# DSCC/CSC/STAT 462 Assignment 3

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

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.

```
shootings <- read.csv("shootings.csv")
xbar = mean(shootings$age)
s = sd(shootings$age)
n = length(shootings$age)
t85_twosided = qt(0.925, df = n - 1)

lb = xbar - t85_twosided * s/sqrt(n)
ub = xbar + t85_twosided * s/sqrt(n)
lb
```

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

```
ub
```

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

```
# check work t.test(shootings$age, conf.level = 0.85)
```

I am 85% confident that the interval (40.186, 43.269) contains the true average age of victims.

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.

```
mu0 = 40
tstat = (xbar - mu0)/(s/sqrt(n))
tstat
```

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

```
pvalue = 2 * (1 - pt(tstat, df = n - 1))
pvalue
```

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

```
# check work t.test(shootings$age, mu=40)
```

$H_0 : \mu = 40$ vs. $H_1 : \mu \neq 40$
The test statistic is 1.6207, and the p-value is 0.1068. Since the p-value is greater than $\alpha$, we fail to reject the null hypothesis. There is not sufficient evidence that the mean age is different from 40.

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.

```
minority_ages <- shootings$age[shootings$minority == "yes"]
nonminority_ages <- shootings$age[shootings$minority == "no"]

n1 = length(minority_ages)
n2 = length(nonminority_ages)

x1 = mean(minority_ages)
x2 = mean(nonminority_ages)

s1 = sd(minority_ages)
s2 = sd(nonminority_ages)

sp = sqrt(((n1 - 1) * s1^2 + (n2 - 1) * s2^2)/(n1 + n2 - 2))
tstat = (x2 - x1)/(sp * sqrt(1/n1 + 1/n2))
tstat
```

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

```
pvalue = 2 * (1 - pt(tstat, df = n1 + n2 - 2))
pvalue
```

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

```
# check work t.test(shootings$age ~ shootings$minority,
# var.equal=T)
```

$H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$
The test statistic is 2.8847 and the p-value is 0.004403. Since the p-value is less than alpha we reject the null hypothesis and conclude that there is a significant difference in the mean age of minority and 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.

```
blackfriday <- read.csv("blackfriday.csv")

xbar = mean(blackfriday$Amount)
s = sd(blackfriday$Amount)
n = length(blackfriday$Amount)
t99_onesided = qt(0.99, df = n - 1)

ub = xbar + t99_onesided * s/sqrt(n)
ub
```

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

```
# check work t.test(blackfriday$Amount, conf.level=0.99,
# alt='less')
```

I am 99% confident that the interval (0, 13717.99) captures the true mean amount of money spent on Black Friday in 2017.

   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.

```
mu0 = 12000
tstat = (xbar - mu0)/(s/sqrt(n))
tstat
```

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

```
pvalue = pt(tstat, df = n - 1)
pvalue
```

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

```
# check work t.test(blackfriday$Amount, mu=12000,
# alt='less')
```

$H_0 : \mu = 12000$ vs. $H_1 : \mu < 12000$. The test statistic for this hypothesis is -0.852 based on 30 degrees of freedom. This results in a p-value of 0.2004. Since the p-value is greater than $\alpha$, I fail to reject the null hypothesis and cannot conclude that spending in 2017 is less than $12000.

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.

```
greek <- read.csv("greek.csv")
t.test(greek$hstuition ~ greek$greek, var.equal = F, alt = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  greek$hstuition by greek$greek
## t = -2.7213, df = 52.121, p-value = 0.00441
## alternative hypothesis: true difference in means between group 0 and group 1 is
## 95 percent confidence interval:
##       -Inf -4328.795
## sample estimates:
## mean in group 0 mean in group 1
##        23477.00        34731.57
```

$H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$.
The test statistic is $t = -2.7213$, which leads to a p-value of $p = 0.00441$. Since $p < \alpha$, we reject the null hypothesis and conclude that the mean amount of high school tuition paid for those not in Greek life is less than the mean amount of high school tuition paid for those 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.

```
t.test(greek$hstuition, alt = "greater", conf.level = 0.9)
```

```
##
##  One Sample t-test
##
## data:  greek$hstuition
## t = 14.105, df = 80, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 90 percent confidence interval:
##  25365.03      Inf
## sample estimates:
## mean of x
##  27923.25
```

I am 90% confident that the interval (26248.15, $\infty$) contains the true mean high school tuition amount.

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.

```
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)
t.test(after, before, paired = T, alt = "greater")
```

```
##
##  Paired t-test
##
## data:  after and 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
```

I am 95% confident that the interval (0.6618768, $\infty$) captures the true mean change breath support time due to the breathing exercise.

   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.

```r
t.test(after, before, paired = T, alt = "greater")
```

```
##
##  Paired t-test
##
## data:  after and 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
```

Let $\delta = \mu_{after} - \mu_{before}$
$H_0 : \delta \leq 0$ vs. $H_1 : \delta > 0$
The test statistic is $t = 2.7872$, which leads to a p-value of 0.01585. Since the p-value is less than $\alpha = 0.1$, I reject the null hypothesis and conclude that mean breath support is higher after the breathing exercise.

5. Let $\mu$ 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.

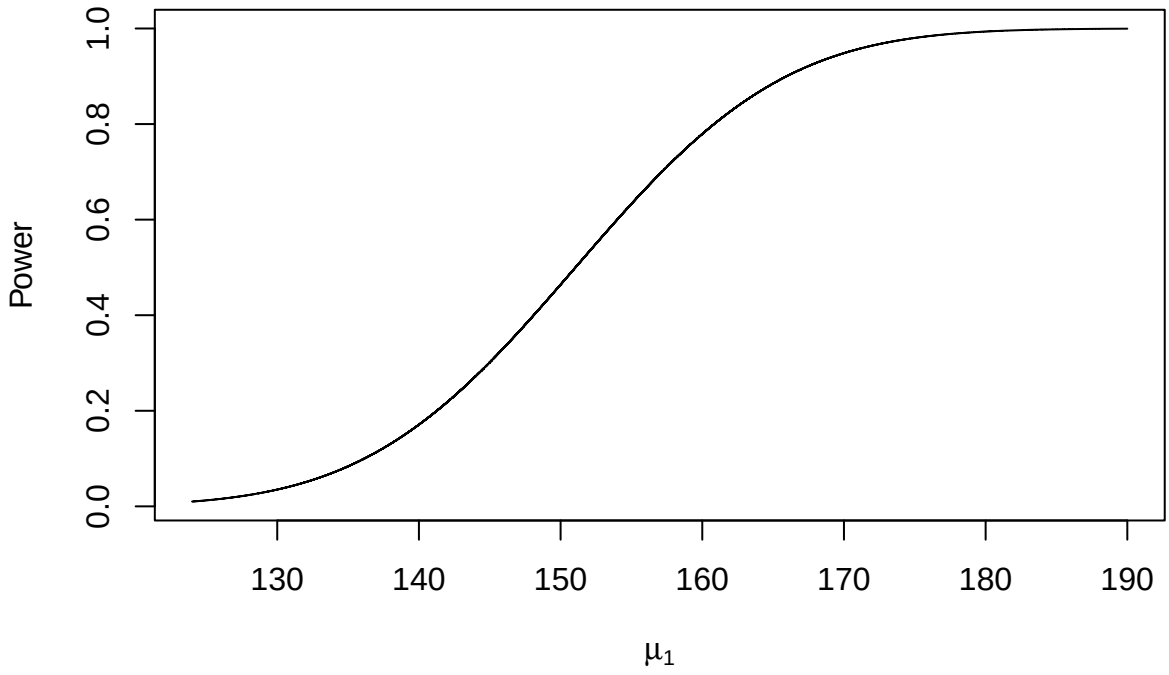
   a. Create a sequence of reasonable alternative values for $\mu$. Take $\mu_1 \in (124, 190)$, using `seq(124,190, by=0.001)` in `R`.

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

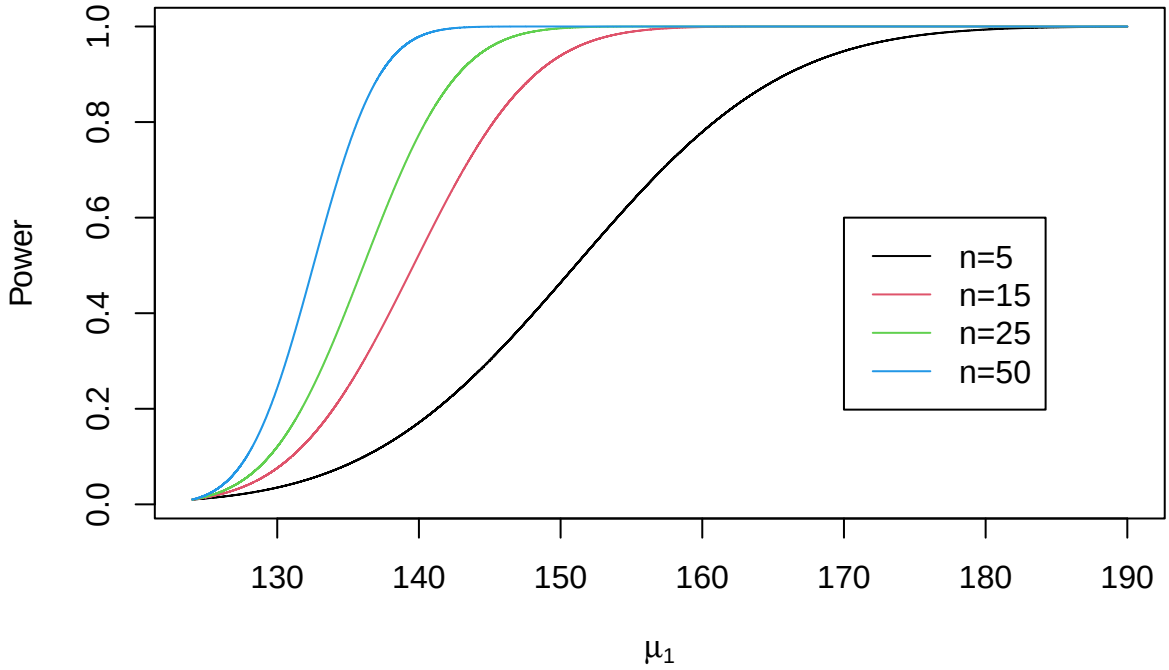
```r
sd <- 26
n <- 5
alpha <- 0.01
mu0 <- 124

#### 'greater than' alternative
x <- mu0 - qnorm(alpha) * sd/sqrt(n)
# plot(mu1, 1-pnorm((x-mu1)/(sd/sqrt(n))), type='l',
# ylab='Power', xlab=expression(mu[1]))
plot(mu1, 1 - pnorm(x, mu1, sd/sqrt(n)), 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.

```
n <- 5
x <- mu0 - qnorm(alpha) * sd/sqrt(n)
plot(mu1, 1 - pnorm((x - mu1)/(sd/sqrt(n))), type = "l", ylab = "Power",
    xlab = expression(mu[1]))
n <- 15
x <- mu0 - qnorm(alpha) * sd/sqrt(n)
lines(mu1, 1 - pnorm((x - mu1)/(sd/sqrt(n))), col = 2)
n <- 25
x <- mu0 - qnorm(alpha) * sd/sqrt(n)
lines(mu1, 1 - pnorm((x - mu1)/(sd/sqrt(n))), col = 3)
n <- 50
x <- mu0 - qnorm(alpha) * sd/sqrt(n)
lines(mu1, 1 - pnorm((x - mu1)/(sd/sqrt(n))), col = 4)
legend(170, 0.6, legend = c("n=5", "n=15", "n=25", "n=50"), col = c(1,
    2, 3, 4), lty = 1)
```




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

```
n = 28
mu1 = 141
x <- mu0 - qnorm(alpha) * sd/sqrt(n)
1 - pnorm((x - mu1)/(sd/sqrt(n)))
```

```
## [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 $\mu_1 = 128$?

```
beta = 0.05
mu1 = 128
ceiling((((qnorm(alpha) + qnorm(beta)) * sd)/(mu1 - mu0))^2)
```

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

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:

```
homeprice <- airbnb$price[airbnb$room_type == "Entire home"]
```

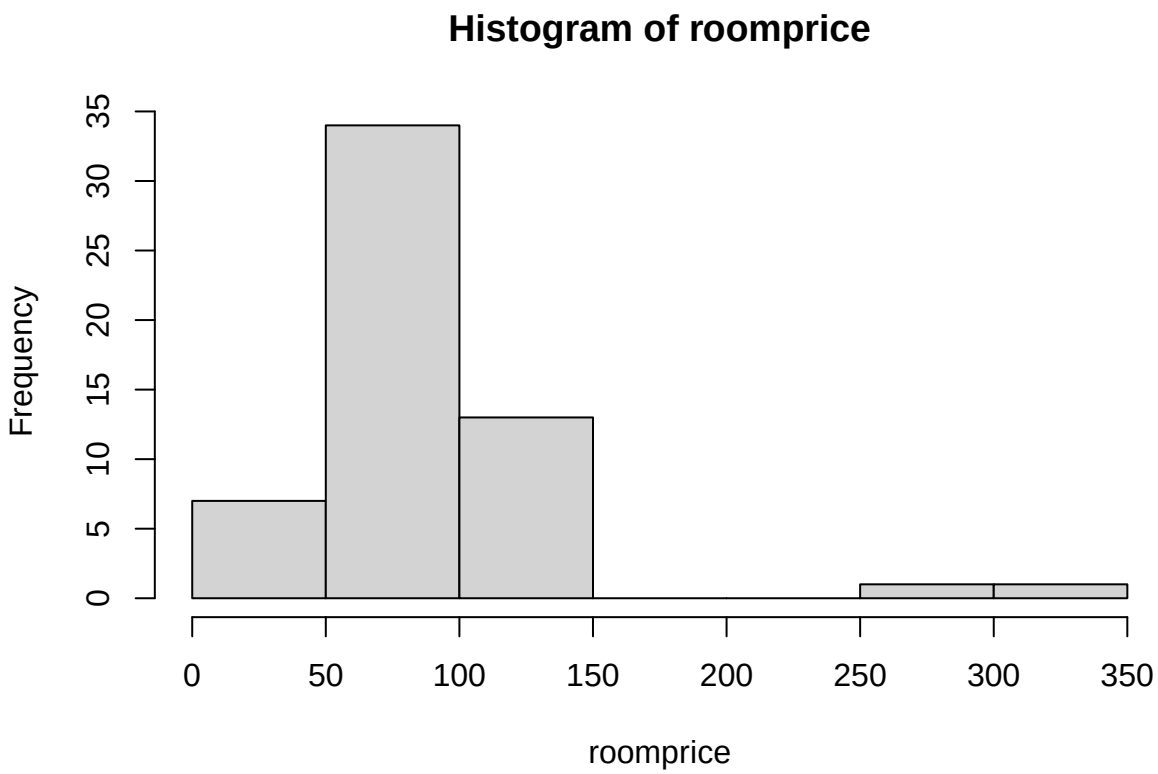
```
airbnb <- read.csv("airbnb.csv")
homeprice <- airbnb$price[airbnb$room_type == "Entire home"]
```

```
roomprice <- airbnb$price[airbnb$room_type == "Private room"]
```

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

```
hist(homeprice)
```


```
hist(roomprice)
```




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. Generally, looking at the histograms, we see some skewness, but a general bell-shape curve. Given that the sample sizes are relatively large, the CLT theorem seems appropriate to use here. One sample size is 27, which is lower than the cutoff of 30, so we proceed with a bit of caution.

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

```
(m1 = mean(homeprice))
```

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

```
(s1 = sd(homeprice))
```

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

```
(n1 = length(homeprice))
```

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

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

```
(m2 = mean(roomprice))
```

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

```
(s2 = sd(roomprice))
```

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

```
(n2 = length(roomprice))
```

```
## [1] 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))
```

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

```
df = ((s1^2/n1) + (s2^2/n2))^2/((s1^2/n1)^2/(n1 - 1) + (s2^2/n2)^2/(n2 -
    1))
2 * (1 - pt(t, df))
```

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

$H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$
The test statistic is 4.094, and the p-value is 0.00033. Since the p-value is smaller than alpha, we reject the null hypothesis and conclude a significant difference in the mean price of entire home rentals and private room rentals.

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?