Assignment3

Alka Santosh Naik

2023-09-19

# Preprocessing

#Alka Santosh Naik  
rm(list=ls())  
library(rio)  
library(moments)  
data=import("6304 Module 3 Assignment Data.xlsx",sheet="midwest")

## New names:  
## • `` -> `...1`

colnames(data)=tolower(make.names(colnames(data)))  
il\_sub=subset(data,data$state=="IL")  
mi\_sub=subset(data,data$state=="MI")  
il\_sample=il\_sub[sample(1:nrow(il\_sub),22),]  
mi\_sample=il\_sub[sample(1:nrow(mi\_sub),22),]  
set.seed(16999752)

# Analysis

1.Show the results of the str() command on both your primary data sets.

2.Using your Illinois 22-county sample construct a 98% confidence interval on the poptotal variable. Give an interpretation of the interval understandable to a lay person with little knowledge of statistics.

str(il\_sample)

## 'data.frame': 22 obs. of 5 variables:  
## $ ...1 : num 33 83 31 66 47 27 79 3 42 4 ...  
## $ county : chr "HAMILTON" "SALINE" "GREENE" "MERCER" ...  
## $ state : chr "IL" "IL" "IL" "IL" ...  
## $ area : num 435 383 543 561 321 ...  
## $ poptotal: num 8499 26551 15317 17290 39413 ...

str(mi\_sample)

## 'data.frame': 22 obs. of 5 variables:  
## $ ...1 : num 14 5 19 58 35 66 41 57 27 49 ...  
## $ county : chr "CLINTON" "BROWN" "DE KALB" "MACON" ...  
## $ state : chr "IL" "IL" "IL" "IL" ...  
## $ area : num 474 306 634 581 178 ...  
## $ poptotal: num 33944 5836 77932 117206 5189 ...

results=t.test(il\_sample$poptotal,conf.level=0.98)  
results

##   
## One Sample t-test  
##   
## data: il\_sample$poptotal  
## t = 3.0404, df = 21, p-value = 0.006219  
## alternative hypothesis: true mean is not equal to 0  
## 98 percent confidence interval:  
## 9408.768 100039.232  
## sample estimates:  
## mean of x   
## 54724

Analysis: From the above data we can say that we are 98% confident that the true population mean of Illinois population falls in the interval: -305533.0 854008.1

3.Referencing Part 1 above, does the mean county population for all 102 counties in Illinois fall within your confidence interval?

mean(il\_sub$poptotal)

## [1] 112064.7

Analysis: Yes the mean county population for all 102 counties in Illinois fall within your confidence interval: -305533.0 854008.1

4.Using your Illinois 22-county sample, can you say (α = .05) that the population mean for all Illinois counties is greater than 90,000?

results2=t.test(il\_sample$poptotal,mu=90000,alternative = c("greater"))  
results2

##   
## One Sample t-test  
##   
## data: il\_sample$poptotal  
## t = -1.9599, df = 21, p-value = 0.9683  
## alternative hypothesis: true mean is greater than 90000  
## 95 percent confidence interval:  
## 23752.29 Inf  
## sample estimates:  
## mean of x   
## 54724

Analysis: As the p-value = 0.2163, we reject the null hypothesis and accept the alternating hypothesis and conclude the true mean is greater than 90,000.

5.Using data on all the Illinois counties, determine how many standard deviations from Illinois’ true county population mean is the total population of White County, Illinois?

6.Consider your Michigan 22-county sample and the poptotal variable. In a simple t.test command (two tailed) what value of mu would result in p = .10.

result5=t.test(il\_sub$poptotal,mu=16522,alternative = c("two.sided"))  
result5

##   
## One Sample t-test  
##   
## data: il\_sub$poptotal  
## t = 1.8881, df = 101, p-value = 0.06189  
## alternative hypothesis: true mean is not equal to 16522  
## 95 percent confidence interval:  
## 11681.6 212447.9  
## sample estimates:  
## mean of x   
## 112064.7

width=result5$conf.int[2]-result5$conf.int[1]  
width

## [1] 200766.2

#answer 6

mean(mi\_sample$poptotal)

## [1] 92593.59

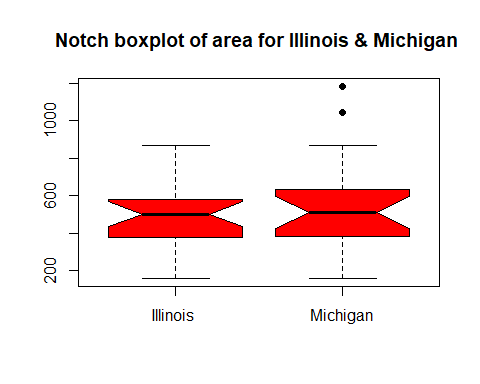
results3=t.test(mi\_sample$poptotal,mu=31500,alternative = c("two.sided"))  
results3

##   
## One Sample t-test  
##   
## data: mi\_sample$poptotal  
## t = 1.5256, df = 21, p-value = 0.142  
## alternative hypothesis: true mean is not equal to 31500  
## 95 percent confidence interval:  
## 9312.565 175874.617  
## sample estimates:  
## mean of x   
## 92593.59

Analysis: In a simple t.test command (two tailed) the value of mu should be aprox 31500 for p = .10.

7.Show comparative notched boxplots of the area of counties in your primary (n=22) Michigan and Illinois data frames. Your boxplots should be displayed side by side in a single graphic. Include labels on the x axis identifying the states included. Include a main title appropriately labeling the graphic. Color your boxplots red. Do these plots indicate a possible difference between the population medians all Illinois counties versus Michigan counties?

boxplot(il\_sample$area,mi\_sample$area,pch=19,  
 names=c("Illinois ","Michigan"),  
 col="Red",  
 notch=TRUE,  
 main="Notch boxplot of area for Illinois & Michigan ")

 As from the notch Analysis: boxplot we can see the notch of Illinois and Michigan are at same level or overlap each other, which indicates no possible difference between the population medians of Illinois counties versus Michigan counties

8.Using your primary (n=22) samples of Michigan and Illinois counties, does there appear to be a statistically significant difference (α = .05) between the population mean land area for Illinois counties versus Michigan counties?

results4=t.test(il\_sample$area,mi\_sample$area)  
results4

##   
## Welch Two Sample t-test  
##   
## data: il\_sample$area and mi\_sample$area  
## t = -0.82932, df = 34.589, p-value = 0.4126  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -189.28348 79.52075  
## sample estimates:  
## mean of x mean of y   
## 490.5327 545.4141

Analysis: From the above result we fail to reject null hypothesis at (α = .05) as there is not enough evidence that true difference in means is not equal to zero.