DSCC/CSC/STAT 462 Assignment 4 - Aradhya Mathur

Due November 3, 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.*

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.)

```
air <- read.csv("airbnb.csv.")
head(air)
```

```
##         id neighbourhood_group   room_type price minimum_nights
## 1  1803165           Manhattan Entire home   799              6
## 2 13410813              Queens Entire home   120              3
## 3   941179            Brooklyn Entire home   150              2
## 4  1256768            Brooklyn Entire home   147              7
## 5  7816449           Manhattan Entire home   500              7
## 6  3415102            Brooklyn Entire home   500              2
##   number_of_reviews reviews_per_month availability_365
## 1                40              0.58              365
## 2                40              1.45              365
## 3                42              0.72              365
## 4                42              0.61              365
## 5                44              0.94              365
## 6                48              0.80              365
```

```
home <- air[air$room_type == "Entire home", ]
price_home = home$price
price_home
```

```
##  [1] 799 120 150 147 500 500 299 180 250 500 250 105 200 150 300  99 895 200 150
## [20] 165 150 105 200  60 125 249 125
```

```r
private <- air[air$room_type == "Private room", ]
price_private = private$price
price_private
```

```
##  [1]  70  68  95  80  75  43 100 109  70 150  85  39 120  89  65  55 100  68 150
## [20]  55 319 110  45  60  54  89  58  59  89  55  80  55  39 129 135 149 259  72
## [39]  75  80 135  50 150 119  70  69  80 125  69  80  77 150  80  48  50  99
```

a. At the $\alpha=0.05$ level, test "by-hand" (i.e. do not use any `.test()` function, b
H0: var of price of entire home equal to var of price of private room
H1: var of price of entire home not equal to var of price of private room

```r
sdhome = sd(price_home)
n1 = 27
sdhome
```

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

```r
sdprivate = sd(price_private)
n2 = 56
sdprivate
```

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

```r
f = (sdhome^2)/(sdprivate^2)
f
```

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

```r
pval = 2 * (1 - pf(f, n1 - 1, n2 - 1))
pval
```

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

P val = 0 which is less than alpha so we reject the null hypothesis

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

H0: variance of price of private room rentals is $= 40^2$ h1: variance of price of private room rentals is significantly different from 40^2

```
fnew = (n2 - 1) * (sdprivate^2)/(40 * 40)
fnew
```

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

```
pval = 2 * (1 - pchisq(fnew, 55))
pval
```

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

P val is less than alpha (reject null hypothesis) so it is significantly different.

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.

```
x = 95
n = 143
pcap = x/n
pcap
```

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

```
z1 = qnorm(0.975)
z1
```

```
## [1] 1.959964
```

```
funct = sqrt((pcap * (1 - pcap))/n)
funct
```

```
## [1] 0.0394892
```

```
interval = z1 * funct
interval
```

```
## [1] 0.07739742
```

```
lower = pcap - interval
lower
```

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

```
upper = pcap + interval
upper
```

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

(0.5869382,0.7417331) is the a two-sided (Wald) 95% confidence interval for the proportion of all teenagers who play video games regularly

```
# prop.test(x=95, n=143, p = NULL, alternative =
# c('two.sided', 'less', 'greater'), conf.level = 0.95,
# correct = TRUE)
```

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

H0: 74% people play regularly H1: 74% people dont play regularly

```
pn = 0.74
deno = sqrt((pn * (1 - pn))/n)
deno
```

```
## [1] 0.03668044
```

```
zb = (pcap - pn)/deno
zb
```

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

```
pval = 2 * (pnorm(zb))
pval
```

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

```
P value is less than alpha and we reject H0.
74% of teenagers don't play video game regularly
```

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

Both confidence interval and hypothesis test provide the same inference but in different manner. If for some percent of teenagers, p value is greater than alpha, then for sure for that value will lie between the confidence interval. Confidence intervals uses data from a sample to estimate a population parameter. While, hypothesis tests uses data from a sample to test a specified hypothesis.

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?

```
m = 0.07
p3 = 0.5
z3 = qnorm(0.995)
z3
```

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

```
num3 = ((z3^2) * (p3) * (1 - p3))
num3
```

```
## [1] 1.658724
```

```
deno3 = m^2
deno3
```

```
## [1] 0.0049
```

```
n3 = num3/deno3
ans3 = ceiling(n3)
ans3
```

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

339 visitors must be included in the study.

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.

H0: Proportion of Australians who have pets is less than or equal to the proportion of Americans who have pets H1: Proportion of Australians who have pets is greater than the proportion of Americans who have pets

```
n1 = 51
x1 = 32
p1cap = x1/n1
n2 = 63
x2 = 27
p2cap = x2/n2
pcap4 = (x1 + x2)/(n1 + n2)
pcap4
```

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

```
num4 = p1cap - p2cap
d4 = pcap4 * (1 - pcap4)
d4_ = (1/n1) + (1/n2)
den4 = d4 * d4_
deno4 = sqrt(den4)
z4 = num4/deno4
z4
```

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

```
pval = (1 - pnorm(z4))
pval
```

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

We reject the null hypothesis as pval $<$ alpha

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

```
tab <- matrix(c(22, 36, 10, 13, 22, 29, 7, 28, 26), nrow = 3,
    ncol = 3)
tab
```

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

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

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

From the test we got X-squared = 20.408 degree of freedom = 4 p-value = 0.0004147

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.

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

```
##    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
```

a. At the $\alpha=0.01$ significance level, use a Wilcox signed-rank test "by-hand" (i.

Ho: Increase equal to zero
H1: Increase not equal to 0

```r
asthma$difference = asthma$air - asthma$sulf.diox
asthma$rank = rank(abs(asthma$difference))
asthma
```

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

```
## 12      12  3.64      3.87      -0.23    5
## 13      13  3.89      9.25      -5.36   14
## 14      14  0.58      6.59      -6.01   15
## 15      15  9.50      6.17       3.33   10
## 16      16  0.93     10.93     -10.00   16
## 17      17  0.49     15.44     -14.95   17
```

```
t_plus = 2 + 6 + 3 + 10
t_plus
```

```
## [1] 21
```

```
t_neg = 4 + 7 + 1 + 8 + 9 + 11 + 12 + 13 + 5 + 14 + 15 + 16 +
    17
t_neg
```

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

```
t = t_plus - t_neg
t
```

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

```
n = 17
sigm = sqrt(((n) * (n + 1) * (2 * n + 1))/6)
sigm
```

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

```
zt = t/sigm
zt
```

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

```
p_ast = 2 * (pnorm(zt))
p_ast
```

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

P value (0.008607429) is less than alpha, we reject the null hypothesis.

b. Run the test again using the exact signed-ranked distribution (i.e., ‘wilcox.test()‘)

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

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

```
P value in the wilcox.test() is less than what we got from part a. But the
inference is still the same.
\vspace{10pt}
```

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]
healthy
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.

   H0: median daily caloric intake of the population of individuals suffering from bulimia is equal to the median caloric intake of the healthy population H1: median daily caloric intake of the population of individuals suffering from bulimia is not equal to the median caloric intake of the healthy population

```
bulimia <- read.csv("bulimia.csv")
bulimic <- bulimia$bulimic
healthy <- bulimia$health[1:15]
n_bul = 23
n_health = 15
wilcox.test(bulimic, healthy, exact = F, correct = F)
```

```
##
##  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
```

```
P value 0.0005927 is less than alpha so we reject the null hypothesis.
```

Short Answers:

- About how long did this assignment take you? Did you feel it was too long, too short, or reasonable? 3 hours. Reasonable length
- Who, if anyone, did you work with on this assignment? No one
- What questions do you have relating to any of the material we have covered so far in class? F test, Chi Square test, wilcoxon signed rank test