DSCC/CSC/STAT 462 Assignment 4

Due November 3, 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.*

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.

```
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
```

```
data<-read.csv("airbnb.csv")
house<-filter(data, room_type == "Entire home")
eh<-house$price
room<-filter(data, room_type == "Private room")
pr <- room$price
n1 <- length(eh)
n2 <- length(pr)
s1 <- sd(eh)
s2 <- sd(pr)
```

```r
#F statistics F_obs
F_obs <- s1^2/s2^2
F_obs
```

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

```r
# p values
2*(1-pf(F_obs,n1-1,n2-1))
```

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

```r
qf(0.975,n1-1,n2-1)
```

```
## [1] 1.879284
```

Comments:

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ vs. } H_1 : \sigma_1^2 \neq \sigma_2^2$$

$$F_{obs} = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,n_2-1}$$

Because $F_{obs} > 1$, we have $2 * (1 - \mathrm{pf}\,(\mathrm{F_{obs}}\,, \mathrm{n_1} - 1, \mathrm{n_2} - 1))$ Because p value $= 0 < \alpha$,or $F_{obs} = 17.40597 \geq F_{n_1-1,n_2-1,1-\alpha/2}$ we 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.

```
b. At the $\alpha=0.05$ level, test "by-hand" (i.e. do not use any `.test()` function, b
```

```r
sigma_0=40
#Calculate the T statistic
T_obs <- (n2-1)*s2^2/sigma_0^2
T_obs
```

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

```r
n2-1
```

```
## [1] 55
```

```r
#p value
2*(1-pchisq(T_obs,n2-1))
```

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

Comments:

$$H_0 : \sigma^2 = 40^2 \text{ vs. } H_1 : \sigma^2 \neq 40^2$$

$$\text{since } T_{obs} = 86 > n - 1 = 55, \text{ we have } 2 * \text{ pchisq (} \mathrm{T_{obs}}\,, \mathrm{n} - 1 \text{)}$$

Because p value=0.01025083<α, we 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.

```
n=143
x=95
# $\hat{p}$
hat.p <- x/n
hat.p
```

```
## [1] 0.6643357
```

```
#check normality assumption
n*x/n
```

```
## [1] 95
```

```
n*(1-x/n)
```

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

```
hat.p-qnorm(0.975)*sqrt(hat.p*(1-hat.p)/n)
```

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

```
hat.p+qnorm(0.975)*sqrt(hat.p*(1-hat.p)/n)
```

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

Comments:

$$\hat{p} = \frac{x}{n}$$

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Two sided 95% confidence interval (0.5869382,0.7417331) of proportion of all teenagers who play video games regularly captures the hat.p(0.6643357), the proportion of all teenagers who play video games regularly. .

```
b. A teen magazine advertises that "74% of teenagers play video game regularly," and you
```

```
n=143
x=95
p0=0.74
hat.p <- x/n
#check normality assumption
n*p0
```

```
## [1] 105.82
```

```r
n*(1-p0)
```

```
## [1] 37.18
```

```r
#z value
z <- (hat.p-p0)/sqrt(p0*(1-p0)/n)
p=2*pnorm(-abs(z))
p
```

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

Comments:

$$H_0 : p == 0.74 \text{ vs. } H_1 : p \neq 0.74$$

Check normality assumption based on p Two-sided hypothesis test for proportions.np_0,$n(1-p_0)$ both larger than 5. p=0.03913182<$\alpha(\alpha = 0.05)$, reject null hypothesis, we have 95% confident that the state 74% of teenagers play video game regularly" is not correct.

```
c. Comment on how comparable the results are from the confidence interval and the hypoth
```

When looking at sample proportion, confidence intervals and hypothesis tests are not equivalent.For this case, the 95% confidence interval contains the true proportion 74% of teenage play video game regularly. However, according to hypothesis test, the proportion of teenager play video game regular is not 74%. Because for proportion (Wald) confidence intervals, we calculate the standard error based on $\hat{p}$ as $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, abd for proportion hypothesis tests, we calculate the standard error based on p_0 as $\sqrt{\frac{p_0(1-p_0)}{n}}$.

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?

```r
qnorm(0.995)
```

```
## [1] 2.575829
```

```r
ceiling((2.576)^2*0.5*(1-0.5)/(0.07)^2)
```

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

Comments: 339 visitors must be included in the study if the executives want 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.

```r
x1 <- 32
x2 <- 27
n1 <- 51
n2 <- 63
hat.p1 <- x1/n1
hat.p2 <- x2/n2
x1+x2
```

```
## [1] 59
```

```r
n1+n2
```

```
## [1] 114
```

```r
#hap.p <- (x1+x2)/(n1+n2)
hat.p <- 59/114
hat.p
```

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

```r
alpha=0.05
# Check normal assumptions
n1*hat.p1
```

```
## [1] 32
```

```r
n1*(1-hat.p1)
```

```
## [1] 19
```

```r
n2*hat.p2
```

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

```r
n2*(1-hat.p2)
```

```
## [1] 36
```

```r
# z-statistic
# z=\frac{\left(\hat{p}_1-\hat{p}_2\right)-\left(p_1-p_2\right)}{\sqrt{\hat{p}(1-\hat{
z=((hat.p1-hat.p2)-0)/sqrt(hat.p*(1-hat.p)*(1/n1+1/n2))
z
```

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

```r
#p value for one-sided hypothesis test
pnorm(-abs(z))
```

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

```r
2*pnorm(-1.675)
```

```
## [1] 0.09393423
```

Comments: One-sieded hypothesis test

$$H_0 : p1 = p2 \text{ vs. } H_1 : p1 \neq p2$$

n1*hat.p1>5 n1*(1-hat.p1)>5 n2$*$ hat.p2>5 n2 $*$ (1$-$ hat.p2)>5 p=0.01730224<α($\alpha = 0.05$), reject null hypothesis, we have 95% confident that the proportion of Australians who have pets is 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

Comment: H_0: severity of COVID-19 symptoms is not associated with age, H_1 :the severity of COVID-19 symptoms is not associated with age.

```r
(22-42*68/193)^2/(42*68/193)+
  (13-42*64/193)^2/(42*64/193)+
  (7-42*61/193)^2/(42*61/193)+
  (36-86*68/193)^2/(86*68/193)+
  (22-86*64/193)^2/(86*64/193)+
  (28-86*61/193)^2/(86*61/193)+
  (10-65*68/193)^2/(65*68/193)+
  (29-65*64/193)^2/(64*65/193)+
  (26-65*61/193)^2/(65*61/193)
```

```
## [1] 20.40833
```

```r
42*68/193
```

```
## [1] 14.79793
```

```r
42*64/193
```

```
## [1] 13.92746
```

```r
42*61/193
```

```
## [1] 13.27461
```

```r
86*68/193
```

```
## [1] 30.30052
```

```r
86*64/193
```

```
## [1] 28.51813
```

```r
86*61/193
```

```
## [1] 27.18135
```

```r
65*68/193
```

```
## [1] 22.90155
```

```r
64*65/193
```

```
## [1] 21.5544
```

```r
65*61/193
```

```
## [1] 20.54404
```

```r
df <- (3-1)*(3-1)
#p value
1-pchisq(20.4083,df)
```

```
## [1] 0.0004147371
```

```r
tab <- matrix(c(22,13,7,36,22,28,10,29,26),ncol=3,byrow=TRUE)
colnames(tab) <- c('Asystomatic','Moderate', 'Severe')
rownames(tab) <- c('[0,18)','[18,55)','55 and older')
tab
```

```
##              Asystomatic Moderate Severe
## [0,18)                22       13      7
## [18,55)               36       22     28
## 55 and older          10       29     26
```

```r
chisq.test(tab,correct = F)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  tab
## X-squared = 20.408, df = 4, p-value = 0.0004147
```

Comment: $H_0 : TheserverityofCOVID-19symptomsandageareindependent.H_1 : theserverityofCOVID-19symptomsandageareindependent$ All the xpected frequency>5 Because p vale=0.0004147< $\alpha = 0.01$, we reject null hypothesis,and we can conclude the severity of COVID-19 symptoms is 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.

```r
table <- read.csv("asthma.csv")
table
```

```
##    subject   air sulf.diox
## 1        1  0.82      0.72
## 2        2  0.86      1.05
## 3        3  1.86      1.40
## 4        4  1.64      2.30
## 5        5 12.57     12.59
## 6        6  1.56      1.42
## 7        7  1.28      2.41
## 8        8  1.08      2.32
## 9        9  4.29      8.19
## 10      10  1.37      6.33
## 11      11 14.68     19.88
## 12      12  3.64      3.87
## 13      13  3.89      9.25
## 14      14  0.58      6.59
## 15      15  9.50      6.17
## 16      16  0.93     10.93
## 17      17  0.49     15.44
```

```
di <- ifelse(is.nan(table$air) ==TRUE | is.nan(table$sulf.diox) ==TRUE,NaN,table$air-tab
di
```

```
##  [1]   0.10  -0.19   0.46  -0.66  -0.02   0.14  -1.13  -1.24  -3.90  -4.96
## [11]  -5.20  -0.23  -5.36  -6.01   3.33 -10.00 -14.95
```

```
df <- data.frame(table,di)
df
```

```
##    subject   air sulf.diox     di
## 1        1  0.82      0.72   0.10
## 2        2  0.86      1.05  -0.19
## 3        3  1.86      1.40   0.46
## 4        4  1.64      2.30  -0.66
## 5        5 12.57     12.59  -0.02
## 6        6  1.56      1.42   0.14
## 7        7  1.28      2.41  -1.13
## 8        8  1.08      2.32  -1.24
## 9        9  4.29      8.19  -3.90
## 10      10  1.37      6.33  -4.96
## 11      11 14.68     19.88  -5.20
## 12      12  3.64      3.87  -0.23
## 13      13  3.89      9.25  -5.36
## 14      14  0.58      6.59  -6.01
## 15      15  9.50      6.17   3.33
## 16      16  0.93     10.93 -10.00
## 17      17  0.49     15.44 -14.95
```

```
df_abs <- abs(df$di)
df_final <- data.frame(df, df_abs)
df_final
```

```
##    subject   air sulf.diox     di df_abs
## 1        1  0.82      0.72   0.10   0.10
## 2        2  0.86      1.05  -0.19   0.19
## 3        3  1.86      1.40   0.46   0.46
## 4        4  1.64      2.30  -0.66   0.66
## 5        5 12.57     12.59  -0.02   0.02
## 6        6  1.56      1.42   0.14   0.14
## 7        7  1.28      2.41  -1.13   1.13
## 8        8  1.08      2.32  -1.24   1.24
## 9        9  4.29      8.19  -3.90   3.90
## 10      10  1.37      6.33  -4.96   4.96
## 11      11 14.68     19.88  -5.20   5.20
## 12      12  3.64      3.87  -0.23   0.23
## 13      13  3.89      9.25  -5.36   5.36
```

```
## 14       14  0.58      6.59  -6.01   6.01
## 15       15  9.50      6.17   3.33   3.33
## 16       16  0.93     10.93 -10.00  10.00
## 17       17  0.49     15.44 -14.95  14.95
```

```
df_rank <- rank(df_final$df_abs)
df_final_rank <- data.frame(df_final, df_rank)
df_final_rank
```

```
##    subject   air sulf.diox     di df_abs df_rank
## 1        1  0.82      0.72   0.10   0.10       2
## 2        2  0.86      1.05  -0.19   0.19       4
## 3        3  1.86      1.40   0.46   0.46       6
## 4        4  1.64      2.30  -0.66   0.66       7
## 5        5 12.57     12.59  -0.02   0.02       1
## 6        6  1.56      1.42   0.14   0.14       3
## 7        7  1.28      2.41  -1.13   1.13       8
## 8        8  1.08      2.32  -1.24   1.24       9
## 9        9  4.29      8.19  -3.90   3.90      11
## 10      10  1.37      6.33  -4.96   4.96      12
## 11      11 14.68     19.88  -5.20   5.20      13
## 12      12  3.64      3.87  -0.23   0.23       5
## 13      13  3.89      9.25  -5.36   5.36      14
## 14      14  0.58      6.59  -6.01   6.01      15
## 15      15  9.50      6.17   3.33   3.33      10
## 16      16  0.93     10.93 -10.00  10.00      16
## 17      17  0.49     15.44 -14.95  14.95      17
```

```
library(dplyr)
table.t1 <- filter(df_final_rank, di >0)
t1 <- data.frame(table.t1)
t1
```

```
##   subject  air sulf.diox   di df_abs df_rank
## 1       1 0.82      0.72 0.10   0.10       2
## 2       3 1.86      1.40 0.46   0.46       6
## 3       6 1.56      1.42 0.14   0.14       3
## 4      15 9.50      6.17 3.33   3.33      10
```

```
table.t2 <- filter(df_final_rank, di <0)
t2 <- data.frame(table.t2)
t2
```

```
##    subject   air sulf.diox     di df_abs df_rank
## 1        2  0.86      1.05  -0.19   0.19       4
## 2        4  1.64      2.30  -0.66   0.66       7
## 3        5 12.57     12.59  -0.02   0.02       1
```

```
## 4         7  1.28       2.41  -1.13   1.13        8
## 5         8  1.08       2.32  -1.24   1.24        9
## 6         9  4.29       8.19  -3.90   3.90       11
## 7        10  1.37       6.33  -4.96   4.96       12
## 8        11 14.68      19.88  -5.20   5.20       13
## 9        12  3.64       3.87  -0.23   0.23        5
## 10       13  3.89       9.25  -5.36   5.36       14
## 11       14  0.58       6.59  -6.01   6.01       15
## 12       16  0.93      10.93 -10.00  10.00       16
## 13       17  0.49      15.44 -14.95  14.95       17
```

```
T <-sum(t1$df_rank)-sum(t2$df_rank)
T
```

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

```
mu <- 0
n <- 17
sigma <- sqrt(n*(n+1)*(2*n+1)/6)
z <- (T-mu)/sigma
z
```

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

```
#p value
2*pnorm(z)
```

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

Comments:

$$H_0 : \mu_T = 0 \text{ vs. } H_1 : \mu_T \neq 0$$

Check normality assumption based on n $Z_T \sim N(0, 1)$ given that n is large enough (typically $n > 12$) Because p=0.008607429 $< \alpha = 0.01$, we reject the null hypothesis. We conclude that the median difference in increase in SAR for the two air conditions is not equal to 0.

```
b. Run the test again using the exact signed-ranked distribution (i.e., `wilcox.test()`)
\vspace{10pt}
```

```
wilcox.test(table$air,table$sulf.diox, paired=T, exact=T, correct=F)
```

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

Comments: p value of the exact signed-ranked distribution is smaller than the p value of the normal distribution approximation (part b).

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:

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

   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.

```
bulimia <- read.csv("bulimia.csv")
cb <- append(bulimia$bulimic, bulimia$healthy[1:15])
cb
```

```
##  [1] 16.5 16.7 16.9 17.0 17.6 18.1 18.4 18.9 18.9 19.6 21.5 21.6 22.9 23.6 24.1
## [16] 24.5 22.1 25.2 25.6 28.0 28.7 29.2 30.9 20.7 22.4 23.1 23.8 24.5 25.3 25.7
## [31] 30.6 30.6 33.2 33.7 36.6 37.1 37.4 40.8
```

```
rank <- rank(cb, na.last = FALSE)
rank
```

```
##  [1]  1.0  2.0  3.0  4.0  5.0  6.0  7.0  8.5  8.5 10.0 12.0 13.0 16.0 18.0 20.0
## [16] 21.5 14.0 23.0 25.0 27.0 28.0 29.0 32.0 11.0 15.0 17.0 19.0 21.5 24.0 26.0
## [31] 30.5 30.5 33.0 34.0 35.0 36.0 37.0 38.0
```

```
W1 <- sum(rank[1:23])
W2 <- sum(rank[24:38])
W1
```

```
## [1] 333.5
```

```
W2
```

```
## [1] 407.5
```

```
W <- min(W1, W2)
n1 <- 23
n2 <- 15
mu <- n1*(n1+n2+1)/2
sigma <- sqrt(n1*n2*(n1+n2+1)/12)
z <- (W-mu)/sigma
z
```

```
## [1] -3.434366
```

```
2*pnorm(z)
```

```
## [1] 0.000593941
```

```
wilcox.test(bulimia$bulimic,bulimia$health[1:15],exact = F, correct = F, alternative = "
```

```
##
##  Wilcoxon rank sum test
##
## data:  bulimia$bulimic and bulimia$health[1:15]
## W = 57.5, p-value = 0.0005927
## alternative hypothesis: true location shift is not equal to 0
```

n1, n2 > 10, so $z_W \sim N(0, 1)$ Two sided hypothesis: H_0 : the population of individuals suffering from bulimia is equal to the median caloric intake of the healthy population H_1 : the population of individuals suffering from bulimia is not equal to the median caloric intake of the healthy population Because p value=0.00059<α=0.01,we reject null hypothesis and conclude that the population of individuals suffering from bulimia is not equal to 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 12 hours. It's doable.
- Who, if anyone, did you work with on this assignment? Prof. Anson and classmates.
- What questions do you have relating to any of the material we have covered so far in class? A bit confused about the meta, power curve, alpha.