllTemp ) # expected value of All temps
sdA<- sd(AllTemp)
BdAll<- 2*sqrt(sdA^2/length(AllTemp)); # computes the Boundary
```

a. We will ignore the FPC for the simple random sample since the total population is all backyard pools which is trivially larger than 20 times our sample size of $n = 31$. The sample mean for an SRS is given by: 48.4096774. The standard deviation for the sample is calculated to be 5.7234229. The boundary, given by the formula: $B = 2\sqrt{\frac{s^2}{n}}$ computes to 2.0559142. We are 95% confident that the true mean backyard pool temperature is captured by the interval: **(46.3537632,50.4655916)**. Since $50 \in (46.3537632, 50.4655916)$, there is not enough evidence to suggest the average pool temperature is different from the recommended 50 degrees.

```r
  Yst <-  mean(mean(Abv),mean(Bel))
StrataBd <- sqrt((sd(Abv)^2/length(Abv)) + (sd(Bel)^2/length(Bel)) )
LwBD <- Yst - StrataBd
UpBD <- Yst + StrataBd
```

b. We will now consider a stratified sample. Note that `Yst` is the expected value for the stratified sample. Consider the formula where N_1 corresponds to the population of above ground pools, N_2 for below ground, and N is the total of all pools:

$$\bar{Y}_{st} = \frac{1}{N}(N_1\bar{Y}_1 + N_2\bar{Y}_2)$$

It is important to note from the problem statement that the "the population of backyard pools is evenly split" across strata. In other words $\frac{N}{2} = N_1 = N_2$, so our above formula can be simplified to:

$$\bar{Y}_{st} = \frac{1}{N}(N_1\bar{Y}_1 + N_2\bar{Y}_2) = \frac{1}{N}(N_1\bar{Y}_1 + N_1\bar{Y}_2) = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)$$

This formula is computed by: `mean(mean(Abv),mean(Bel))`. So the average pool temperature in the stratified random sample is: 47.4733333.

We will now find the bound on the error of estimation. Since the population of above and below ground pools is reasonably much larger than 20 times either sample size, $n_1 = 15$ and $n_2 = 16$ respectively, we can safely ignore FPC. This also aids in our calculations since we do not have a particular value for N_1, N_2 or N. Our formula for the bound is:

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

Note that s_i stands for the ith strata's standard deviation and n_i corresponds to the sample size taken in the ith strata. Recall from above that $\frac{N}{2} = N_1 = N_2$, it follows that when we square the expression $N_1^2 = N_2^2 = \frac{N^2}{4}$. It follows that our bound can be simplified to:

$$B = 2\sqrt{\frac{1}{N^2}\sum_1^2 \frac{N^2}{4} * \frac{s_i^2}{n_i}} = 2\sqrt{\frac{1}{4}\sum_1^2 \frac{s_i^2}{n_i}} = \sqrt{\sum_1^2 \frac{s_i^2}{n_i}}$$

This formula for the bound is computed in the code chuck above and stored as `StrataBd`. So our bound on the error of estimation is: **2.0167036**.

We can create the confidence interval be using the form `(LwBD,UpBD)`, where the Lower bound was computed by subtracting the bound from the average value, and the upper bound was calculated with the average value plus the bound. Since 50 is not contained in the open interval (45.4566297 , 49.490037), we can say with 95% confidence that the average pool temperature is different from the recommended value of 50.

c. Treating the sample as stratified does improve our estimation of the average temperature. This can be seen by the improvement lowering in the bound on when we used a stratified sample. As seen above when we conisdered a random sample, we have a bound of 2.0559142, and under a stratified sample we have a smaller bound of 2.0167036 which is an improvement.

```
ExpVal <- mean(Abv) - mean(Bel)
varA <- (sd(Abv)^2/length(Abv))
varB<- (sd(Bel)^2/length(Bel))

VarDiff <- varA + varB
BdDiff <- 2*sqrt(VarDiff)
LDiff <- ExpVal - BdDiff
UDiff <- ExpVal + BdDiff
```

d. We can compare differences by considering the difference of the random variables X_1 and X_2 which represnt above and below ground pools respectively. We compute the expected value of the difference by computing the difference of the expected values thru the formula: `ExpVal <- mean(Abv) - mean(Bel)`. Note that the expected value is: -1.8141667. We will now compute the bound on the error of estimation with the general formula $B = 2\sqrt{Var(X_1 - X_2)}$ By properties of variance $Var(X_1 - X_2) = Var(X_1) + Var(X_2)$, which can be computed by R commands of the same name: `varA + varB`. We will treat each strata as a distinct sample and therefore use the SRS formula for variance to compute each variance. For example, `varA <- (sd(Abv)^2/length(Abv))` is used to compute the variance for the above ground pools. The bound on the error of estimation is, $B = 4.0334073$. The confidence interval was created in a similar fashion from above: (-5.8475739, 2.2192406). Since 0 is contained in the interval, we cannot conclude that there is a statistically significant difference in above and and in ground pool temperatures.
   We can not be sure about the reliability of this solution until we are able to determine if the samples are independent of one another. If they were not independent our variance calculation would need to be factored into the value of our variance. We also have small sample sizes for each strata (that are < 30) which makes applying the CLT more difficult. It is hard to generalize that the two samples and the random variable $X_1 - X_2$ are normally distributed.