Assignment 7

Grisham Nathan

November 4, 2017

## 8.

# a)

library(tidyverse)

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

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

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

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

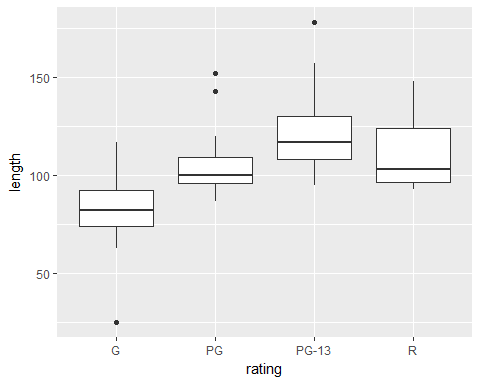
## Parsed with column specification:  
## cols(  
## length = col\_integer(),  
## rating = col\_character()  
## )

movie %>% group\_by(rating) %>% summarize(count=n(), median=median(length))

## # A tibble: 4 x 3  
## rating count median  
## <chr> <int> <int>  
## 1 G 15 82  
## 2 PG 15 100  
## 3 PG-13 15 117  
## 4 R 15 103

# b)

ggplot(data=movie,aes(x=rating, y=length)) +geom\_boxplot()



It is not reasonable to run ANOVA here because the data is not normally distributed. We conclude this from the existence of outliers and the unequal whiskers in the boxplots.

# c)

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)  
median\_test(movie,length,rating)

## $table  
## above  
## group above below  
## G 2 13  
## PG 7 7  
## PG-13 12 3  
## R 8 6  
##   
## $test  
## what value  
## 1 statistic 13.752380952  
## 2 df 3.000000000  
## 3 P-value 0.003262334

Given the small p-value, I conclude that the data being above/below overall median depends on the rating. At least one of the ratings has a median that is different from the others.

## 

## d)

**Code**

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 print data=movie;

proc npar1way median;

var length;

class rating;

**Output**

| **Median Scores (Number of Points Above Median) for Variable length Classified by Variable rating** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **rating** | **N** | **Sum of Scores** | **Expected Under H0** | **Std Dev Under H0** | **Mean Score** |
| **G** | 15 | 2.00 | 7.50 | 1.662778 | 0.133333 |
| **PG-13** | 15 | 12.00 | 7.50 | 1.662778 | 0.800000 |
| **PG** | 15 | 7.50 | 7.50 | 1.662778 | 0.500000 |
| **R** | 15 | 8.50 | 7.50 | 1.662778 | 0.566667 |
| **Average scores were used for ties.** | | | | | |

| **Median One-Way Analysis** | |
| --- | --- |
| **Chi-Square** | 13.9701 |
| **DF** | 3 |
| **Pr > Chi-Square** | 0.0029 |



Yes, it does give similar results with R. In both R and SAS, we get very close Chi-Square Statistic, same degrees of freedom and very low P-values.

## Analysis of Two Measurement Methods for the Atomic Weight of Carbon

# Introduction

The purpose of this study is to find out which one of the two methods better accurately measure the atomic weight of Carbon. Each observation of data specifies the method of measurement and the atomic weight of the carbon using that method of measurement.

# Data Output and Analysis

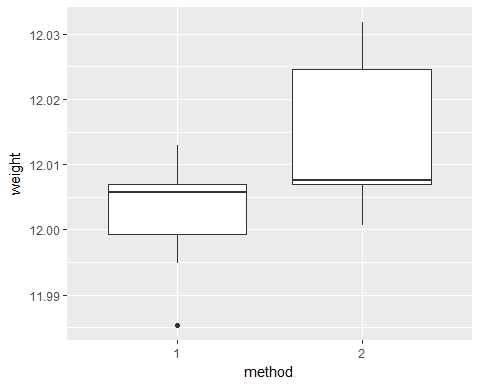
url="http://www.utsc.utoronto.ca/~butler/c32/carbon.txt"  
carbon=read\_delim(url," ")

## Parsed with column specification:  
## cols(  
## method = col\_integer(),  
## weight = col\_double()  
## )

carbon

## # A tibble: 15 x 2  
## method weight  
## <int> <dbl>  
## 1 1 12.0129  
## 2 1 12.0072  
## 3 1 12.0064  
## 4 1 12.0054  
## 5 1 12.0016  
## 6 1 11.9853  
## 7 1 11.9949  
## 8 1 11.9985  
## 9 1 12.0077  
## 10 1 12.0061  
## 11 2 12.0318  
## 12 2 12.0246  
## 13 2 12.0069  
## 14 2 12.0006  
## 15 2 12.0075

carbon$method = as.character(carbon$method)  
ggplot(carbon,aes(x=method, y=weight))+geom\_boxplot()



I converted the data type of the method column from numeric to characters to create this boxplot. I see that most of the weights in method 1 are lower than method 2 and are closer to the true atomic weight, which is 12. The medians of both methods are very close to each other.

# 

# Significance Testing of the Results

t.test(weight~method, data = carbon)

##   
## Welch Two Sample t-test  
##   
## data: weight by method  
## t = -1.817, df = 5.4808, p-value = 0.1238  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.027777288 0.004417288  
## sample estimates:  
## mean in group 1 mean in group 2   
## 12.00260 12.01428

median\_test(carbon,weight,method)

## $table  
## above  
## group above below  
## 1 3 6  
## 2 4 1  
##   
## $test  
## what value  
## 1 statistic 2.80000000  
## 2 df 1.00000000  
## 3 P-value 0.09426431

From conducting the t-test, we see that there is a large p-value. As a result, we accept our null hypothesis that the mean weight of method 1 is equal to mean weight of method 2.

Assuming that our data is not normally-distributed, the Mood's Median test is now conducted. From the test, we see that there is a large p-value. Therefore, we accept the null hypothesis that the median weight of method 1 is equal to median weight of method 2.

The hypothesis tests which have been conducted, indicate that there is statistical association between method 1 and method 2. As a result, our analysis should be sound and reasonable.