DSCC/CSC/STAT 462 Assignment 3

Due October 20, 2022 by 11:59 p.m.

Qirong Huang

Please complete this assignment using `RMarkdown`, and submit the knitted PDF. *For all hypothesis tests, state the hypotheses, report the test statistic and p-value, and comment on the results in the context of the problem.*

In order to run hypothesis tests and construct confidence intervals, you may find the `z.test` and/or `t.test` functions in `R` to be useful. For documentation, run `?z.test` and/or `?t.test` in the console.

1. Recently there has been much concern regarding fatal police shootings, particularly in relation to a victim's race (with "victim" being used generally to describe the person who was fatally shot). Since the start of 2015, the Washington Post has been collecting data on every fatal shooting in America by a police office who was on duty. A subset of that data is presented in the dataset "shootings.csv."
    a. Construct a two-sided 85% confidence interval "by-hand" (i.e. do not use the `t.test()` function, but still use `R`) on the mean age of victims. Interpret the result.

```
shooting<-read.csv("shootings.csv")
#sample mean
xbar<-mean(shooting$age)
xbar
```

```
## [1] 41.72778
```

```
#sample deviation
sigma<-sd(shooting$age)
sigma
```

```
## [1] 14.30312
```

```
#sample number
n=180
#determine the z_alpha/2 for two sided
z<-qnorm(1-0.15/2)
z
```

```
## [1] 1.439531
```

```
xbar-z*sigma/sqrt(n)
```

```
## [1] 40.19311
```

```
xbar+z*sigma/sqrt(n)
```

```
## [1] 43.26245
```

So a two-sided 85% confidence interval of the age of victims is (40.19311,43.26245) Which means we have 85% confident that the interval (40.19311,43.26245) contains the ture population mean age μ of the victims.

b. A recent census study indicates that the average age of Americans is 40 years old. Co

```
shooting<-read.csv("shootings.csv")
#sample mean
xbar <- mean(shooting$age)
#sample deviation
s <- sd(shooting$age)
#sample number
n=180
#population mean
mu_0 <- 40
# two side hypothesis: $H_0$:$\mu$=40, $H_1$:$\mu$ $\neq$ 40.
t<-(xbar-mu_0)/s*sqrt(n)
t
```

```
## [1] 1.620666
```

```
t1 <- qt(0.975,df=n-1)
t1
```

```
## [1] 1.973305
```

```
p<-2*(1-pt(t,df=n-1))
p
```

```
## [1] 0.1068496
```

```
alpha=0.05
```

This is a two side hypothesis. The null hypothesis(H_0) is the average age of victims are the same as 40 years old(μ=40). And the alternative hypothesis(H_1) is the average age of vicitms are different from 40 years old($\mu \neq 40$). Since 1.6207<1.9733 (0.1068>0.05),we fail to reject H0, there is insufficient evidence to conclude that average age of victims is significantly different from 40 years old.

c. At the $\alpha=0.01$ significance level, test "by-hand" (i.e. do not use the `t.test(
\vspace{5pt}

```r
shooting<-read.csv("shootings.csv")
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
```

```r
minority<-filter(shooting, minority == "yes")
nonminority<-filter(shooting, minority == "no")
xbar1<-mean(minority$age)
xbar2<-mean(nonminority$age)
n_1<-nrow(minority)
n_2<-nrow(nonminority)
sigma_1<-sd(minority$age)
sigma_2<-sd(nonminority$age)
xbar1
```

```
## [1] 36.72917
```

```r
xbar2
```

```
## [1] 43.54545
```

```r
n_1
```

```
## [1] 48
```

```r
n_2
```

```
## [1] 132
```

```r
sigma_1
```

```
## [1] 13.5407
```

```r
sigma_2
```

```
## [1] 14.18706
```

```r
#s_p^2=\frac{\left(n_1-1\right) s_1^2+\left(n_2-1\right) s_2^2}{\left(n_1-1\right)+\le
#sample variance
a=((48-1)*13.5407^2+(132-1)*14.18706^2)/((48-1)+(132-1))
a
```

```
## [1] 196.5404
```

```
#t=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{s_p^2\left(\
t=((36.72917-43.54545)-0)/sqrt(196.540428*(1/48+1/132))
t
```

```
## [1] -2.884648
```

```
p<-2*pt(t,(48-1+132-1))
print(p)
```

```
## [1] 0.004402599
```

Comment: two side hypothesis, H_0:μ_1=μ_2, H_1:$\mu_1 \neq \mu_2$.p=0.004402599< α, null hypothesis is rejected. The average age of minority victims is different than the average age of non-minority victims.

```
shooting<-read.csv("shootings.csv")
t.test(shooting$age~shooting$minority, mu=0, alternative = "two.sided", conf=0.99, var.e
```

```
##
##  Two Sample t-test
##
## data:  shooting$age by shooting$minority
## t = 2.8847, df = 178, p-value = 0.004403
## alternative hypothesis: true difference in means between group no and group yes is no
## 99 percent confidence interval:
##   0.6637988 12.9687770
## sample estimates:
##  mean in group no mean in group yes
##          43.54545          36.72917
```

2. In the dataset named "blackfriday.csv," there is information relating to the amount of money that a sample of $n = 31$ consumers spent shopping on Black Friday in 2017.
   a. A company is interested in determining an upper-bound on the mean amount of money spent on Black Friday in order to determine maximum effects on the economy. Construct a one-sided upper-bound 99% lower confidence interval "by-hand" (i.e. do not use the `t.test()` function, but still use `R`) for the mean amount of money spent on Black Friday. Interpret the results.

```
data <-read.csv("blackfriday.csv")
xbar <-mean(data$Amount)
xbar
```

```
## [1] 11087.65
```

```
s <- sd(data$Amount)
s
```

```
## [1] 5959.942
```

```
n <- length(data$Amount)
n
```

```
## [1] 31
```

```
alpha <- 0.01
# one-sided upper-bound 99% lower confidence interval
xbar-qt(alpha,n-1)*s/sqrt(n)
```

```
## [1] 13717.99
```

The sample mean 11087.65 contains within the 99% confident interval of the upper bound 13717.99.

```
b. Suppose that in 2018, the average amount spent shopping on Black Friday was \$12000.
\vspace{5pt}
```

```
data <-read.csv("blackfriday.csv")
xbar <-mean(data$Amount)
sigma <- sd(data$Amount)
n=31
alpha <- 0.05
mu_0=12000
t <- (xbar-mu_0)/sigma*sqrt(n)
t
```

```
## [1] -0.8523199
```

```
p= pt(t,n-1)
p
```

```
## [1] 0.2003949
```

Comments: H_0: $\mu \geq 12000$, H_1: $\mu < 12000$, Since p=0.2003949 $> \alpha$, fail to reject null hypothesis. There is no sufficient evidence to conclude that the mean amount spent shopping on Black Friday in 2017 is less than 12000 at the α=0.05 significance level.

3. The Duke Chronicle collected data on all 1739 students listed in the Class of 2018's "Freshmen Picture Book." In particular, the Duke Chronicle examined hometowns, details about the students' high schools, whether they won a merit scholarship, and their sports team involvement. Ultimately, the goal was to determine trends between those who do and do not join Greek life at the university. A subset of this data is contained in the file named "greek.csv." The variable `greek` is an indicator that equals 1 if the student is involved in Greek life and 0 otherwise. The variable `hstuition` gives the amount of money spent on the student's high school tuition.
    a. At the $\alpha = 0.1$ significance level, test whether the average high school tuition for a student who does not partake in Greek life is less than the average high school tuition for a student who does partake in Greek life. Assume unequal variances.

```
data <- read.csv("greek.csv")
t.test(data$hstuition~data$greek, mu=0,alt="less",conf=.9,var.eq=F,paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  data$hstuition by data$greek
## t = -2.7213, df = 52.121, p-value = 0.00441
## alternative hypothesis: true difference in means between group 0 and group 1 is less
## 90 percent confidence interval:
##       -Inf -5886.364
## sample estimates:
## mean in group 0 mean in group 1
##        23477.00        34731.57
```

Comments: Since p=0.00441<0.1, we reject null hyothesis. We——

```
b. Construct a one-sided, lower-bound 90% confidence interval on the mean amount of high
\vspace{5pt}
```

```
n=length(data$hstuition)
xbar<-mean(data$hstuition)
xbar
```

```
## [1] 27923.25
```

```
#alpha <- 0.1
#one-sided, lower-bound 90% confidence interval on the mean amount of high school tuit
mean(data$hstuition)-qt(0.9,n-1)*sd(data$hstuition)/sqrt(n)
```

```
## [1] 25365.03
```

Comments: 90% confident that sample mean are located in

4. Seven trumpet players are given a new breathing exercise to help with their breath support. The trumpet players are asked to play a C note for as long as they can both before and after the breathing exercise. The time (in seconds) that they can hold the note for are presented below. Assume times are normally distributed.

Subject	1	2	3	4	5	6	7
Before	9.1	11.2	11.9	14.7	11.7	9.5	14.2
After	10.7	14.2	12.4	14.6	16.4	10.1	19.2

   a. Construct a one-sided lower-bound 95% confidence interval for the mean after-before change time holding a note. Interpret your interval.

```
pre <- c(9.1,11.2,11.9,14.7,11.7,9.5,14.2)
post <- c(10.7, 14.2, 12.4, 14.6, 16.4, 10.1, 19.2)
change<-post-pre
change
```

```
## [1]  1.6  3.0  0.5 -0.1  4.7  0.6  5.0
```

```
xbar<-mean(change)
xbar
```

```
## [1] 2.185714
```

```
sigma<-sd(change)
sigma
```

```
## [1] 2.074792
```

```
n=7
t=qt(0.95,n-1)
t
```

```
## [1] 1.94318
```

```
##Lower One-Sided Confidence Interval: \bar{x}_d \pm t_{\alpha} \frac{s_d}{\sqrt{n}}
lower.bound= xbar-t*sigma/sqrt(n)
lower.bound
```

```
## [1] 0.6618768
```

This is an paried sample of before and after time of holding a note. I am 95% confidence that the lower one-sided confidence interval capture the true mean after-before change time holding a note.

b. Perform an appropriate test at the $\alpha = 0.1$ significance level to determine if the mean time holding a note is greater after the exercise than before.

```
#\bar{x}_d \pm t_{\alpha } \frac{s_d}{\sqrt{n}}
b=t.test(post,pre,conf=0.9,pair=T, alternative = "greater")
b
```

```
##
##  Paired t-test
##
## data:  post and pre
## t = 2.7872, df = 6, p-value = 0.01585
## alternative hypothesis: true mean difference is greater than 0
## 90 percent confidence interval:
##  1.056661      Inf
## sample estimates:
## mean difference
##        2.185714
```

```
#H_0: \mu_d \leq 0 \text { vs. } H_1: \mu_d>0
#One-sided paired t test
```

```
#t=\frac{\bar{x}_d-\mu_d}{s_d / \sqrt{n}}
t=(xbar-0)/(sigma/sqrt(n))
t
```

```
## [1] 2.787198
```

```
# p value
p=1-pt(t,n-1)
p
```

```
## [1] 0.01584723
```

Comments: Since p value=0.01585$<$ α =0.1, reject null hypothesis, there is sufficient evidence to conclude that, the mean time holding a note is greater after the exercise than before.

5. Let μ be the average amount of time in minutes spent on social media apps each day. Based on an earlier study, it is hypothesized that $\mu = 124$ minutes. It is believed, though, that people are spending increasingly more time on social media apps during the pandemic. We sample n people and determine the average amount of time spent on social media apps per day in order to test the hypotheses $H_0 : \mu \leq 124$ vs. $H_1 : \mu > 124$, at the $\alpha = 0.01$ significance level. Suppose we know that $\sigma = 26$ minutes.
   a. Create a sequence of reasonable alternative values for μ. Take $\mu_1 \in (124, 190)$, using `seq(124,190, by=0.001)` in `R`.

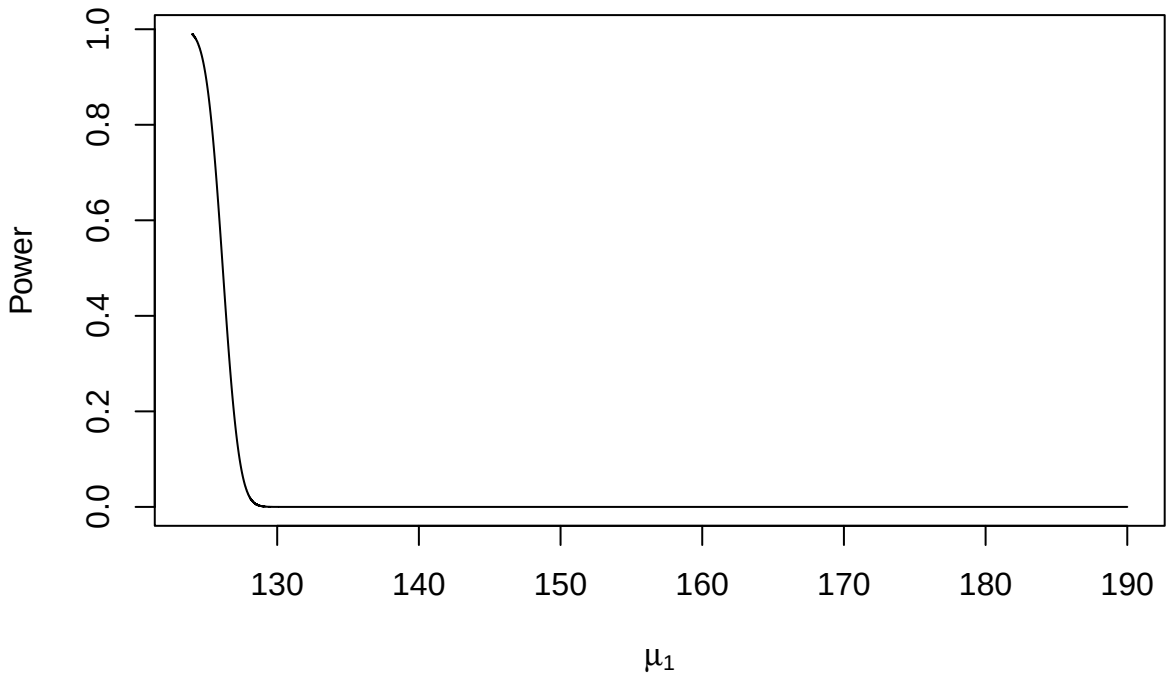
```
mu1<-seq(124,190, by=0.001)
```

```
b. Use `R` to draw a power curve for when $n=5$. You may find the `plot()` function usef
```

```
mu=124
z=qnorm(1-0.01)
n=5
xbar=mu+(sigma/sqrt(n))*z
xbar
```

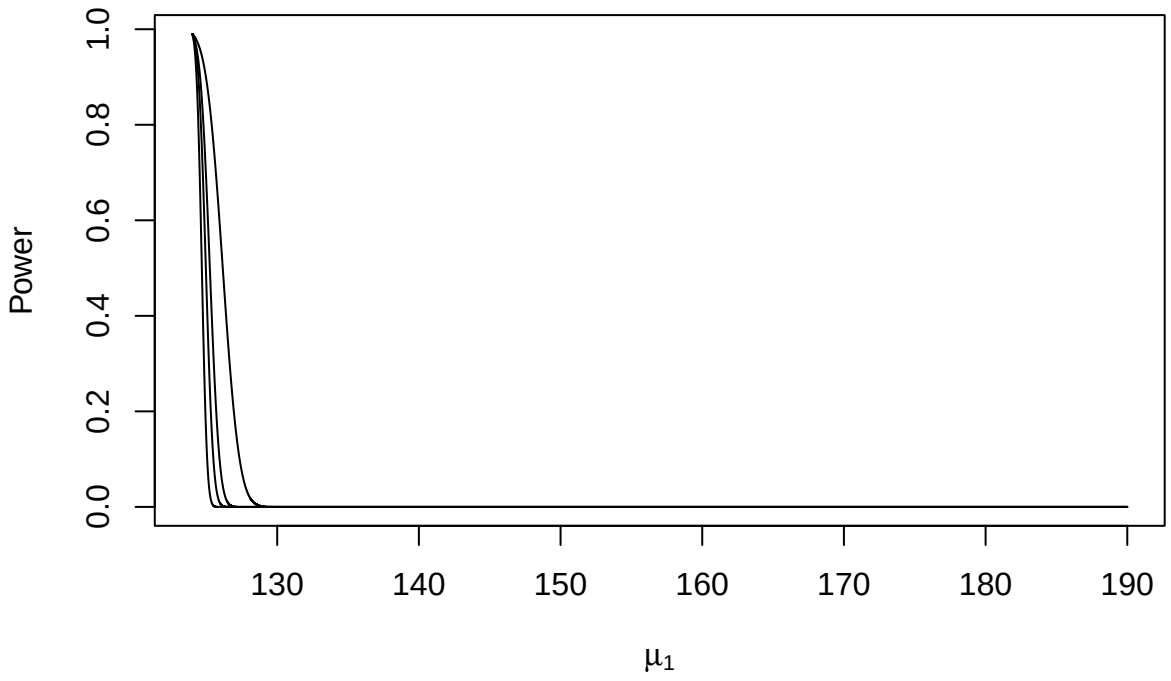
```
## [1] 126.1586
```

```
plot(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu[
```



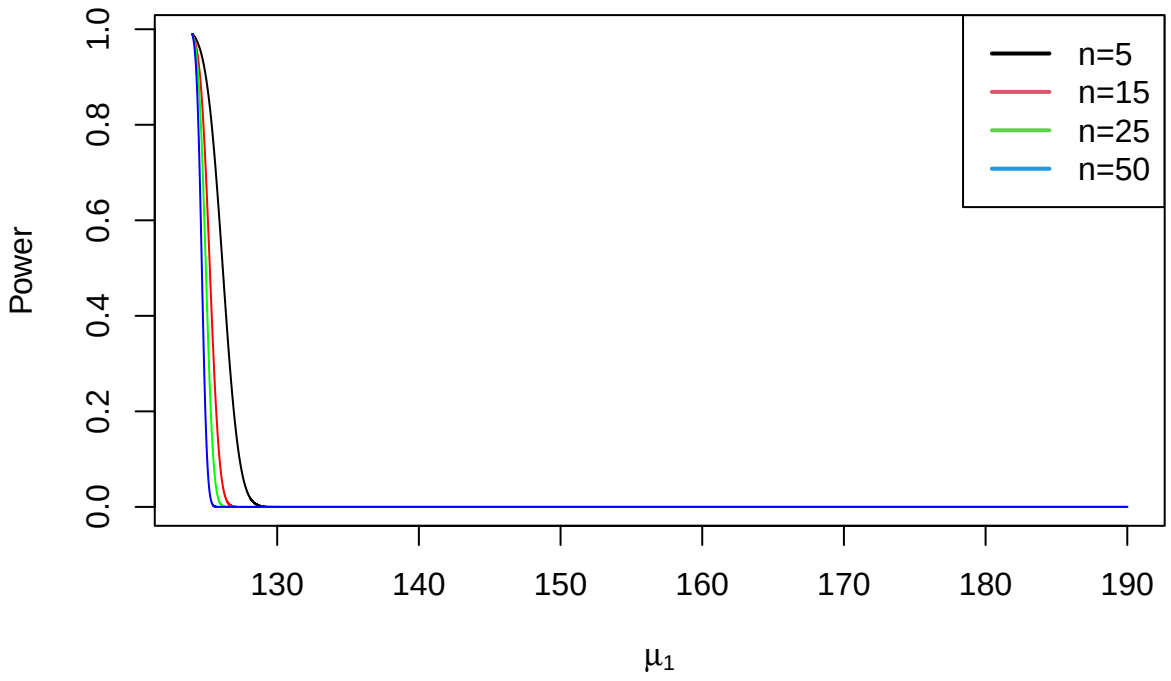
c. Using the same general plot as part b, draw power curves for when the sample size equ

```
n=5
xbar=mu+(sigma/sqrt(n))*z
plot(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu[
n=15
xbar=mu+(sigma/sqrt(n))*z
lines(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu
n=25
xbar=mu+(sigma/sqrt(n))*z
lines(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu
n=50
xbar=mu+(sigma/sqrt(n))*z
lines(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu
```



b. Make the curve for each of these a different color, and add a legend to distinguish t

```
n=5
xbar=mu+(sigma/sqrt(n))*z
plot(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu[
n=15
xbar=mu+(sigma/sqrt(n))*z
lines(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu
n=25
xbar=mu+(sigma/sqrt(n))*z
lines(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu
n=50
xbar=mu+(sigma/sqrt(n))*z
lines(mu1,pnorm(xbar,mu1,sigma/sqrt(n)),type = "l", ylab = "Power", xlab = expression(mu
legend(x = "topright",
       legend = c("n=5", "n=15", "n=25", "n=50"), col=c(1, 2, 3, 4, 5), lwd = 2)
```



d. What is the power of this test when $\mu_1=141$ and $n=28$?

```
alpha=0.01
sigma=26
mu1=141
mu=124
n=28
z=qnorm(1-0.01)
xbar=mu+(sigma/sqrt(n))*z
z1=(xbar-mu1)/(sigma/sqrt(n))
beta=pnorm(z1)
#power
1-beta
```

```
## [1] 0.8714938
```

e. How large of a sample size is needed to attain a power of 0.95 when the true mean a
\vspace{5pt}

```
alpha=0.01
beta=1-0.95
sigma=26
mu=124
mu1=128
n=28
z=qnorm(1-0.01)
xbar=mu+(sigma/sqrt(n))*z
#z2=(xbar-mu1)/(sigma/sqrt(n))
(sigma*qnorm(1-0.95)/(xbar-mu1))^2
```

```
## [1] 33.12482
```

Comments: The sample size of at least 33.12482 is needed to attain a power of 0.95 when the true mean amount of time on social media apps is $\mu_1 = 128$.

6. When it is time for vacation, many of us look to Air BnB for renting a room/house. Data collected on $n = 83$ Air BnB listings in New York City are contained in the file "airbnb.csv." Read this file into R.
    a. Create two new variables: one for the price of full house rentals and one for the price of private room rentals. You can use code such as this to subset:

```
data <- read.csv("airbnb.csv")
house <- filter(data, room_type == "Entire home")
hp <- house$price
room <- filter(data, room_type == "Private room")
pr <- room$price
```

b. Make a histogram for each of the new variables from part a to visualize their distrib

```
n1 <- length(hp)
n1
```

```
## [1] 27
```

```
k1 = ceiling(log2(n1))+1
binwidth1 <- (max(hp)-min(hp))/k1
binwidth1
```

```
## [1] 139.1667
```

```
n2 <- length(pr)
n2
```

```
## [1] 56
```

```
k2 = ceiling(log2(n2))+1
binwidth2 <- (max(pr)-min(pr))/k2
binwidth2
```

```
## [1] 40
```

```
library(ggplot2)
ggplot(house, aes(x=hp)) +
  geom_histogram(binwidth = 140)  +
labs(x="Price", title="The price of the full house rentals") +
  theme(plot.title = element_text(hjust = 0.5))
```


```
ggplot(room, aes(x=pr)) +
  geom_histogram(binwidth = 40)  +
labs(x="Price", title="The price of the private room rentals") +
  theme(plot.title = element_text(hjust = 0.5))
```

c. Discuss why we generally can apply the central limit theorem to analyze these two var

Comment: According to the histogram, both distributions are asymmetric, so they are not normal distribution. However when sample size n $>$ 30, we can apply the central limit theorem.

d. Calculate the mean, standard deviation, and sample size for the price of full home re

```
#mean for the price of full home rentals
xbar1 <- mean(hp)
print(xbar1)
```

```
## [1] 258.2593
```

```
#standard deviation for the price of full home rentals
sigma1 <- sd(hp)
print(sigma1)
```

```
## [1] 208.2271
```

```
#sample size for the price of full home rental.
n1 <- length(hp)
print(n1)
```

```
## [1] 27
```

Comment: the mean, standard deviation, and sample size for the price of full home rentals are 258.2593, 208.2271, 27.

e. Calculate the mean, standard deviation, and sample size for the price of private room

```
#mean for the price of private room rentals
xbar2 <- mean(pr)
print(xbar2)
```

```
## [1] 91.92857
```

```
#standard deviation for the price of full home rentals
sigma2 <- sd(pr)
print(sigma2)
```

```
## [1] 49.91005
```

```
#sample size for the price of full home rental.
n2 <- length(pr)
print(n2)
```

```
## [1] 56
```

Comment: the mean, standard deviation, and sample size for the price of private room rentals 91.92857, 49.91005, 56.

f. At the $\alpha=0.05$ significance level, test "by-hand" (i.e. do not use the `t.test(

```
#This is two sample with unequal variances, we use Welch t-test.
#t=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{s_1^2}
t=((xbar1-xbar2)-0)/sqrt({sigma1}^2/n1+{sigma2}^2/n2)
t
```

```
## [1] 4.094341
```

```
#\nu=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac
nu=({sigma1}^2/n1+{sigma2}^2/n2)^2/((({sigma1}^2/n1)^2/(n1-1)+({sigma2}^2/n2)^2/(n2-1)))
nu
```

```
## [1] 27.45038
```

```
p=2*(1-pt(t,nu))
p
```

```
## [1] 0.0003360658
```

p=0.0003360658 <0.05, we reject null hypothesis, we have 95% sure that the average price of renting an entire home in NYC is different from the average price of renting a private room

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable?

- Who, if anyone, did you work with on this assignment?
- What questions do you have relating to any of the material we have covered so far in class?