R Assignment 4

Alireza Sheikh-Zadeh, PhD

Document format: Follow the instructions given on the web page. Always review your solution word document before submission.

Plagiarism: You are not allowed to share your write-up with your peers. It's okay to advise your peers about how to solve a problem, but you never share your own write-up.

Problem 1: 25 points

Problem 2: 20 points

Problem 3: 15 points

Problem 4: 20 points

Format: 20 points

Problem 1

A quality characteristic of interest for a tea-bag-filling process is the weight of the tea in the individual bags. The label weight on the package indicates that the mean amount is 5.37 grams of tea in a bag. Problems arise if the bags are under-filled or if the mean amount of tea in a bag exceeds the label weight. The accompanying data are the weights, in grams, of a sample of 50 tea bags produced in one hour by a single machine (data file Teabags.csv).

```
data <- read.csv("https://goo.gl/ZCVUpc")
Xbar <- mean(data$Teabags)
sd <- sd(data$Teabags)
n = length(data$Teabags)
```

a) Construct a 95% confidence interval estimate for the population mean weight of the tea bags. Interpret the interval. (10 points)

```
# statistics +-CV*SE
# t_0.025

lb = Xbar - qt(0.975, df = 49) * sd/sqrt(n)
ub = Xbar + qt(0.975, df = 49) * sd/sqrt(n)
c(lb,ub)

## [1] 5.471323 5.531477

t.test(data$Teabags, conf.level = .95)

##
##  One Sample t-test
```

```
##
## data:  data$Teabags
## t = 367.58, df = 49, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.471323 5.531477
## sample estimates:
## mean of x
##    5.5014

# This confidence interval demonstrates the values that the true mean must
fall within.
```

b) Is the company meeting the requirement set forth on the label that the mean amount of tea in a bag is 5.37 grams? (5 points)

```
# No. The interval estimate is 5.471323 and 5.531477 which is higher than
5.37 grams. We reject the null because 5.37 grams does not fall between the
confidence interval.
```

c) Explain how to understand the 95% confidence interval via simulation. Use simulation in your answer. (10 points)

```
n = 50
nsim = 10000
true_mean=5.37
sd <- sd(data$Teabags)
ntot = n*nsim
rv = rnorm(ntot, true_mean, sd)
rvm = matrix(rv, nrow = nsim)
xbars = apply(rvm, 1, mean)
stdevs = apply(rvm, 1, sd)
lowers = xbars - qt(1-.05/2, n-1)*stdevs/sqrt(n)
uppers = xbars + qt(1-.05/2, n-1)*stdevs/sqrt(n)
mean( isIntheRange <- (lowers < true_mean & uppers > true_mean))

## [1] 0.9513

CI <- data.frame(lowers, uppers, isIntheRange)

# Going through the simulation of 10,000 samples. The confidence interval is
set to 95%. This means that there is a 5% chance that the True mean does not
fall within the upper and lower bound. Calculating the true means that falls
within the bound is at 95%. Looking at the Dataframe you can see the each
samples that provides a upper and lower bound with 5% percent not in the
range and is False and those that are within that is true at 95%.
```

Problem 2

This data is taken from one of the MBA classes at TTU and asked my students whether they had had breakfast that day? In the following code, we extract the breakfast data of male and female students.

```
mba <- read.csv("http://tiny.cc/fa18classData" )


male <- mba$today.breakfast[mba$gender=="Male"]

female <- mba$today.breakfast[mba$gender=="Female"]

tabMale <- table(male)
tabFemale <- table(female)
```

a) How many students are male and how many are female? (5 points)

```
# These are the amount of males and females in order

length(male)

## [1] 26

length(female)

## [1] 19
```

b) What proportion of male and female students said Yes for having breakfast that day? (5 points)

```
p.male = 18/26
p.male

## [1] 0.6923077

p.female = 12/19
p.female

## [1] 0.6315789

prop.table(tabMale)

## male
##        No       Yes
## 0.3076923 0.6923077

prop.table(tabFemale)

## female
##        No       Yes
## 0.3684211 0.6315789
```

```
# Males that said yes are proportionally 0.6923077
# Females that said yes are proportionally 0.6315789
```

c) Conduct a two sample proportion test; is there significant evidence that the proportion of male students had breakfast that date is different from female students with alpha = 0.05? Show your work (e.g., what is the test statistic, p-value?) (10 points)

```
SE = sqrt((p.male) * (1-p.male)/26 + (p.female)*(1-p.female)/19)
Phat.diff = 0
Zstat = (p.male - p.female)/SE
Zstat

## [1] 0.4247734

CV = qnorm(0.975)

Zstat > CV

## [1] FALSE

Zstat < -CV

## [1] FALSE

pvalue <- 2 * pnorm(-Zstat, 0, 1)
pvalue

## [1] 0.6710019

# Fail to Reject the null hypothesis
```

Problem 3

A manufacturing company is interested in whether they can save money by adopting a shorter training period while still achieving desired outcomes for employees. Researchers sampled 15 employees to participate in traditional 3-day training and 15 to participate in revised 2-day training. After the training was complete, the researchers compared exit test scores between the two groups (scores are shown in the following data).

```
score <- read.csv("http://tiny.cc/training_data")
```

a) In order to compare the two methods of training, what type of test we need to use? Are the data of two training methods dependent on each other? Why? (5 points)

```
#t test will be the choice and the two training are independent. They are two different set of samples.
```

b) At alpha = 0.05 and assuming that the population is normally distributed, is there significant evidence that the two methods achieve different results? (10 points)

```
library(e1071)
null.diff = 5
xy <- mean(score$traditional.training)
```

```r
xx <- mean(score$revised.training)
xm <- sd(score$traditional.training)
xw <- sd(score$revised.training)

n1 = 10
xbar1 = xy
s1 = xm

n2 = 10
xbar2 = xx
s2 = xw

k = kurtosis(score$traditional.training)
k2 = kurtosis(score$revised.training)
n1 > 10*abs(k); n > 10*abs(k2)

## [1] FALSE

## [1] TRUE

xbar.diff = xbar1-xbar2

muXbar.diff = null.diff
sigmaXbar.diff = sqrt(s1^2/n1 + s2^2/n2)

Tstat = (xbar.diff-null.diff)/sqrt(s1^2/n1 + s2^2/n2)
Tstat

## [1] -3.927127

df.t <- function(s1, s2, n1, n2){
  nom = (s1^2/n1 + s2^2/n2)^2
  denom = (s1^2/n1)^2 / (n1-1) + (s2^2/n2)^2 / (n2-1)
}

df = df.t(s1,s2,n1,n2)
df

## [1] 0.928612

alpha = 0.05
qt(1-alpha,df)

## [1] 7.203087

Tstat > qt(1-alpha, df)

## [1] FALSE

# lecture part 6 in module 8
```

```r
t.test(score$traditional.training, score$revised.training, alternative =
"two.sided", conf.level = .95)

##
##  Welch Two Sample t-test
##
## data:  score$traditional.training and score$revised.training
## t = -1.7784, df = 27.898, p-value = 0.08624
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.3126034  0.4459367
## sample estimates:
## mean of x mean of y
##  49.80000  52.73333

CV = qt(0.975, df = 27.898)
CV

## [1] 2.048745

tStat = 1.7784
tStat < -CV

## [1] FALSE

tStat > CV

## [1] FALSE

# Fail to reject the null. t = -1.7 is not smaller than -2
```

Problem 4

Use the TTU graduate student exit survey data.

```r
grad <- read.csv("http://westfall.ba.ttu.edu/isqs6348/Rdata/pgs.csv", header
= T)
```

Two variables of interest are FacTeaching, a 1,2,3,4,5 rating of teaching at TTU by the student, and COL, the college from which the student graduated.

a) Test the independence between FacTeaching and COL variables at alpha = 0.05. (10 points)

```r
tb <- table(grad$COL, grad$FacTeaching)
tb

##
##          1   2   3   4   5
##   AG     4  15  26  78  56
##   AR     3   4   6  16   4
##   AS    12  24 124 290 171
```

```
##   BA      9  28  44 116  66
##   DUAL    0   0   2   0   0
##   ED      3   6  26 113  93
##   EN      5  36  65 168  86
##   GR      0   3   8  27  15
##   HS      1   5  17  41  33
##   MC      0   0   3  25   6
##   VPA     4   7  10  37  44
```

```
chiTest <- chisq.test(tb)
```

```
## Warning in chisq.test(tb): Chi-squared approximation may be incorrect
```

```
chiTest
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tb
## X-squared = 106.54, df = 40, p-value = 5.864e-08
```

```
round(chiTest$p.value, 5)
```

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

```
ChiStat <- chiTest$statistic

CV = qchisq(0.975, df = 40)
CV
```

```
## [1] 59.34171
```

```
ChiStat > CV
```

```
## X-squared 
##      TRUE
```

```
# Teaching quality in different college is different. Reject H0 and the data
is dependent

# There is significant dependence in the DATA
```

b) Remove a row or column of the contingency table having a very low count. After removing the outlier data that you discovered, re-construct the independence test again. This answer is more precise. (10 points)

```
tab <- table(grad$COL, grad$FacTeaching)
tab.clean <- tab[-5,]
tab.clean
```

```
## 
##          1   2   3   4   5
##   AG     4  15  26  78  56
```

```
##   AR     3   4   6  16   4
##   AS    12  24 124 290 171
##   BA     9  28  44 116  66
##   ED     3   6  26 113  93
##   EN     5  36  65 168  86
##   GR     0   3   8  27  15
##   HS     1   5  17  41  33
##   MC     0   0   3  25   6
##   VPA   4   7  10  37  44

#---------------------------------------------



chiTest2 <- chisq.test(tab.clean)

## Warning in chisq.test(tab.clean): Chi-squared approximation may be
incorrect

chiTest2

##
##  Pearson's Chi-squared test
##
## data:  tab.clean
## X-squared = 96.526, df = 36, p-value = 1.957e-07

round(chiTest2$p.value, 5)

## [1] 0

ChiStat2 <- chiTest2$statistic

CV2 = qchisq(0.975, df = 36)
CV2

## [1] 54.43729

ChiStat2 > CV2

## X-squared
##      TRUE

# After cleaning out the DATA we notice that the p value has decreased and
the data is dependent due to the chi square of 0.
```