R Assignment 4

Jonathan De Los Santos

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

# Find t of (α/2) = .05/2 = 0.025
# Find lower and upper bounds
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
```

```
# Histogram  with values
#hist(rt(1000, df = 49))
#abline(v = c(qt(0.025, df = 49), qt(0.975, df = 49)), col = 'red')

# We can also use this super convenient function that our professor
#...didn't teach us until just now!!
# It was probably a good idea because
#...I definitely would have used this lazy method instead :)
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
```

With 95% confidence, the mean amount of tea in teabags is between 5.47 and 5.53 grams.

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) The mean amount listed on the label is not in the range of 5.47 and 5.53. The bags are on average being overfilled.

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

```
# Note: Realize, of course, that the distributions that produce the data cannot possi
bly be normal, but do the simulation using normal distribution anyway.

# Create normal random numbers based on the mean and stdev of the data.
# set a sample size equal to the size of the data:
n = 50
# simulate 10000 samples:
nsim = 10000
true_mean = 5.37
ntot = n*nsim
rv = rnorm(ntot, true_mean, sd)
# convert data to a matrix
rvm = matrix(rv, nrow = nsim)
# find the means and sd of rows. To do so you can use apply function:
xbar = apply(rvm, 1, mean)
stdev = apply(rvm, 1, sd)
# Then find 95% lower bound and upper bound for all simulated samples:
lower = xbar - qt(1-.05/2, n-1)*stdev/sqrt(n)
upper = xbar + qt(1-.05/2, n-1)*stdev/sqrt(n)
# We assume that the true mu is
# mu = true_mean(data$Teabags)
# What proportion of the constructed intervals contain the true mu.
mean(lower < true_mean & upper > true_mean)
```

```
## [1] 0.9464
```

```
CI <- data.frame(lower, upper)
```

Here we are analyzing the means resulting from 10,000 simulations of our teabag data.The result is the proportion means contained in our interval is .9473 which is less than 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)
tabMale
```

```
## male
##  No Yes
##   8  18
```

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

```
## female
##  No Yes
##   7  12
```

a) How many students are male and how many are female? (5 points) There are 36 males and 19 females for a total of 55 students.

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

```
# you can use prop.tabl() function: E.g., prop.table(tabMale)
prop.table(tabMale)
```

```
## male
##        No       Yes
## 0.3076923 0.6923077
```

```
prop.table(tabFemale)
```

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

```
# Or use a proportion
p.male = 8/26
p.female = 12/19
```

This means .692 of males and 0.632 of females said yes to having breakfast that day.

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)

```
# H0: p1 - p2 = 0
# H1: p1 - p2 not-equal 0

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] -2.265458
```

```
# Critical value (CV) = z_0.25
CV = qnorm(.975)

# Check both for two-sided
Zstat > CV
```

```
## [1] FALSE
```

```
Zstat < -CV
```

```
## [1] TRUE
```

We fail to reject the null hyothesis that there is no difference between male and female students who had breakfast. (I'm getting different results for Zstat for some reason. I've double checked my p.male/female inputs and the standard error equation so I'm not sure what I'm doing wrong. The professor's answer is)

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) We do not know the standard deviation so we will need a t-test, it is two-tail because no specific comparison is asked for.

These data are not dependent on each other, the training methods do not influence each other's 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)

```
# If you assume Mu1 and Mu2 are the true mean scores for traditional and revised trai
ning respectively, then you can set your hypothesis as follows:
# H0: Mu1 - Mu2 = 0
# H1: Mu1 - Mu2 not= 0
t.test(score$traditional.training, score $revised.training)
```

```
##
##  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
```

This t-score is within the CV bounds, therefore we fail to reject the null hypothesis and the trainings do not result in statistically significant tests.

Problem 4

Use the TTU graduate student exit survey data.

```
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)

```
# In this problem, we test the independence between two categorical variables. We did
this in Module 2 (conditional probability) by creating a contingency table and checki
ng whether P(A and B) = P(A)*P(B) or not. This time we want to implement a hypothesis
test to conclude whether the dependence between two variables is significant or not.
# Null hypothesis assumes two variables are independent.
# Alternative hypothesis assumes they are dependent to each other.
# You need to make a contingency table, then run a chi-squared test. Lecture 11, part
8 helps you to do this.

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
```

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

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

```
chiTest$statistic
```

```
## X-squared
##  106.5411
```

```
ChiStat <- chiTest$statistic

CV = qchisq(.975, df = 40)
CV
```

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

```
ChiStat > CV
```

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

The p-value is extremely low and Chi Statistic is greater than the critical value. there we can reject the null hypothesis, the variables are significantly dependent.

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)

```
# Hint: If you put colleges in rows of your contingency table, then the fifth row is associated with dual college and you can remove it as an outlier. This is how to remove row 5:
tab <- table(grad$COL, grad$FacTeaching)
# tab[-5,] = remove row 5
tab.clean <- tab[-5,]
#that's how you do it. Then repeat part a.
chisq.test(tab.clean)
```

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

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

This data is more precise without the “dirty” data obscuring the result.