DSCC/CSC/STAT 462 Assignment 4

Due November 3, 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.*

1. Recall the "airbnb.csv" dataset from HW3. Data collected on $n = 83$ Air BnB listings in New York City are contained in the file "airbnb.csv." Read this file into R and, just as in HW3, create two new variables, one for the price of full house rentals and one for the price of private room rentals. (It may be useful to revisit some of your code from that assignment.)

    a. At the $\alpha = 0.05$ level, test "by-hand" (i.e. do not use any `.test()` function, but still use `R`) whether the variance of price of entire home rentals is significantly different from the variance of price of private home rentals.

    Answer:

    Approach 1: compare price of entire home to price of private home

```
rm(list = ls())
data <- read.csv("airbnb(1).csv")
full <- data %>%
    filter(room_type == "Entire home")
private <- data %>%
    filter(room_type == "Private room")
s1 <- sd(full$price)
s2 <- sd(private$price)
n1 <- nrow(full)
n2 <- nrow(private)
f <- (s1/s2)^2
f
```

```
## [1] 17.40597
```

```
p <- 2 * (1 - pf(f, n1 - 1, n2 - 1))
p
```

```
## [1] 0
```

Conclusion: F = 17.40597, with df1 = 26 and df2 = 55, p-value = 0 < alpha = 0.05, reject the null hypothesis and conclude that the variance of price of entire home rentals is significantly different from the variance of price of private home rentals.

Approach 2: compare price of private home to price of entire home

```
rm(list = ls())
data <- read.csv("airbnb(1).csv")
full <- data %>%
    filter(room_type == "Entire home")
private <- data %>%
    filter(room_type == "Private room")
s1 <- sd(full$price)
s2 <- sd(private$price)
n1 <- nrow(full)
n2 <- nrow(private)
f <- (s2/s1)^2
f
```

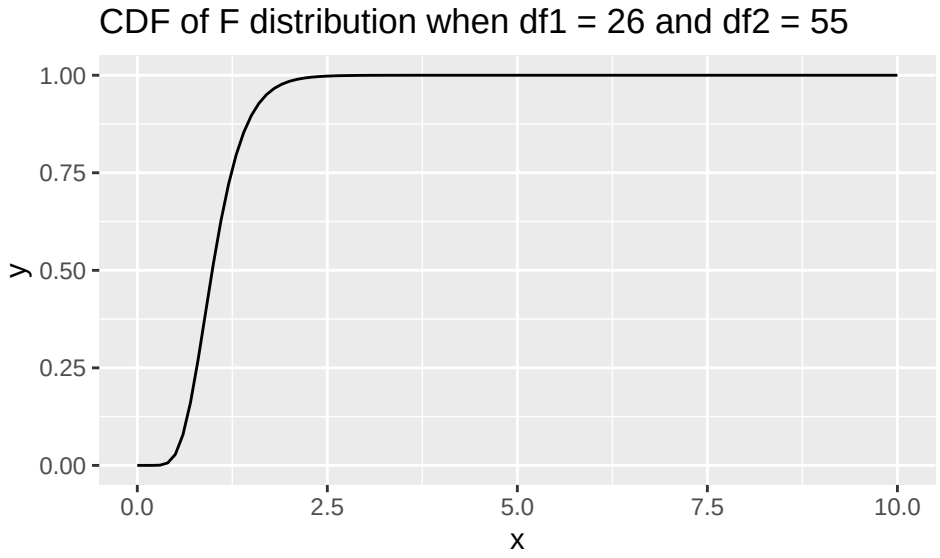
```
## [1] 0.05745154
```

```
p <- 2 * (pf(f, n2 - 1, n1 - 1))
p
```

```
## [1] 7.168664e-18
```

Conclusion: F = 0.05745154, with df1 = 55 and df2 = 26, p-value = 7.168664e-18 < alpha = 0.05, reject the null hypothesis and conclude that the variance of price of entire home rentals is significantly different from the variance of price of private home rentals.

```
a <- c(0:10)
p <- ggplot(data.frame(x = a), aes(x))
p + geom_function(fun = pf, args = list(df1 = 26, df2 = 55)) +
    labs(title = "CDF of F distribution when df1 = 26 and df2 = 55")
```



b. At the $\alpha = 0.05$ level, test "by-hand" (i.e. do not use any `.test()` function, but still use `R`) whether the variance of price of private room rentals is significantly different from 40^2.

Answer:

```
rm(list = ls())
data <- read.csv("airbnb(1).csv")
private <- data %>%
    filter(room_type == "Private room")
s1 <- sd(private$price)
n1 <- nrow(private)
t <- (n1 - 1) * s1^2/(40^2)
t
```

```
## [1] 85.62857
```

```
p <- 2 * (1 - pchisq(t, n1 - 1))
p
```

```
## [1] 0.01025083
```

Conclusion: $T_{obs} = 85.62857$, degree of freedom = n1 - 1 = 55, p-value = 0.01025083 < alpha = 0.05. Reject the null hypothesis and conclude that the variance of price of private room rentals is significantly different from 40^2.

2. A gaming store is interested in exploring the gaming trends of teenagers. A random sample of 143 teenagers is taken. From this sample, the gaming store observes that 95 teenagers play videos games regularly. For all parts of this problem, do the calculation "by-hand" (i.e. do not use the `prop.test()` or `binom.test()` functions, but still use `R`).

a. Construct a two-sided (Wald) 95% confidence interval for the proportion of all teenagers who play video games regularly. Interpret the interval.

```
rm(list = ls())
n <- 143
x <- 95
mean <- x/n
# use sample proportion to estimate population proportion
upper <- mean + qnorm(0.975) * sqrt(mean * (1 - mean)/n)
lower <- mean - qnorm(0.975) * sqrt(mean * (1 - mean)/n)
upper
```

```
## [1] 0.7417331
```

```
lower
```

```
## [1] 0.5869382
```

Answer: np = 95 > 5 and n(1-p) = 48 > 5. With that we conclude that the two-sided (Wald) 95% confidence interval for for the proportion of all teenagers who play video games regularly is (0.5869382, 0.7417331).

b. A teen magazine advertises that "74% of teenagers play video game regularly," and you want to see if this claim is true. Perform a hypothesis test at the $\alpha = 0.05$ significance level to test whether this claim is correct.

```
p0 <- 0.74
z <- (mean - p0)/sqrt(p0 * (1 - p0)/n)
p <- 2 * pnorm(-abs(z))
z
```

```
## [1] -2.062798
```

```
p
```

```
## [1] 0.03913182
```

Answer: z = -2.062798, p-value = 0.03913182 < alpha = 0.05. We reject the null hypothesis and conclude that this claim is not correct.

c. Comment on how comparable the results are from the confidence interval and the hypothesis test in examining the teen magazine's claim. Explain.

Answer: In the confidence interval, we estimate that the population proportion fall within (0.5869382, 0.7417331) with 95% confidence. However, in the hypothesis test, we conclude that the possibility that we get the sampling mean is lower than 1 - 95% if the population proportion = 0.74, which falls within the 95% CI. From that we can tell hypothesis test and CI are NOT mathematically equivalent in examining proportions, unlike examining means.

3. Researchers at a Las Vegas casino want to determine what proportion of its visitors smoke while in the casino. Casino executives are planning to conduct a survey, and they are willing to have a margin of error of 0.07 in estimating the true proportion of visitors who smoke. If the executives want to create a two-sided (Wald) 99% confidence interval, how many visitors must be included in the study?

```
rm(list = ls())
m <- 0.07
p <- 0.5
n <- ceiling(((qnorm((1 - 0.99)/2))^2) * p * (1 - p)/(m^2))
n
```

```
## [1] 339
```

Answer: 339 visitors must be included in the study to create a two-sided (Wald) 99% confidence interval.

4. Are people in Australia more likely to have pets than people in America? Of a sample of 51 Australians, 32 indicated having a pet. In an independent sample of 63 Americans, 27 indicated having a pet. Test "by-hand" (i.e. do not use the `prop.test()` or `binom.test()` functions, but still use `R`) at the $\alpha = 0.05$ significance level whether the proportion of Australians who have pets is greater than the proportion of Americans who have pets.

```
rm(list = ls())
# Australians
p1 <- 32/51
n1 <- 51
# Americans
p2 <- 27/63
n2 <- 63
# p_hat
p_hat <- (n1 * p1 + n2 * p2)/(n1 + n2)
p_hat
```

```
## [1] 0.5175439
```

```
# hypothesis test
s <- sqrt(p_hat * (1 - p_hat) * (1/n1 + 1/n2))
z <- (p1 - p2)/s
z
```

```
## [1] 2.112957
```

```
p_value <- 1 - pnorm(z)
p_value
```

```
## [1] 0.01730224
```

Answer: $\hat{p} = 0.5175439$, z = 2.112957, p-value = 0.01730224 < alpha = 0.05, we conclude that the proportion of Australians who have pets is significantly greater than the proportion of Americans who have pets

5. Researchers are interested in exploring severity of COVID-19 symptoms by age group. A sample of 193 patients at a health clinic were asked their age and have their symptoms categorized as "asymptomatic," "moderate," or "severe." The results are presented in the table below. Conduct an appropriate test (you do not need to do this test "by-hand" and can use the `chisq.test()` function) at the $\alpha = 0.01$ significance level to determine whether severity of COVID-19 symptoms is associated with age.

Age (years)	Asymptomatic	Moderate	Severe	Total
$[0, 18)$	22	13	7	42
$[18, 55)$	36	22	28	86
55 and older	10	29	26	65
Total	68	64	61	193

```
rm(list = ls())
A <- c(22, 36, 10)
M <- c(13, 22, 29)
S <- c(7, 28, 26)
mat_obs <- matrix(c(A, M, S), nrow = 3, ncol = 3)
mat_obs
```

```
##      [,1] [,2] [,3]
## [1,]   22   13    7
## [2,]   36   22   28
## [3,]   10   29   26
```

```
chisq.test(mat_obs, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat_obs
## X-squared = 20.408, df = 4, p-value = 0.0004147
```

Answer: p-value = 0.0004147 < alpha = 0.01, we conclude that severity of COVID-19 symptoms is significantly associated with age.

6. A study was conducted to investigate the respiratory effects of sulphur dioxide in subjects with asthma. During the study, two measurements were taken on each subject. First, investigators measured the increase in specific airway resistance (SAR)—a measure of broncho-constriction—from the time when the individual is at rest until after he/she has been exercising for 5 minutes (variable: `air`). The second measurement is the increase in SAR for the same subject after he/she has undergone a similar 5 minute exercise conducted in an atmosphere of 0.25 ppm sulfur dioxide (variable: `sulf.diox`). Ultimately, we are interested in examining the `air-sulf.diox` difference. For the

17 subjects enrolled in the study, the two measurements are presented in dataset "asthma.csv" on Blackboard.

a. At the $\alpha = 0.01$ significance level, use a Wilcox signed-rank test "by-hand" (i.e. do not use the `wilcox.test()` function, but still use R) to test the null hypothesis that the median difference in increase in SAR for the two air conditions is equal to 0 against the two-sided alternative hypothesis that it is not equal to 0. What do you conclude? Perform this test using a normal distribution approximation.

```
rm(list = ls())
data <- read.csv("asthma.csv")

# Calculate the difference of Before - After, but not After
# - Before

data <- data %>%
    mutate(difference = air - sulf.diox, abs = abs(difference)) %>%
    arrange(abs) %>%
    mutate(index = if_else(difference > 0, row_number(), -row_number()))

# T = T+ - T-

t <- abs(sum(data[data$index > 0, "index"])) - abs(sum(data[data$index <
    0, "index"]))
n <- nrow(data)
sigma <- sqrt(n * (n + 1) * (2 * n + 1)/6)
t
```

```
## [1] -111
```

```
n
```

```
## [1] 17
```

```
sigma
```

```
## [1] 42.24926
```

```
z <- (t - 0)/sigma
z
```

```
## [1] -2.627265
```

```
p <- 2 * pnorm(z)
p
```

```
## [1] 0.008607429
```

Answer: T = -111, n = 17 > 12 therefore $Z \sim N(0,1)$, $\sigma = 42.24926$, $z = -2.627265$, $p = 0.008607429 < \alpha = 0.01$. Therefore we reject the null hypothesis and conclude

that the median difference in increase in SAR for the two air conditions is significantly different from 0.

b. Run the test again using the exact signed-ranked distribution (i.e., `wilcox.test()`). How does the p-value differ from the result in part b?

```
p_alt <- wilcox.test(data$air, data$sulf.diox, paired = T, exact = T)
p_alt
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  data$air and data$sulf.diox
## V = 21, p-value = 0.006653
## alternative hypothesis: true location shift is not equal to 0
```

Answer: the p-value is higher when using the exact signed-ranked distribution.

7. The data in the file "bulimia.csv" are taken from a study that compares adolescents who have bulimia to healthy adolescents with similar body compositions and levels of physical activity. The data consist of measures of daily caloric intake for random samples of 23 bulimic adolescents and 15 healthy adolescents.

a. Read the data into R. To do so, use code such as this:

```
rm(list = ls())
bulimia <- read.csv("bulimia.csv")
bulimic <- bulimia$bulimic
healthy <- bulimia$health[1:15]
```

b. Test the null hypothesis that the median daily caloric intake of the population of individuals suffering from bulimia is equal to the median caloric intake of the healthy population. Conduct a two-sided test at the $\alpha = 0.01$ significance level (you do not need to do this test "by hand"; i.e., you may use a `.test()` function). Use a normal approximation for the distribution of the test statistic.

```
wilcox.test(bulimic, healthy, correct = F, paired = F, exact = F,
    alternative = "two.sided")
```

```
##
##  Wilcoxon rank sum test
##
## data:  bulimic and healthy
## W = 57.5, p-value = 0.0005927
## alternative hypothesis: true location shift is not equal to 0
```

Answer: p-value $= 0.0006262 <$ alpha $= 0.01$, we reject the null hypothesis and conclude that the median daily caloric intake of the population of individuals suffering from bulimia is significantly different from the median caloric intake of the healthy population.

Short Answers:

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

  About 8 hours including reviewing lectures.

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

  Discussed with classmates.

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

  No question for now.