DSCC/CSC/STAT 462 Assignment 3

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

Daxiang Na

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.

```
df <- read.csv("shootings.csv")
average_age <- mean(df$age)
s <- sd(df$age)
n <- nrow(df)
m <- function(x, n, s) {
    q <- 1 - (1 - x)/2
    a <- qt(q, n - 1) * (s/sqrt(n))
    return(a)
}
m <- m(0.85, n, s)
lower_bound <- average_age - m
upper_bound <- average_age + m
lower_bound
```

```
## [1] 40.18649
```

```
upper_bound
```

```
## [1] 43.26906
```

Answer: The two sided 85% confidence interval on the mean age of victims is (40.18649, 43.26906). Which means that we are 85% sure that the age of victims is captured by the interval (40.18649, 43.26906).

b. A recent census study indicates that the average age of Americans is 40 years old. Conduct a hypothesis test "by-hand" (i.e. do not use the `t.test()` function, but still use `R`) at the $\alpha = 0.05$ significance level to see if the average age of victims is significantly different from 40 years old.

Answer:

1. Conditions: we have a normal population
2. Parameters of interest: μ = average age of victims
3. Significance level: 0.05
4. null hypothesis: $\mu = \mu_0 = 40$ years old.
5. alternative hypothesis: $\mu \neq \mu_0 = 40$ years old.
6. Which test and test statistic: σ is unknown, so $t = (x - \mu_0)/(s/\sqrt{n}) = 1.62$
7. calculate p-value

```
t <- (average_age - 40)/(s/sqrt(n))
p1 <- 2 * (1 - pt(t, n - 1))
p1
```

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

p-value = 2 * Pr(T > 1.62) = 2 * (1 - pt(t, n - 1)) = 0.1068496 > alpha = 0.05

8. Fail to reject H_0.
9. Conclusion: the average age of victims is not 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()` function, but still use `R`) whether the average age of minority victims is different than the average age of non-minority victims. Assume equal variances.

Answer:

1. Conditions: two normal populations with equal variance.
2. Parameters of interest:

μ_1 = the average age of minority victims

μ_2 = the average age of non-minority victims

3. Significance level: 0.01
4. null hypothesis: $\mu_1 = \mu_2$
5. alternative hypothesis: $\mu_1 \neq \mu 2$
6. Which test and test statistic:

   z-test

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

7. calculate p-value

```
m1 <- mean(df[df$minority == "yes", "age"])
m1
```

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

```
m2 <- mean(df[df$minority == "no", "age"])
m2
```

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

```
n1 <- nrow(df[df$minority == "yes", ])
n1
```

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

```
n2 <- nrow(df[df$minority == "no", ])
n2
```

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

```
s1 <- sd(df[df$minority == "yes", "age"])
s1
```

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

```
s2 <- sd(df[df$minority == "no", "age"])
s2
```

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

```
# here in this test, since the standard deviations of those
# two groups are very close, I assume those two groups has
# equal and unknown variance.  The pooled variance
```

```r
sp2 <- ((n1 - 1) * s1^2 + (n2 - 1) * s2^2)/(n1 - 1 + n2 - 1)
sp2
```

```
## [1] 196.5405
```

```r
z <- (m1 - m2)/sqrt(sp2 * (1/n1 + 1/n2))
z
```

```
## [1] -2.884651
```

```r
p2 <- 2 * pt(z, n1 + n2 - 2)
p2
```

```
## [1] 0.00440256
```

8. Conclusion: z = -2.884651, n(minority) = 48, n(non_minority) = 132, df = 178, pooled variance = 196.5405, p-value = 0.00440256 < alpha = 0.01, reject H_0, we conclude that the average age of minority victims is significantly different than the average age of non-minority victims.

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.

   Answer:

```r
rm(list = ls())
df <- read.csv("blackfriday.csv")
x_bar <- mean(df$Amount)
s <- sd(df$Amount)
n <- 31
m <- function(x, n, s) {
    q <- 1 - (1 - x)
    a <- qt(q, n - 1) * (s/sqrt(n))
    return(a)
}
m <- m(0.99, n, s)
upper <- x_bar + m
upper
```

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

The one-sided upper-bound 99% lower confidence interval is $13717.99. That means we are 99% sure that the amount of money spent on Black Friday per person is equal or

below $13717.99

b. Suppose that in 2018, the average amount spent shopping on Black Friday was $12000. Based on your sample, is there evidence to conclude that the mean amount spent shopping on Black Friday is 2017 is less than $12000? Conduct an appropriate hypothesis test "by-hand" (i.e. do not use the `t.test()` function, but still use `R`) at the $\alpha = 0.05$ significance level.

Answer:

1. Conditions: two normal populations with equal variance.
2. Parameters of interest:

μ = the average amount spent shopping on Black Friday in 2017

μ_0 = the average amount spent shopping on Black Friday in 2018 = 12000

3. Significance level: 0.05
4. null hypothesis: $\mu \geq \mu_0$
5. alternative hypothesis: $\mu < \mu 0$
6. Which test and test statistic: t test

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

```
t <- (x_bar - 12000)/(s/sqrt(n))
t
```

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

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

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

7. Conclusion: since $p = 0.2003949 > \alpha$, we fail to reject the null hypothesis $\mu \geq \mu_0$, and conclude that the mean amount spent on shopping on Black Friday 2017 is less than it was in 2018.

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.

```
rm(list = ls())
df <- read.csv("greek.csv")
greek_1 <- df %>%
    filter(greek == 1)
greek_0 <- df %>%
    filter(greek == 0)
m1 <- mean(greek_1$hstuition)
m0 <- mean(greek_0$hstuition)
n1 <- nrow(greek_1)
n0 <- nrow(greek_0)
s1 <- sd(greek_1$hstuition)
s0 <- sd(greek_0$hstuition)
t <- (m0 - m1)/sqrt(s0^2/n0 + s1^2/n1)
v <- (s0^2/n0 + s1^2/n1)^2/((s0^2/n0)^2/(n0 - 1) + (s1^2/n1)^2/(n1 -
    1))
p <- pt(t, v)
t
```

```
## [1] -2.721299
```

```
v
```

```
## [1] 52.12135
```

```
p
```

```
## [1] 0.004409615
```

Answer: t = -2.721299, degree of freedom = 52.12135, p-value = 0.004409615 < alpha = 0.1, we reject the null hypothesis, and conclude that 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.

b. Construct a one-sided, lower-bound 90% confidence interval on the mean amount of high school tuition paid by Duke students. Interpret the result.

```
m <- mean(df$hstuition)
s <- sd(df$hstuition)
n <- nrow(df)
lower_bound <- m - qt(0.9, n - 1) * s/sqrt(n)
lower_bound
```

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

Answer: the lower-bound 90% conficence interval on the mean amount of high school tuition paid by Duke students is $25365.03

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.

```
rm(list = ls())
subject <- c(1:7)
before <- c(9.1, 11.2, 11.9, 14.7, 11.7, 9.5, 14.2)
after <- c(10.7, 14.2, 12.4, 14.6, 16.4, 10.1, 19.2)
df <- data.frame(subject, before, after)
df[, "change"] <- df$after - df$before
m <- mean(df$change)
s <- sd(df$change)
n <- nrow(df)
lower_bound <- m - qt(0.95, n - 1) * s/sqrt(n)
lower_bound
```

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

Answer: the one-sided lower-bound 95% confidence interval for the mean after-before change time holding a note, which means that the time change of 95% players are above this level.

   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.

```
t.test(df$after, df$before, mu = 0, alt = "greater", paired = T)
```

```
##
##  Paired t-test
##
## data:  df$after and df$before
## t = 2.7872, df = 6, p-value = 0.01585
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  0.6618768       Inf
```

```
## sample estimates:
## mean difference
##        2.185714
```

Answer: the p-value = 0.01585 > alpha = 0.1, we fail to reject the null hypothesis, which stating that the mean time holding a note did not change after the exercise. We conclude that the mean time holding a note not significantly greater after the exercise than before.

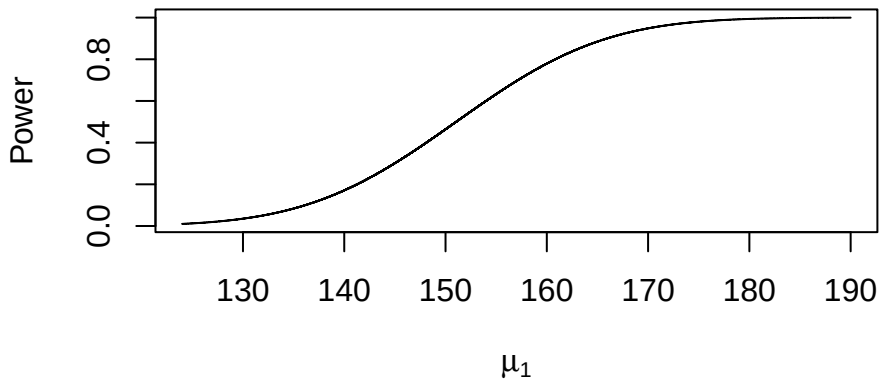
5. Let μ be the average amount of time in minutes spent on social media apps each day. Based on an earlier study, it is hypothesized that $\mu = 124$ minutes. It is believed, though, that people are spending increasingly more time on social media apps during the pandemic. We sample n people and determine 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`.

```r
rm(list = ls())
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 useful. In particular, `plot(mu1, __, type = "l", ylab = "Power", xlab = expression(mu[1]))` could be a useful starting point for formatting.

```r
mu <- 124
s <- 26
n <- 5
f <- function(x, n) {
    cutoff <- mu + qnorm(1 - 0.01) * s/sqrt(n)
    y <- 1 - (1 - pnorm((x - cutoff)/(s/sqrt(n))))
    return(y)
}
plot(mu1, f(mu1, 5), type = "l", ylab = "Power", xlab = expression(mu[1]))
```

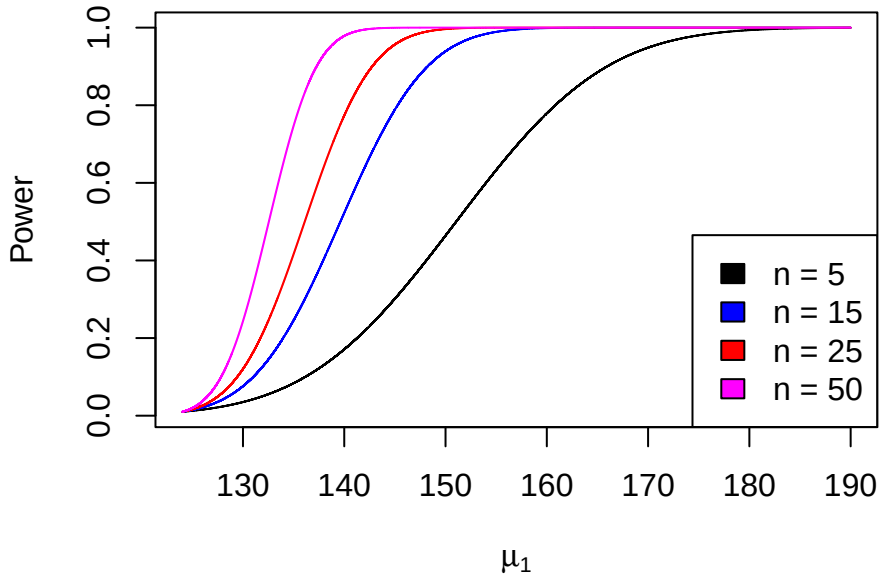


c. Using the same general plot as part b, draw power curves for when the sample size equals $n = 5, 15, 25, 50$. You can do this using the `lines()` function in place of when you used `plot()` in part b. Make the curve for each of these a different color, and add a legend to distinguish these curves.

Answer:

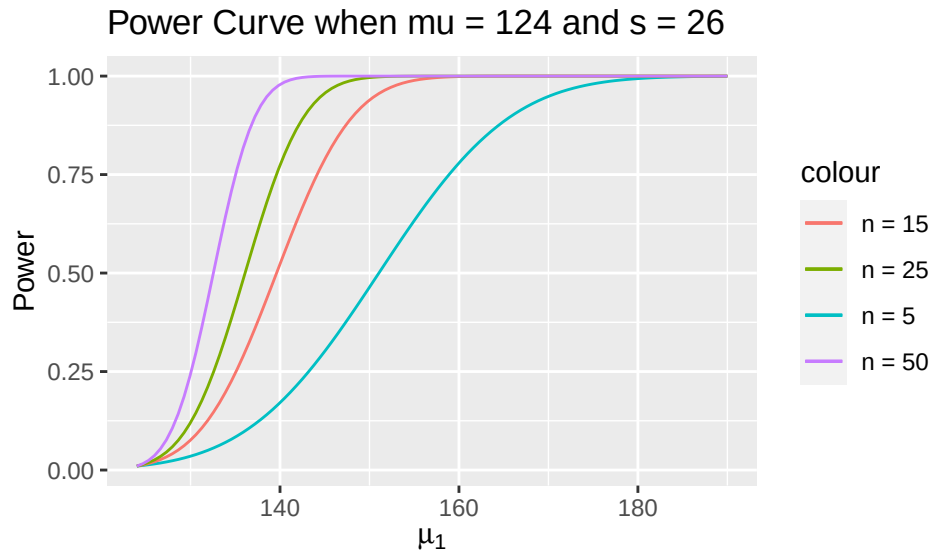
Using Base R

```
plot(mu1, f(mu1, 5), type = "l", ylab = "Power", xlab = expression(mu[1]),
    col = "black")
lines(mu1, f(mu1, 15), type = "l", ylab = "Power", xlab = expression(mu[1]),
    col = "blue")
lines(mu1, f(mu1, 25), type = "l", ylab = "Power", xlab = expression(mu[1]),
    col = "red")
lines(mu1, f(mu1, 50), type = "l", ylab = "Power", xlab = expression(mu[1]),
    col = "magenta")
legend(x = "bottomright", legend = c("n = 5", "n = 15", "n = 25",
    "n = 50"), fill = c("black", "blue", "red", "magenta"))
```



Using ggplot2

```
p <- ggplot(data.frame(x = mu1), aes(x))
p + geom_function(aes(colour = "n = 5"), fun = f, args = list(n = 5)) +
    geom_function(aes(colour = "n = 15"), fun = f, args = list(n = 15)) +
    geom_function(aes(colour = "n = 25"), fun = f, args = list(n = 25)) +
    geom_function(aes(colour = "n = 50"), fun = f, args = list(n = 50)) +
    labs(x = expression(mu[1]), y = "Power", title = "Power Curve when mu = 124 and
```



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

```
power <- f(141, 28)
power
```

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

e. How large of a sample size is needed to attain a power of 0.95 when the true mean amount of time on social media apps is $\mu_1 = 128$?

Answer:

- parameters: alpha = 0.01, power = 0.95 (beta = 0.05), $\mu_1 = 128$
- cutoff = 124 + qnorm(1-0.01)*s/sqrt(n)
- power = 1-(1-pnorm((x - cutoff)/(s/sqrt(n)))) = 0.95
- x = 128
- get n

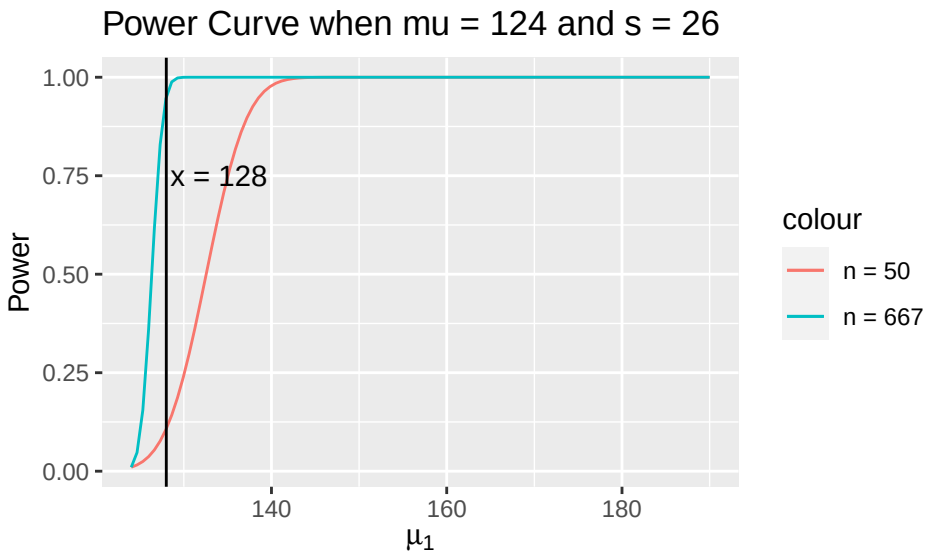
```
n = ((s/(128 - 124)) * (qnorm(0.95) + qnorm(0.99)))^2
ceiling(n)
```

```
## [1] 667
```

Conclusion: Sample size $n = 667$ is needed to attain a power of 0.95 when the true mean amount of time on social media apps is $\mu_1 = 128$

Double check with plot:

```
p + geom_function(aes(colour = "n = 50"), fun = f, args = list(n = 50)) +
    geom_function(aes(colour = "n = 667"), fun = f, args = list(n = 667)) +
    geom_vline(aes(xintercept = 128)) + annotate("text", x = 128 +
    6, y = 0.75, label = "x = 128") + labs(x = expression(mu[1]),
    y = "Power", title = "Power Curve when mu = 124 and s = 26")
```



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.

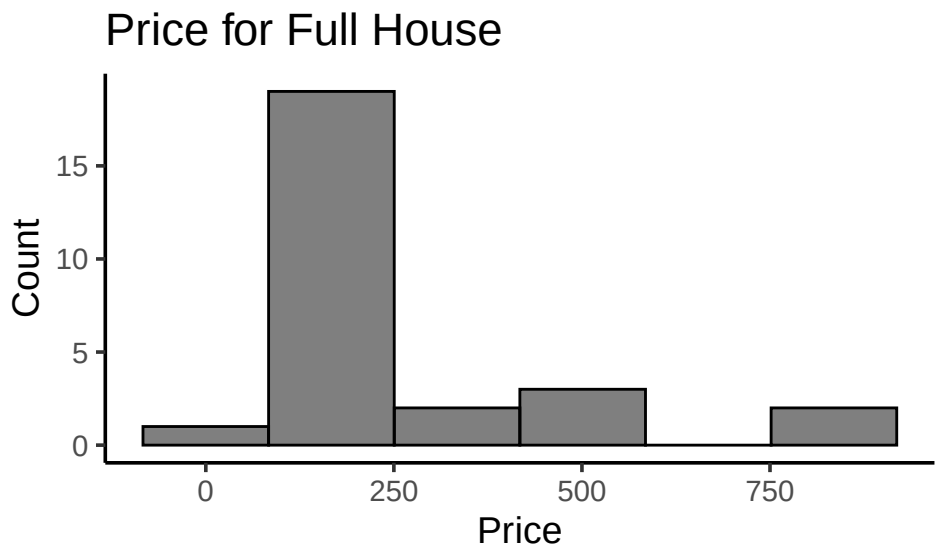
   a. Create two new variables: one for the price of full house rentals and one for the price of private room rentals. You can use code such as this to subset:

```
rm(list = ls())
df <- read.csv("airbnb.csv")
full <- df %>%
    filter(room_type == "Entire home")
private <- df %>%
    filter(room_type == "Private room")
```

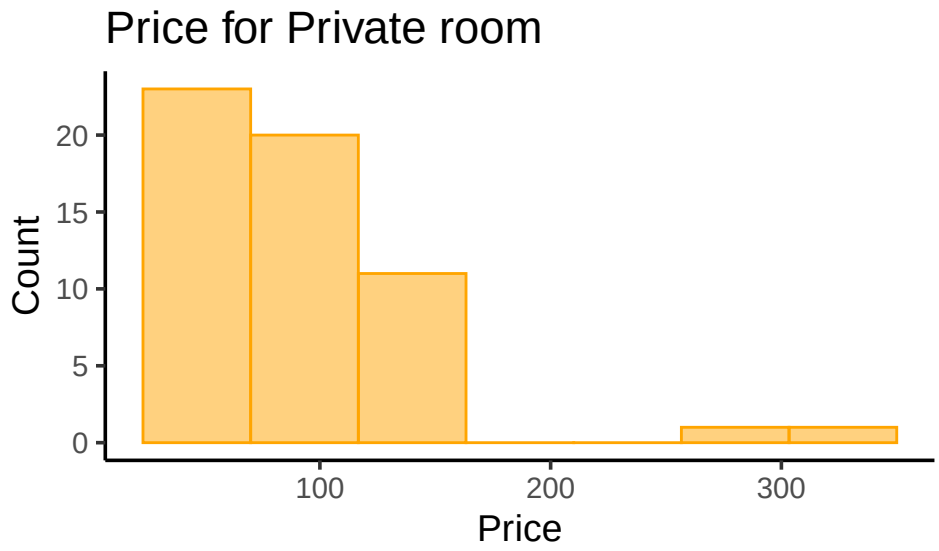
   In this answer, I subset the dataframe instead of the prices only: this is for the convience of calculation and plotting in later questions.

   b. Make a histogram for each of the new variables from part a to visualize their distributions. You can use base R or ggplot2.

```
k <- ceiling(log2(nrow(full)) + 1)
p <- ggplot(full) + geom_histogram(aes(price), bins = k, color = "black",
    fill = "black", alpha = 0.5) + labs(x = "Price", y = "Count",
    title = "Price for Full House") + theme_classic(base_size = 14)
p
```




```
h <- ceiling(log2(nrow(private)) + 1)
q <- ggplot(private) + geom_histogram(aes(price), bins = h, color = "orange",
    fill = "orange", alpha = 0.5) + labs(x = "Price", y = "Count",
    title = "Price for Private room") + theme_classic(base_size = 14)
q
```



c. Discuss why we generally can apply the central limit theorem to analyze these two variables. You should mention the histogram and the sample size, along with any potential reservations you have about using the CLT here.

Answer: Unlike the shape of normal distribution, the histogram of those two variables definitely do not look symmetrical, and it is interesting that both of them appear to be right-skewed. However, the sample size of those two variables are close or greater than 30, which allowed us to apply central limit theorem to analyze these two variables.

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

```
m1 <- mean(full$price)
s1 <- sd(full$price)
n1 <- nrow(full)
m1
```

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

```
s1
```

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

```
n1
```

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

Answer: for the price of full home rentals, mean = 258.2593, standard deviation = 208.2271, sample size = 27.

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

```
m2 <- mean(private$price)
s2 <- sd(private$price)
n2 <- nrow(private)
m2
```

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

```
s2
```

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

```
n2
```

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

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

f. At the $\alpha = 0.05$ significance level, test "by-hand" (i.e. do not use the `t.test()` function, but still use `R`) whether the average price of renting an entire home in NYC is different from the average price of renting a private room. Use unequal variances.

```
t <- (m1 - m2)/sqrt(s1^2/n1 + s2^2/n2)
v <- (s1^2/n1 + s2^2/n2)^2/((s1^2/n1)^2/(n1 - 1) + (s2^2/n2)^2/(n2 -
    1))
p <- 2 * (1 - pt(t, v))
t
```

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

```
v
```

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

```
p
```

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

Answer: t = 4.094341, degree of freedom = 27.45038, p-value = 0.0003360658 < alpha = 0.05, we reject the null hypothesis and conclude that the average price of renting an entire home in NYC is significantly 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?

Pretty long, about a week including reviewing contents learnt in class.

- Who, if anyone, did you work with on this assignment?

Myself.

- What questions do you have relating to any of the material we have covered so far in class?

For 1c, what is the best test?