Jordan Mayer
STAT 350
Lab 08
April 5, 2018

Part B. Average Test Score across the regions in the United States

1. Code

```
#####
# Jordan Mayer
# STAT 350
# Lab 07
# March 29, 2018
#####

# setup
setwd("W:/Courses Spring 2018/STAT 350/STAT 350 Labs/Lab 08")
  # set working directory
library(ggplot2)  # set up ggplot2 for plotting
graphics.off()  # close any open figures
USData <- read.table("US_Data.txt", header=TRUE, sep="\t")  # get US data
US_clean <- USData[complete.cases(USData),]  # clean US Data
US_NE <- subset(US_clean, Region == "NE") # subset for Northeast region only
US_NC <- subset(US_clean, Region == "NC") # subset for North Central region
only
US_SO <- subset(US_clean, Region == "SO") # subset for South region only
US_WE <- subset(US_clean, Region == "WE") # subset for West region only

attach(US_clean)

### PART B ###
# data of interest: Average Test Score (TestScore) across regions of US
(Region)
# create side-by-side boxplots and effects plot
title = "Average Test Score by Region"
# side-by-side boxplots
box <- ggplot(US_clean, aes(x=Region,y=TestScore))+
  geom_boxplot()+
  stat_boxplot(geom="errorbar")+
  stat_summary(fun.y=mean,col="black",geom="point",size=3)+
  ggtitle(title)
ggsave(box,filename="box.jpg",width=6,height=6)
# effects plot
effects <- ggplot(data=US_clean,aes(x=Region,y=TestScore))+
  stat_summary(fun.y=mean,geom="point")+
  stat_summary(fun.y=mean,geom="line",aes(group=1))+
  ggtitle(title)
ggsave(effects,filename="effects.jpg",width=6,height=6)
```

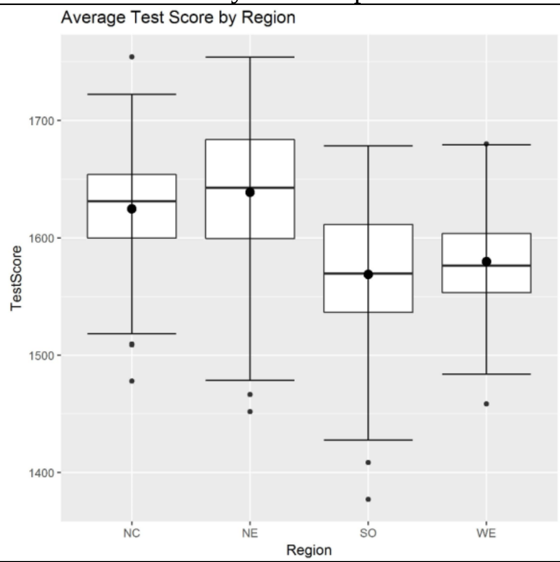
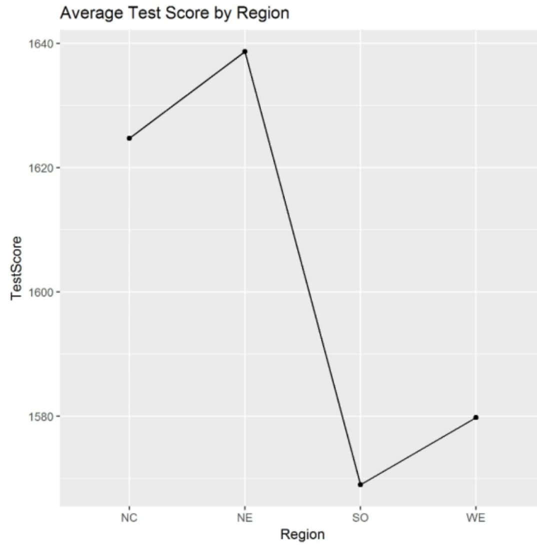
Jordan Mayer
STAT 350
Lab 08
April 5, 2018

```
# display sample statistics
tapply(TestScore, Region, length)  # display sample sizes
tapply(TestScore, Region, mean)  # display sample means
tapply(TestScore, Region, sd)  # display sample standard deviations

# check normality via histograms
# calculate theoretical density curves
xbar <- tapply(TestScore, Region, mean)
sd <- tapply(TestScore, Region, sd)
detach(US_clean)
US_clean$normal.density <- apply(US_clean, 1, function(x) {
  dnorm(as.numeric(x["TestScore"]),
        xbar[x["Region"]], sd[x["Region"]])
})
# create histograms
hist <- ggplot(US_clean,aes(x=TestScore))+
  geom_histogram(aes(y=..density..),bins=sqrt(nrow(US_clean))+2,
                 fill="grey",col="black")+
  facet_grid(Region ~ .)+
  geom_density(col="red",lwd=1)+
  geom_line(aes(y=normal.density),col="blue",lwd=1)+
  ggtitle(title)
ggsave(hist,filename="hist.jpg",width=6,height=6)

# check normality via normal probability plots
US_clean$intercept <- apply(US_clean, 1, function(x){xbar[x["Region"]]})
US_clean$slope <- apply(US_clean, 1, function(x){sd[x["Region"]]})
# create normal probability plots
qq <- ggplot(US_clean,aes(sample=TestScore))+
  stat_qq()+
  facet_grid(Region ~ .)+
  geom_abline(data=US_clean,aes(intercept=intercept,slope=slope))+
  ggtitle(title)
ggsave(qq,filename="qq.jpg",width=6,height=6)

# perform ANOVA significance test
fit <- aov(TestScore ~ Region, data=US_clean)
summary(fit)

# perform multiple-comparison via Tukey procedure
test.Tukey <- TukeyHSD(fit, conf.level=0.999)
test.Tukey
```

Jordan Mayer
STAT 350
Lab 08
April 5, 2018

2. Initial information

Plots:

Side-by-side boxplots	Effects plot
	
From this figure, it appears that all of the means could be the same due to the amount of overlap in the boxes. It is worth noting, however, that it is not entirely clear how much overlap, if any, exists between the North Central and Northeast boxplots and the West boxplot.	From this figure, it appears that the means are statistically different, with the Northeast region having the highest mean and the South region having the lowest.

Jordan Mayer
STAT 350
Lab 08
April 5, 2018

Code outputs:

```
> # display sample statistics
> tapply(TestScore, Region, length)  # display sample sizes
 NC  NE  SO  WE 
249 311 318 220 
> tapply(TestScore, Region, mean)  # display sample means
      NC       NE       SO       WE 
1624.718 1638.707 1568.982 1579.798 
> tapply(TestScore, Region, sd)  # display sample standard deviations
      NC       NE       SO       WE 
43.89106 59.09264 54.26599 36.53068 
```

Tabulated:

Region	**Sample size**	**Sample mean**	**Sample standard deviation**
North Central	249	1624.718	43.89106
Northeast	311	1638.707	59.09264
South	318	1568.982	54.26599
West	220	1579.798	36.53068

Jordan Mayer
STAT 350
Lab 08
April 5, 2018

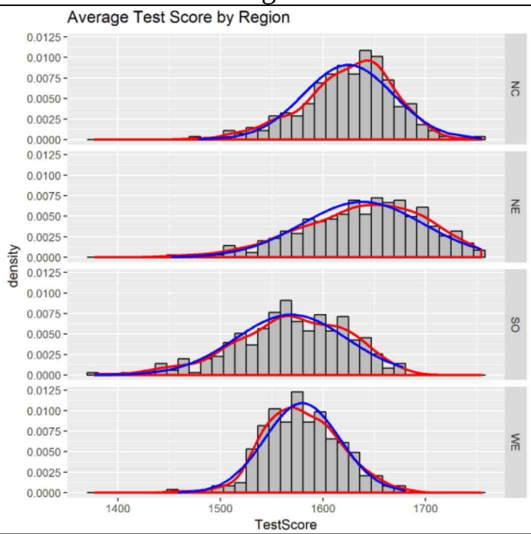
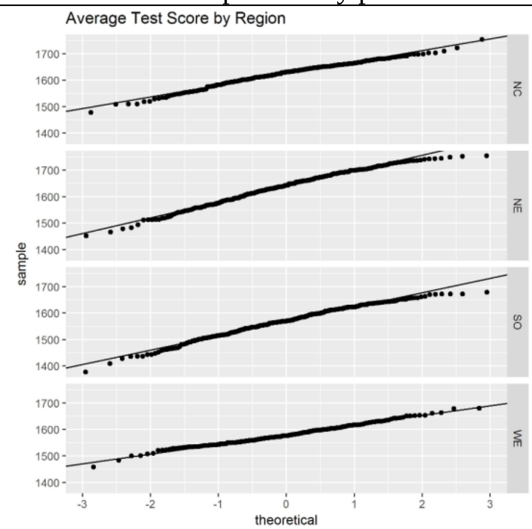
3. ANOVA Assumptions

1. Samples are independent SRSs (Simple Random Samples).

We cannot confirm this assumption graphically or numerically, but we can assume that it is true.

2. Populations are normally distributed.

We can examine this assumption using histograms and normal probability plots of the different populations.

<table>
<tr><th>histograms</th><th>normal probability plots</th></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>Based on this figure, the populations do appear to be normally distributed. Note that there is relatively little skew right or left in these histograms and that the tails do not appear to be extremely heavy or light.</td><td>Based on this figure, the populations do appear to be normally distributed. Note how most points on the normal probability plot are very close to the line of normality.</td></tr>
</table>

3. Populations have equal variance.

We can confirm this assumption using our sample variances, tabulated in Question 2. Specifically:

$$\frac{s_{max}}{s_{min}} = \frac{59.09264}{36.53068} = 1.618 < 2$$

Therefore, this assumption is valid.

Jordan Mayer
STAT 350
Lab 08
April 5, 2018

4. ANOVA significance test

1. Parameters of interest
   $\mu_{NC} = population\ mean\ Average\ Test\ Score\ in\ North\ Central\ Region$
   $\mu_{NE} = population\ mean\ Average\ Test\ Score\ in\ Northeast\ Region$
   $\mu_{SO} = population\ mean\ Average\ Test\ Score\ in\ South\ Region$
   $\mu_{WE} = population\ mean\ Average\ Test\ Score\ in\ West\ Region$
2. Hypotheses
   $H_0: \mu_{NC} = \mu_{NE} = \mu_{SO} = \mu_{WE}$
   $H_a: at\ least\ two\ \mu_i s\ are\ different$
3. Test statistic (F), degrees of freedom (Df), and p-value (Pr(>F))
   Code output:

```
> # perform ANOVA significance test
> fit <- aov(TestScore ~ Region, data=US_clean)
> summary(fit)
              Df  Sum Sq Mean Sq F value Pr(>F)    
Region         3 1000083  333361   130.9 <2e-16 ***
Residuals   1094 2786009    2547                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

   Values:
   $F_{ts} = 130.9$
   $DF_1 = 3$
   $DF_2 = 1094$
   $p = 2 * 10^{-16}$
4. Conclusion
   $\alpha = 0.001$

This data provides evidence (p-value = 2e-16) to the claim that the population mean Average Test Score of at least one of the US Regions is different from the rest.

This is consistent with the results of Question 2. From the effects plot, it did appear that the population means were statistically different, and the boxplots were unclear. The objective results of the ANOVA test have cleared up the uncertainties of the subjective results of the initial information.

Jordan Mayer
STAT 350
Lab 08
April 5, 2018

5. Tukey multiple-comparison test

We will perform this multiple-comparison test using the Tukey method because we want to compare all means in a pairwise fashion.

Code output:

```
> # perform multiple-comparison via Tukey procedure
> test.Tukey <- TukeyHSD(fit, conf.level=0.999)
> test.Tukey
  Tukey multiple comparisons of means
    99.9% family-wise confidence level

Fit: aov(formula = TestScore ~ Region, data = US_clean)

$Region
           diff         lwr       upr     p adj
NE-NC  13.98957   -2.174295  30.15344 0.0063070
SO-NC -55.73542  -71.819987 -39.65085 0.0000000
WE-NC -44.92018  -62.507781 -27.33259 0.0000000
SO-NE -69.72499  -84.883714 -54.56627 0.0000000
WE-NE -58.90975  -75.654819 -42.16469 0.0000000
WE-SO  10.81524   -5.853296  27.48377 0.0696635
```

To determine which pairs are significantly different, we could see if 0 is in the interval from "lwr" to "upr" (in which case there is no evidence for a difference), or we could simply check whether "p adj" is less than our significance level, 0.001 (in which case there is evidence for a difference). Whichever method we choose, we have evidence that the following pairs of Regions have different population mean Average Test Scores: (NC, SO), (NC, WE), (NE, SO), (NE, WE).

We can also represent these findings visually:

$\bar{x}_{SO}$	$\bar{x}_{WE}$	$\bar{x}_{NC}$	$\bar{x}_{NE}$
1568.982	1579.798	1624.718	1638.707

Our test and the corresponding figure tell us that the South and West regions have the same mean test score, as do the North Central and Northeast regions. However, the North Central and Northeast regions have significantly different mean test scores from the South and West regions. From our test, this is clear because the pairs (NC, SO), (NC, WE), (NE, SO), and (NE, WE) all have "p adj" values below 0.001 and increments ("lwr", "upr") that do not include 0. From our figure, this is also clear because the South and West sample means have a horizontal line beneath them, as do the North Central and Northeast sample means; but there is no horizontal line joining the North Central or Northeast values with the South or West values. In practical terms, this tells us that the South and West regions should improve their test preparation programs in order to achieve test scores closer to the North Central and Northeast regions.

Jordan Mayer
STAT 350
Lab 08
April 5, 2018

6. Explanation and conclusions

Our original goal was to compare the Average Test Score for the college entrance exam across the four regions of the United States. After validating the ANOVA assumptions, we performed an ANOVA analysis to infer about the average test scores in each region. These results showed no statistical difference between test scores in the South and West regions and no difference between scores in the North Central and Northeast regions; however, we do have evidence that the average test scores in the South and West regions are lower than those in the North Central and Northeast regions. However, it would be unwise to generalize this data to other college entrance exams, as different exams can differ substantially in many significant ways, such as the format, the length, and, most importantly, the material being tested.