Project 3 - MC DATA 101

Juan J. Nunez

May 09, 2018

First I'm going to start by turning an SPSS file (i.e., .sav) into an RMarkdown file (i.e., .Rda). The data I will be using is from the World Bank. For more information on how to download World Bank data, please visit https://data.worldbank.org

Once the data is downloaded to a .SAV file, it is easy to use it using R by bringing up the 'foreign' package.

```
setwd("C:\\Users\\Juan Nunez\\Desktop\\MC_DATA_101\\ASSIG_3_DATA101")
```

Now I look at the dimensions of the ASSIG3_DATA data frame.

```
library(foreign)
NEWLES <- read.spss("C:\\Users\\Juan Nunez\\Desktop\\MC_DATA_101\\ASSIG_3_DATA101\\LES101.SAV", use.valu

ASSIG3_DATA <- NEWLES

save(ASSIG3_DATA, file = "ASSIG3_DATA.Rda")

dim(ASSIG3_DATA)
```

```
## [1] 148 356
```

I see that there are 148 rows and 356 columns.

Codebook:

V1 COUNTRY NUMBER ; V2 ABBREVIATED COUNTRY NAME ; V3 COUNTRY NAME ; V5 % ADULT FEMALE ILLITERACY 1990 ; V12 ENERGY CONSUMPTION/CAPITA 1991 ; V14 INFANT MORTALITY RATE 1991 ; V168 FEMALE SECODARY SCHOOL ENROLLMENT GROSS 1980 ; V133 CIVIL LIBERTIES 1991 ; V188 WORLD AS 5 REGIONS ;

Now I can begin by doing a simple hypothesis test using one sample.

```
INFMORTCLEAN <-  na.omit(ASSIG3_DATA$V14)
mean(INFMORTCLEAN)
```

```
## [1] 53.69565
```

```
sd(INFMORTCLEAN)
```

```
## [1] 43.06973
```

```
firstttest <- t.test(ASSIG3_DATA$V14, alternative = "two.sided", mu = 50, conf.int = 0.95)
firstttest
```

```
##
##  One Sample t-test
##
## data:  ASSIG3_DATA$V14
## t = 1.008, df = 137, p-value = 0.3152
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  46.44572 60.94559
## sample estimates:
## mean of x
```

```
##  53.69565
```

Based on the information I see above, I fail to reject the null hypothesis that the mean is equal to 50. Let's calculate the t statistic ourselves:

The numerator is the true mean divided by the value that determines the null hypothesis. The denominator is the standard deviation of the variable divided by the square root of the sample size.

```
(53.69565 - 50) / (sd(INFMORTCLEAN) / sqrt(length(INFMORTCLEAN)))
```

```
## [1] 1.007995
```

Everything seems to be in order.

Now we will do what is called an independent sample t test. The variables we are going to be working with are a dummy of the variable civil liberties where 1 is those cases with civil liberties that are equal to or higher than the mean. Infant mortality rate will be the other variable. We have to make sure to use listwise deletion for this to work.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
SMALL_DATA <- select(ASSIG3_DATA, V14, V133)

SMALL_DATA <- na.omit(SMALL_DATA)

mean(SMALL_DATA$V133)
```

```
## [1] 3.882353
```

```
SMALL_DATA <- mutate(SMALL_DATA, CIV_DUM = as.factor(SMALL_DATA$V133 >= mean(SMALL_DATA$V133)))

SMALL_DATA <- select(SMALL_DATA, -V133)

head(SMALL_DATA)
```

```
##   V14 CIV_DUM
## 1  28    TRUE
## 2  64    TRUE
## 3 130    TRUE
## 4  25   FALSE
## 5   8   FALSE
## 6   8   FALSE
```

```
class(SMALL_DATA$V14)
```

```
## [1] "numeric"
```

```
class(SMALL_DATA$CIV_DUM)
```

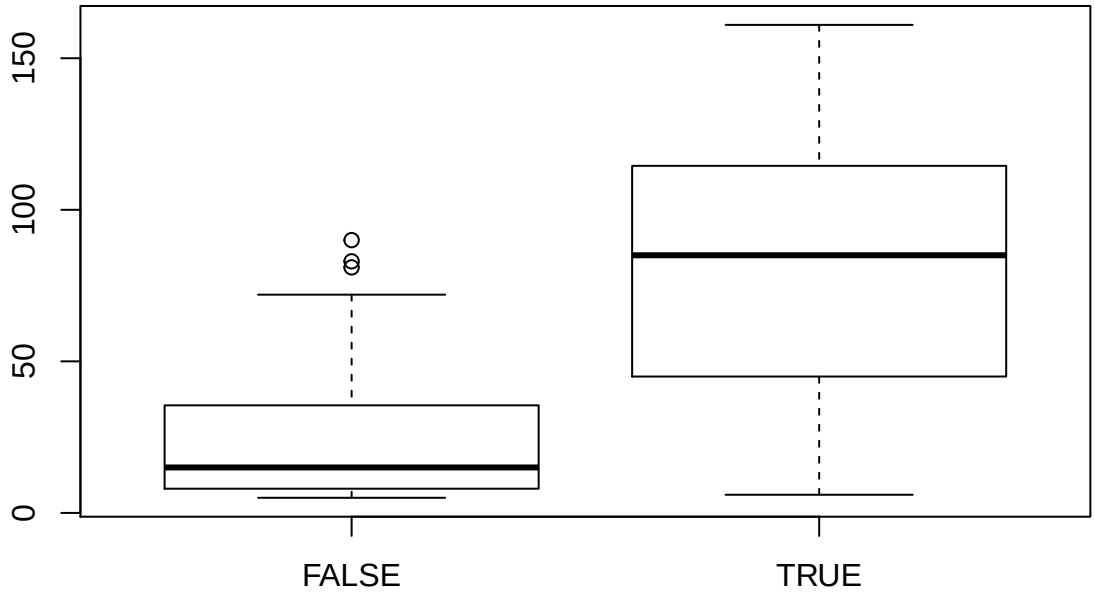
```
## [1] "factor"
```

```r
levels(SMALL_DATA$CIV_DUM)
```

```
## [1] "FALSE" "TRUE"
```

Now that I have my two variables, I can check if the means of infant mortality rate for the cases above or below the mean on civil liberites are significantly different. I can do this by using t test again. First I look at this association using a boxplot.

```r
boxplot(SMALL_DATA$V14 ~ SMALL_DATA$CIV_DUM)
```



Now we can use statistics to test whether there is or not a difference in infant mortality rate between the two groups of civil liberties.

H0 = Mean infant mortality rate of those countries that are above the mean of civil liberties is equal to the mean infant mortality rate of those countries that are below the mean of civil liberties. We use a two-sided test to test this hypothesis assuming non-equal variances. We do not assume that the variance between the FALSE and the TRUE groups is equal because of theory. We say unpaired because the countries being evaluated are not the same (i.e., the variables are not measured more than once for each case).

```r
t.test(SMALL_DATA$V14~SMALL_DATA$CIV_DUM, mu = 0, alt = "two.sided", conf.int = .95, var.eq=FALSE, pair
```

```
##
##  Welch Two Sample t-test
##
## data:  SMALL_DATA$V14 by SMALL_DATA$CIV_DUM
## t = -9.6247, df = 114.46, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
##  -67.97256 -44.76864
## sample estimates:
## mean in group FALSE  mean in group TRUE
##             24.8125             81.1831
```

Based on the p-value above, we reject the null hypothesis and find evidence to support the alternative hypothesis that one mean is significantly higher than the other. If we didn't have theory to guide us, we would have to do a few checks to see if we assume that the variances are equal or not. First we would look at the boxplot above. Then we would look at the variance of the two groups. Finally, we use Leven's test of equal variance where H0 = The population variances are equal.

```
var(SMALL_DATA$V14[SMALL_DATA$CIV_DUM=="TRUE"])
```

```
## [1] 1635.037
```

```
var(SMALL_DATA$V14[SMALL_DATA$CIV_DUM=="FALSE"])
```

```
## [1] 541.1769
```

We see that the variances are not equal. Let's look at the Leven's test. To use Leven's test, we have to use the package called 'car'.

```
##install.packages("car")
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.4
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.4.4
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
leveneTest(SMALL_DATA$V14~SMALL_DATA$CIV_DUM)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   1  26.181 1.232e-06 ***
##       117
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we can see that with a small p-value, we reject the null hypothesis and conclude that the variances between the two groups are not equal. So we use the non-equal variance option for our analysis.

Let's suppose we want to take samples from our population and then measure if there is a difference between the mean of the sample means and the population mean. What I'm going to do is first create a function that does the sampling, and then compare the means. There are several ways of taking a random sample from the dataframe. I will show two below.

```
set.seed(2000)
n = 100
index <- sample(1:nrow(SMALL_DATA), n, replace = TRUE)
sample_SMALL_DATA<-SMALL_DATA[index, ]
```

```
SampleTake = function(df,n) {
   return (SMALL_DATA[sample(nrow(df), n,  replace = TRUE),])
}

SMALL_SAMPLE<-SampleTake(SMALL_DATA, 100)
```

Now That I have my samples, I will do a t-test on the infant mortality rate variable.

```
MORTSAMP<-sample_SMALL_DATA$V14
MORTttest <- t.test(MORTSAMP, alternative = "two.sided", mu = 54, conf.int = 0.95)
MORTttest
```

```
##
##  One Sample t-test
##
## data:  MORTSAMP
## t = 0.79769, df = 99, p-value = 0.427
## alternative hypothesis: true mean is not equal to 54
## 95 percent confidence interval:
##  48.36257 67.21743
## sample estimates:
## mean of x
##     57.79
```

```
mean(SMALL_DATA$V14)
```

```
## [1] 58.44538
```

```
mean(MORTSAMP)
```

```
## [1] 57.79
```

We see above that the mean of V14 (Infant Mortality Rate) of the sample is 57.79 and of that of the population is 58.45. The confidence interval for the ttest is from 49.68 to 67.36. In this occassion, I fail to reject the null hyothesis (t = 0.8, p = 0.43). The difference between the population mean and the sample mean is not significant. I want to know use the CIVDUM variable as a dummy. Remember that true is those countries above the mean, meaning those that have low civil liberties.

```
table(sample_SMALL_DATA$CIV_DUM)
```

```
##
## FALSE  TRUE
##    39    61
```

There are 63 countries that have the same or more (which is less) civil liberties than the mean and there are 37 countries that are below the mean (in this case this means more liberites).

```
library(dplyr)
  CIVSAMP<-sample_SMALL_DATA %>% group_by(CIV_DUM)
  CIVSAMP%>%summarise(mean(V14))
```

```
## # A tibble: 2 x 2
##   CIV_DUM `mean(V14)`
##   <fct>         <dbl>
## 1 FALSE          18.4
## 2 TRUE           83.0
```

As we could predict, the countries that have less civil liberties have more infant mortality. The countries that have less civil liberties have more infant mortality. Remeber that TRUE is less civil liberties. We can use the

t-test to determine

```
BADCIV<-sample_SMALL_DATA%>%filter(CIV_DUM == "TRUE")%>%select(V14)
GOODCIV<-sample_SMALL_DATA%>%filter(CIV_DUM == "FALSE")%>%select(V14)

t.test(BADCIV, GOODCIV, mu = 0, alt = "two.sided", conf.int = .95, var.eq=FALSE, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  BADCIV and GOODCIV
## t = 10.527, df = 80.752, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  52.35425 76.75966
## sample estimates:
## mean of x mean of y
##  82.96721  18.41026
```

We see that we reject the null hypothesis and find evidence to support the alternative hypothesis. In this situation the difference in the mean of infant mortality rate between the countries with high and low civil liberties is significant. Let's look at the population.

```
  CIVDAT<-SMALL_DATA %>% group_by(CIV_DUM)
  CIVDAT%>%summarise(mean(V14))
```

```
## # A tibble: 2 x 2
##   CIV_DUM `mean(V14)`
##   <fct>         <dbl>
## 1 FALSE          24.8
## 2 TRUE           81.2
```

```
  BADCIV2<-SMALL_DATA%>%filter(CIV_DUM == "TRUE")%>%select(V14)
GOODCIV2<-SMALL_DATA%>%filter(CIV_DUM == "FALSE")%>%select(V14)

t.test(BADCIV2, GOODCIV2, mu = 0, alt = "two.sided", conf.int = .95, var.eq=FALSE, paired=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  BADCIV2 and GOODCIV2
## t = 9.6247, df = 114.46, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  44.76864 67.97256
## sample estimates:
## mean of x mean of y
##   81.1831   24.8125
```

We see that the population also has a signficant different infant mortality rate for the countries that are high or low in civil liberties. Now let's do a loop to figure out how many of our sample means are in the confidence interval.

```
library(dplyr)
LIT <- select(ASSIG3_DATA, V133, V5)

LIT <- na.omit(LIT)
head(LIT)
```

```
##   V133 V5
## 1    7 86
## 3    4 55
## 4    7 72
## 5    3  5
## 7    1  2
## 8    1  2
```

```
str(LIT)
```

```
## 'data.frame':    101 obs. of  2 variables:
##  $ V133: num  7 4 7 3 1 1 5 1 4 5 ...
##  $ V5  : num  86 55 72 5 2 2 78 2 84 75 ...
##  - attr(*, "variable.labels")= Named chr  "COUNTRY NUMBER" "ABBREVIATED COUNTRY NAME" "COUNTRY NAME"
##   ..- attr(*, "names")= chr  "V1" "V2" "V3" "V4" ...
##  - attr(*, "na.action")=Class 'omit'  Named int [1:47] 2 6 9 11 18 32 33 39 40 44 ...
##   .. ..- attr(*, "names")= chr [1:47] "2" "6" "9" "11" ...
```

```
count = 0
lit_mean = mean(LIT$V5)
set.seed(2000)
 n<-100
 m<-100

for (i in 1:m){
          index2<-sample(1:nrow(LIT), n, replace = TRUE)
          NEWSAMPLE<-LIT[index2,]
          V5<-NEWSAMPLE$V5
          result<-t.test(V5)



          lower_bound_conf<-result$conf[1]
          upper_bound_conf<-result$conf[2]



          if (lower_bound_conf <= lit_mean && lit_mean <= upper_bound_conf){
            count = count +1
          }
}


count
```

```
## [1] 93
```

We see that 93 of the 100 samples have confidence intervals that contain the true mean of the population.