Assignment #6

Grisham Nathan

Nov 1, 2017

library(devtools)  
install\_github("nxskok/smmr")

## Skipping install of 'smmr' from a github remote, the SHA1 (8030a7d1) has not changed since last install.  
## Use `force = TRUE` to force installation

library(smmr)

## 4.

## a)

library(tidyverse)

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: readr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Warning: package 'purrr' was built under R version 3.4.2

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

url="http://www.utsc.utoronto.ca/~butler/c32/forehead.txt"  
forehead=read\_delim(url," ")

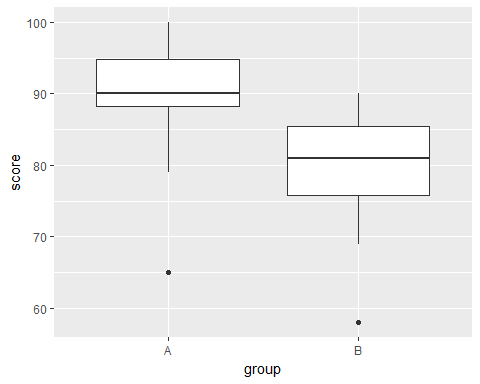
## Parsed with column specification:  
## cols(  
## group = col\_character(),  
## score = col\_integer()  
## )

forehead %>% slice(1:10)

## # A tibble: 10 x 2  
## group score  
## <chr> <int>  
## 1 A 88  
## 2 A 89  
## 3 A 79  
## 4 A 100  
## 5 A 98  
## 6 A 89  
## 7 A 65  
## 8 A 94  
## 9 A 95  
## 10 A 91

# b)

ggplot(forehead,aes(x=group,y=score))+geom\_boxplot()



# c)

Their is an outlier in each group and the boxplot for each group is not symmetric. As a result, this violates the assumption of the t-test that data must be normally distributed. As a result, we can conclude that our two-sample t-test is not valid.

# d)

m=median(forehead$score)  
m

## [1] 87.5

tab=with(forehead,table(group,score>m))  
tab

##   
## group FALSE TRUE  
## A 2 8  
## B 7 1

chisq.test(tab,correct=F)

## Warning in chisq.test(tab, correct = F): Chi-squared approximation may be  
## incorrect

##   
## Pearson's Chi-squared test  
##   
## data: tab  
## X-squared = 8.1, df = 1, p-value = 0.004427

Given the small p-value, we reject the null hypothesis of no association between group and the scores being above/below the overall median. As a result, both groups have different group medians. I recommend that the video be watched at 8:30 am, as Group A tends to have more scores above median than group B.

# e)

median\_test(forehead,score,group)

## $table  
## above  
## group above below  
## A 8 2  
## B 1 7  
##   
## $test  
## what value  
## 1 statistic 8.100000000  
## 2 df 1.000000000  
## 3 P-value 0.004426526

The Chi Squared Statistic, degrees of freedom and P-Value are the same as in d).

# f)

| **Obs** | **group** | **score** |
| --- | --- | --- |
| **1** | A | 88 |
| **2** | A | 89 |
| **3** | A | 79 |
| **4** | A | 100 |
| **5** | A | 98 |
| **6** | A | 89 |
| **7** | A | 65 |
| **8** | A | 94 |
| **9** | A | 95 |
| **10** | A | 91 |
| **11** | B | 87 |
| **12** | B | 69 |
| **13** | B | 78 |
| **14** | B | 79 |
| **15** | B | 83 |
| **16** | B | 90 |
| **17** | B | 85 |
| **18** | B | 58 |

filename myurl url "http://www.utsc.utoronto.ca/~butler/c32/forehead.txt";

proc import

datafile=myurl

dbms=dlm

out=forehead

replace;

delimiter=' ';

getnames=yes;

proc print data=forehead;

# g)

| **Median Scores (Number of Points Above Median) for Variable score Classified by Variable group** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **group** | **N** | **Sum of Scores** | **Expected Under H0** | **Std Dev Under H0** | **Mean Score** |
| **A** | 10 | 8.0 | 5.0 | 1.084652 | 0.8000 |
| **B** | 8 | 1.0 | 4.0 | 1.084652 | 0.1250 |
| **Average scores were used for ties.** | | | | | |

| **Median Two-Sample Test** | |
| --- | --- |
| **Statistic** | 1.0000 |
| **Z** | -2.7659 |
| **One-Sided Pr < Z** | 0.0028 |
| **Two-Sided Pr > |Z|** | 0.0057 |

| **Median One-Way Analysis** | |
| --- | --- |
| **Chi-Square** | 7.6500 |
| **DF** | 1 |
| **Pr > Chi-Square** | 0.0057 |



proc npar1way median;

var score;

class group;

The Chi-Square statistic of 7.65 is very close to 8.1. The small P-Value of 0.0057 is pretty close to 0.0044. The results I got from R are very similar to the results I got from SAS. Whether you use the results from either R or SAS, we will still conclude that the data being above/below overall median depends on the group.

## 5.

# a)

| **Analysis Variable : length** | | |
| --- | --- | --- |
| **rating** | **N Obs** | **N** |
| G | 15 | 15 |
| PG | 15 | 15 |
| PG-13 | 15 | 15 |
| R | 15 | 15 |

filename myurl url "http://www.utsc.utoronto.ca/~butler/c32/movie-lengths.csv";

proc import

datafile=myurl

dbms=csv

out=movie

replace;

getnames=yes;

proc means n;

var length;

class rating;

I have the correct ratings and I have the correct amount of each rating, as in the CSV file.

# b)

| **Class Level Information** | | |
| --- | --- | --- |
| **Class** | **Levels** | **Values** |
| **rating** | 4 | G PG PG-13 R |

|  |  |
| --- | --- |
| **Number of Observations Read** | 60 |
| **Number of Observations Used** | 60 |

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 3 | 14624.40000 | 4874.80000 | 11.72 | <.0001 |
| **Error** | 56 | 23294.53333 | 415.97381 |  |  |
| **Corrected Total** | 59 | 37918.93333 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **length Mean** |
| --- | --- | --- | --- |
| 0.385675 | 19.32606 | 20.39544 | 105.5333 |

| **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **rating** | 3 | 14624.40000 | 4874.80000 | 11.72 | <.0001 |



# 

|  |  |
| --- | --- |
| **Note:** | This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Alpha** | | | | 0.05 | |
| **Error Degrees of Freedom** | | | | 56 | |
| **Error Mean Square** | | | | 415.9738 | |
| **Critical Value of Studentized Range** | | | | 3.74468 | |
| **Minimum Significant Difference** | | | | 19.72 | |
| **Means with the same letter are not significantly different.** | | | | |
| **Tukey Grouping** | **Mean** | **N** | **rating** | |
| A | 123.400 | 15 | PG-13 | |
| A |  |  |  | |
| A | 111.200 | 15 | R | |
| A |  |  |  | |
| A | 106.933 | 15 | PG | |
|  |  |  |  | |
| B | 80.600 | 15 | G | |

proc anova;

class rating;

model length=rating;

means rating / tukey;

# c)

We reject the null hypothesis that the mean lengths of movies of all four rating types are equal, since P-value is less than 0.0001. Also the means for PG-13, R and PG are not significantly different. The rating G is significantly different from rest of the ratings.

# d)

url="http://www.utsc.utoronto.ca/~butler/c32/movie-lengths.csv"  
movie=read\_csv(url)

## Parsed with column specification:  
## cols(  
## length = col\_integer(),  
## rating = col\_character()  
## )

movie.aov=aov(length~rating, data=movie)  
summary(movie.aov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## rating 3 14624 4875 11.72 4.59e-06 \*\*\*  
## Residuals 56 23295 416   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(movie.aov)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = length ~ rating, data = movie)  
##   
## $rating  
## diff lwr upr p adj  
## PG-G 26.333333 6.613562 46.053104 0.0044541  
## PG-13-G 42.800000 23.080229 62.519771 0.0000023  
## R-G 30.600000 10.880229 50.319771 0.0007379  
## PG-13-PG 16.466667 -3.253104 36.186438 0.1327466  
## R-PG 4.266667 -15.453104 23.986438 0.9397550  
## R-PG-13 -12.200000 -31.919771 7.519771 0.3660019

The results in R and SAS are the same. The results from ANOVA table in both R and SAS are the same. The differences calculated in R are the same if I manually calculated from my SAS results.