**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)

# 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)

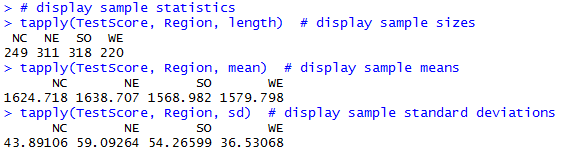
test.Tukey

**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. |

Code outputs:



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 |

**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.

|  |  |
| --- | --- |
| histograms | normal probability plots |
|  |  |
| 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. | 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. |

3. Populations have equal variance.

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

Therefore, this assumption is valid.

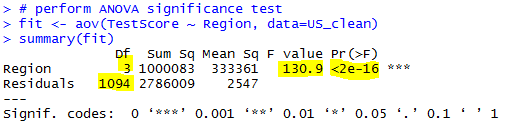
**4. ANOVA significance test**

1. Parameters of interest

2. Hypotheses

3. Test statistic (F), degrees of freedom (Df), and p-value (Pr(>F))

Code output:



Values:

4. Conclusion

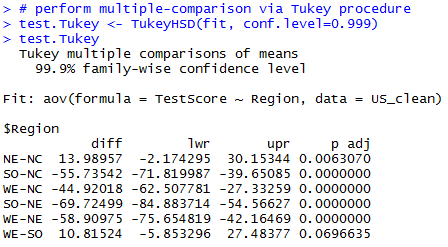
This data provides evidence (p-value = 2e-16) to the claim that the population mean Average Test Score of at least one of the US Regions is different from the rest.

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.

**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:



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:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
| 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.

**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.