Homework 9

Max Ryoo

Problem 1

Part A

```
set.seed(12181998)
setwd("/Users/maxryoo/Documents/Fall 2018/STAT3080/HW9")
datafatal <- read.csv("fatal accidents.csv")
months <- sort(unique(datafatal$Month))
counts <- sapply(1:length(months), function(x)
  length(which(datafatal$Month == months[x] & datafatal$State == "Virginia")))
fatal_bymonth <- data.frame(Month = months, Counts=counts)
print(fatal_bymonth)
```

```
##    Month Counts
## 1      1     47
## 2      2     51
## 3      3     55
## 4      4     53
## 5      5     58
## 6      6     65
## 7      7     50
## 8      8     71
## 9      9     50
## 10    10     66
## 11    11     69
## 12    12     87
```

I sorted teh months from 1 - 12. I then used the sapply function to return a vector containing the counts of entries if I subset by a certain month. I made a dataframe with those two vectors as fata_bymonth

Part B

```
fallandwinter <- fatal_bymonth$Counts[fatal_bymonth$Month %in% c(9,10,11,12,1,2)]
springandsummer <- fatal_bymonth$Counts[fatal_bymonth$Month %in% c(3,4,5,6,7,8)]
test.stat <- median(fallandwinter) - median(springandsummer)
print(test.stat)
```

```
## [1] 2
```

I subseted to make two vectors for the fall/winter months and springandsummer for spring/winter months. I took the median of both the vectors and subtracted it to get a difference of 2.

Part C

```
boot_samps <- replicate(10000, sample(months, replace=T))
boot_samps1 <- boot_samps[c(1:2,9:12),]
boot_samps2 <- boot_samps[c(3:8),]
medians1 <- apply(boot_samps1, 2, median)
medians2 <- apply(boot_samps2, 2, median)
boot_null <- medians1 - medians2
pval<-sum(boot_null >= test.stat) / 10000
print(pval)
```

```
## [1] 0.2743
```

To do this section I used two samples as if they were from the same population. I found the proportion of the bootstrapped samples where the meidan of half of the sample is greater than the median of the other half, which means winter vs summer. This was determined by ther test statistics. The pval will thus become 0.2708

Part D

The null hypothesis that we had was the there were no differences in the median number of accidents that were significant of fatal accidents in Virginia compared to the months of fall/winter and spring/summer. The althernative is that there was a difference where the assumption is the number of fatal accidents is greater in fall/winter compared to spring/summer. The p-value was shown to be 0.27, which means that the probability of this happening at random is 27%, which is a high percentage and it is just by chance we see this phenomena. The p-value is not low enough for us to say that this value is statistically significant, which means we fail to reject the null hypothesis.

Problem 2

Part A

```
data1 <- read.csv("data1.csv")
correlation <- cor(data1$V1, data1$V2)
print(correlation)
```

```
## [1] 0.524066
```

I read in the csv file and found the correlation through the corr function of the two columns.

Part B

```
samp <- replicate(10000, sample(1:nrow(data1), 13))
p_values <- apply(samp, 2, function(x)
t.test(data1$V1[x], data1$V2[x], alternative="two.sided", paired=T)$p.value)
proportion2.1 <- length(which(p_values <= 0.05))/10000
print(proportion2.1)
```

```
## [1] 0.048
```

We ran a Monte Carlo simulation of 10,000 repeitions of paired t test. we took 13 pairs to see if there werea difference between this and the population means.In this we saw that out of our values the proportion lower than 0.05 was 0.0504, which is close to the theortical threshold for rejection of the null. Becuase of this value being similar we fail to reject the null hypothesis of there being a difference becuase of the p-value being very close to 0.05

Part C

```
samp <- replicate(10000, sample(1:nrow(data1), 13))
p_values <- apply(samp, 2, function(x)
t.test(data1$V1[x], data1$V2[x], alternative="two.sided")$p.value)
proportion2.2<- length(which(p_values <= 0.05))/10000
print(proportion2.2)
```

```
## [1] 0.0098
```

We did a similar process as we did in part B. We can see that the probability of Type 1 error is at 0.0079, which menas only .79 percent of two-tailed t-test have p-values lower than 0.05, which would make a type1 error.

Problem 3

A

```
data2 <- read.csv("data2.csv")
correlation2 <- cor(data2$V1, data2$V2)
print(correlation2)
```

```
## [1] -0.52036
```

I read in the csv file and found the correlation through the corr function of the two columns, which was -0.52036.

B

```
samp <- replicate(10000, sample(1:nrow(data2), 13))
p_values <- apply(samp, 2, function(x)
t.test(data2$V1[x], data2$V2[x], alternative="two.sided", paired=T)$p.value)
proportion3.1 <- length(which(p_values <= 0.05))/10000
print(proportion3.1)
```

```
## [1] 0.0506
```

Our proportion of type 1 error for this data sert is 0.0508, which is very close to the theoretical value of 0.05. Same methodology was used as in Problem 2B.

C

```
samp <- replicate(10000, sample(1:nrow(data2), 13))
p_values <- apply(samp, 2, function(x)
t.test(data2$V1[x], data2$V2[x], alternative="two.sided")$p.value)
proportion3.2 <- length(which(p_values <= 0.05))/10000
print(proportion3.2)
```

```
## [1] 0.1051
```

When we do this for this data set with two sample t test we get that the proportion of p values less than 0.05 is 0.1126, which is higher than the expected of 0.05. This means that we would have falsely rejected the null hypothesis.

Problem 4

A

```
data3 <- read.csv("data3.csv")
correlation3 <- cor(data3$V1, data3$V2)
print(correlation3)
```

```
## [1] 0.002426237
```

The population correlation was 0.002426237 for this data.

B

```
samp <- replicate(10000, sample(1:nrow(data3), 13))
p_values <- apply(samp, 2, function(x)
t.test(data3$V1[x], data3$V2[x], alternative="two.sided", paired=T)$p.value)
proportion4.1 <- length(which(p_values <= 0.05))/10000
print(proportion4.1)
```

```
## [1] 0.0519
```

Our proportion of type 1 error for this data sert is 0.0478, which is very close to the theoretical value of 0.05. Same methodology was used as in Problem 2B.

C

```
samp <- replicate(10000, sample(1:nrow(data3), 13))
p_values <- apply(samp, 2, function(x)
t.test(data3$V1[x], data3$V2[x], alternative="two.sided")$p.value)
proportion4.2 <- length(which(p_values <= 0.05))/10000
print(proportion4.2)
```

```
## [1] 0.0469
```

Our proportion of type 1 error for this data sert is 0.0446, which is very close to the theoretical value of 0.05. Same methodology was used as in Problem 2C.

Problem 5

```
tableoutput5 <- rbind(Data1 = c(proportion2.1,proportion2.2),
                      Data2 = c(proportion3.1,proportion3.2),
                      Data3 = c(proportion4.1,proportion4.2))
colnames(tableoutput5) <- c("pair","two-sample")
print(tableoutput5)
```

```
##         pair two-sample
## Data1 0.0480     0.0098
## Data2 0.0506     0.1051
## Data3 0.0519     0.0469
```

A conclusion that we can see is that for all data sets it can be seen that the pair test seemed to hold values that were close to the theoretical value of 0.05. Also another key feauture to look at is that if the data set doesn't have a high correlation it wasn't show as much of a difference bewteen pair&two-sample compared to the datasets that have a high correlaiton. The genearl idea is that the pair test seemed to be more accuarte. ## References 1. https://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm 2. https://www.statisticshowto.datasciencecentral.com/standardized-test-statistic/