Statistical Anova Analysis

Patrick Asztabski

4/7/2022

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
```

Cleaning our data

Down below is our functions. Everything below is already pre-run and nothing needs to be touched to examine the entire structure of this given code

```
remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}
```

This part is a portion of our data that is cleaned via R Programming

```
#####################################################################
heartData <- read.csv("heart_updated.csv")
strokeData <- read.csv("stroke_updated.csv")

strokeData[strokeData == "American Indian and Alaskan Native"] <- "AI/AN"
strokeData[strokeData == "Asian and Pacific Islander"] <- "A/PI"
heartData[heartData == "American Indian and Alaskan Native"] <- "AI/AN"
heartData[heartData == "Asian and Pacific Islander"] <- "A/PI"

#SUBSETTING DATA FOR OUTLIER REMOVAL
#####################################################################
whiteheartData = subset(x = heartData, subset = Race.Ethnicity == "White")
whitestrokeData = subset(x = strokeData, subset = Race.Ethnicity == "White")
BlackheartData = subset(x = heartData, subset = Race.Ethnicity == "Black")
BlackstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "Black")
APIheartData = subset(x = heartData, subset = Race.Ethnicity == "A/PI")
APIstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "A/PI")
HispanicheartData = subset(x = heartData, subset = Race.Ethnicity == "Hispanic")
HispanicstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "Hispanic")
AIANheartData = subset(x = heartData, subset = Race.Ethnicity == "AI/AN")
AIANstrokeData = subset(x = strokeData, subset = Race.Ethnicity == "AI/AN")

#REMOVING OUTLIERS
#####################################################################
whiteheartData$Deaths.per.100.000 <- remove_outliers(whiteheartData$Deaths.per.100.000)
whitestrokeData$Deaths.per.100.000 <- remove_outliers(whitestrokeData$Deaths.per.100.000)
whitestrokeData <- na.omit(whitestrokeData)
whiteheartData <- na.omit(whiteheartData)

BlackheartData$Deaths.per.100.000 <- remove_outliers(BlackheartData$Deaths.per.100.000)
BlackstrokeData$Deaths.per.100.000 <- remove_outliers(BlackstrokeData$Deaths.per.100.000)
BlackheartData <- na.omit(BlackheartData)
BlackstrokeData <- na.omit(BlackstrokeData)

APIheartData$Deaths.per.100.000 <- remove_outliers(APIheartData$Deaths.per.100.000)
APIstrokeData$Deaths.per.100.000 <- remove_outliers(APIstrokeData$Deaths.per.100.000)
APIheartData <- na.omit(APIheartData)
APIstrokeData <- na.omit(APIstrokeData)

HispanicheartData$Deaths.per.100.000 <- remove_outliers(HispanicheartData$Deaths.per.100.000)
HispanicstrokeData$Deaths.per.100.000 <- remove_outliers(HispanicstrokeData$Deaths.per.100.000)
HispanicstrokeData <- na.omit(HispanicstrokeData)
HispanicheartData <- na.omit(HispanicheartData)

AIANheartData$Deaths.per.100.000 <- remove_outliers(AIANheartData$Deaths.per.100.000)
AIANstrokeData$Deaths.per.100.000 <- remove_outliers(AIANstrokeData$Deaths.per.100.000)
AIANheartData <- na.omit(AIANheartData)
AIANstrokeData <- na.omit(AIANstrokeData)

#MERGING DATA
#####################################################################
mergedheartData <- rbind(whiteheartData,BlackheartData, APIheartData, HispanicheartData, AIANhea
```

```
rtData)
mergedstrokeData <- rbind(whitestrokeData,BlackstrokeData, APIstrokeData, HispanicstrokeData, AI
ANstrokeData)

#CREATING A REGION COLUMN BASED ON STATES AND THEIR RESPECTIVE REGIONS
#####################################################################
mergedheartData$region <- with(mergedheartData,
                          ifelse(LocationAbbr %in% c("AR","TN","LA","MS","AL","FL","GA","SC","NC"
),'Southeast',
                                 ifelse(LocationAbbr %in% c("HI","AK","OR","WA","ID","UT","NV","C
A","AZ"), 'Pacific West',
                                        ifelse(LocationAbbr %in% c("MT","ND","WY","SD","NE","CO",
"KS","NM","TX","OK"), 'Plains',
                                               ifelse(LocationAbbr %in% c("MN","WI","IA","IL","M
O","IN","KY","OH","MI"), 'MidWest',
                                                      ifelse(LocationAbbr %in% c("ME","NY","VT",
"NH","MA","MA","CT","RI","NJ","WV","VA","DE","MD", "DC", "PA"), 'Northeast',
                                                             ifelse(LocationAbbr %in% c("US"), 'O
verall', 'NA')
                                                      )
                                               )
                                        )
                                 )
                          )
)

mergedstrokeData$region <- with(mergedstrokeData,
                          ifelse(LocationAbbr %in% c("AR","TN","LA","MS","AL","FL","GA","SC","NC"
),'Southeast',
                                 ifelse(LocationAbbr %in% c("HI","AK","OR","WA","ID","UT","NV","C
A","AZ"), 'Pacific West',
                                        ifelse(LocationAbbr %in% c("MT","ND","WY","SD","NE","CO",
"KS","NM","TX","OK"), 'Plains',
                                               ifelse(LocationAbbr %in% c("MN","WI","IA","IL","M
O","IN","KY","OH","MI"), 'MidWest',
                                                      ifelse(LocationAbbr %in% c("ME","NY","VT",
"NH","MA","MA","CT","RI","NJ","WV","VA","DE","MD", "DC", "PA"), 'Northeast',
                                                             ifelse(LocationAbbr %in% c("US"), 'O
verall', 'NA')
                                                      )
                                               )
                                        )
                                 )
                          )
)

mergedheartData <- filter(mergedheartData, region != "Overall")
mergedstrokeData <- filter(mergedstrokeData, region != "Overall")
mergedheartData$region <- gsub("Overall", "U.S. Overall", mergedheartData$region)
mergedstrokeData$region <- gsub("Overall", "U.S. Overall", mergedstrokeData$region)
```

ANOVA TEST

```
# Hypotheses from Mean Model
# H_0: mu_A/PI = mu_AI/AN = mu_Black = mu_Hispanic = mu_White
# H_1: at least one of the means is different

# ANOVA method
dataAnova <- aov(mergedheartData$Deaths.per.100.000 ~ mergedheartData$Race.Ethnicity)
dataAnova2 <- aov(mergedstrokeData$Deaths.per.100.000 ~ mergedstrokeData$Race.Ethnicity)
summary(dataAnova)
```

```
##                                   Df    Sum Sq  Mean Sq F value Pr(>F)
## mergedheartData$Race.Ethnicity     4 112890782 28222695    4382 <2e-16 ***
## Residuals                      10607  68315735     6441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

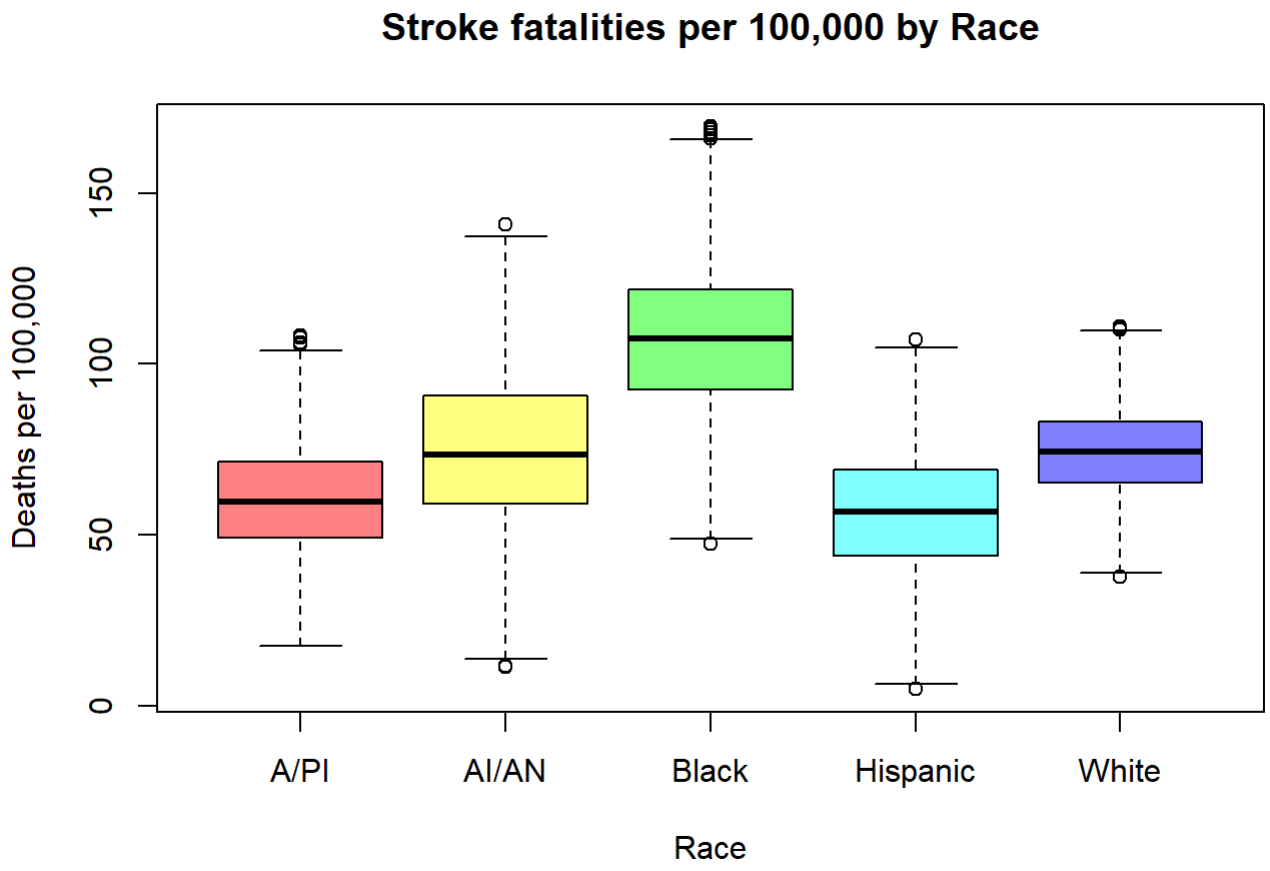
```
summary(dataAnova2)
```

```
##                                   Df  Sum Sq Mean Sq F value Pr(>F)
## mergedstrokeData$Race.Ethnicity    4 2136372  534093    1693 <2e-16 ***
## Residuals                       6579 2075351     315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

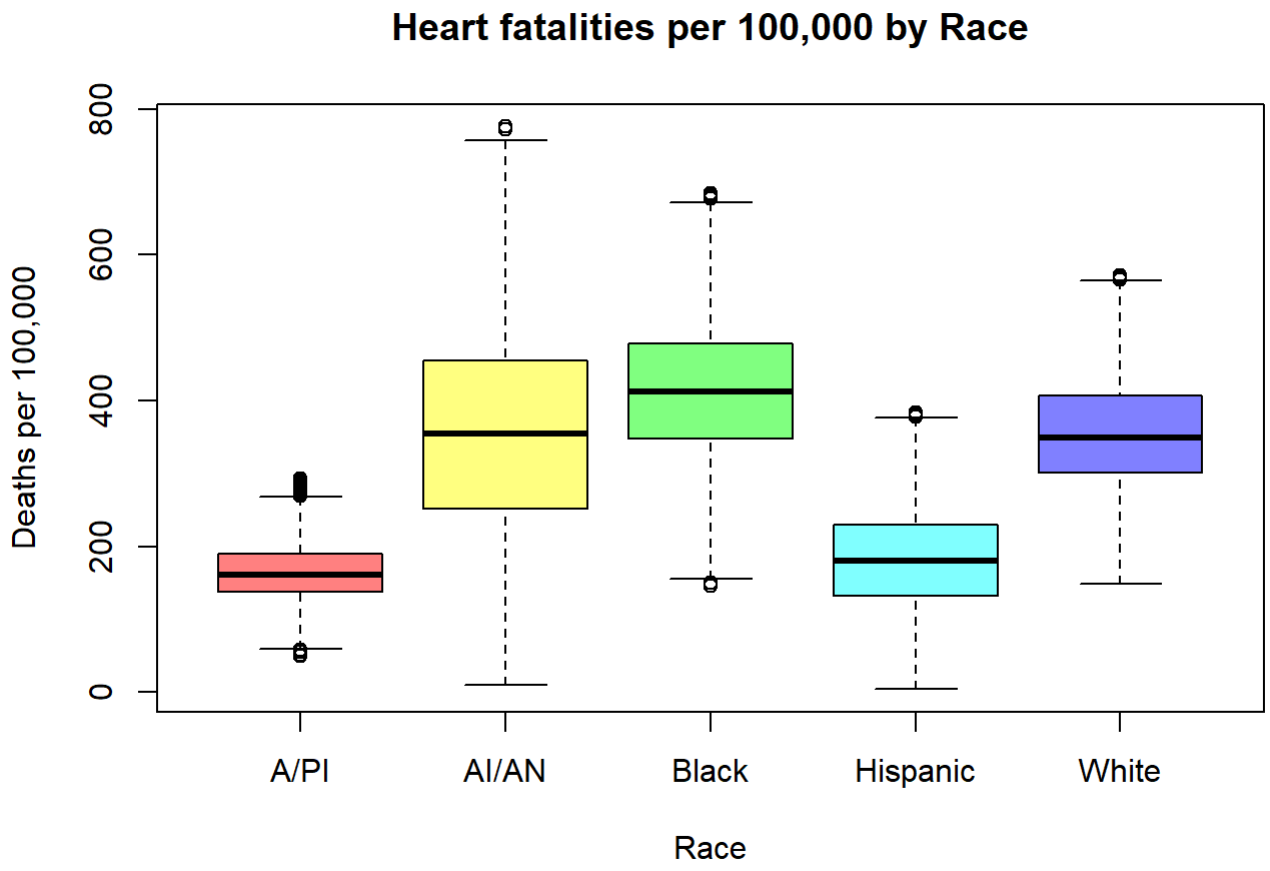
```
# p-value is represented by Pr(>F). It is very small for both tests, reject H_0.
# Conclude at least one mean is different statistically. There is a significant
# difference in the mortality trend between races.
```

##Exploratory Data Analysis

```
par(mfrow = c(1,1)) #can adjust this for 1,1 for individual view
rainbowcols <- rainbow(6,s = 0.5)
boxplot(mergedstrokeData$Deaths.per.100.000 ~ mergedstrokeData$Race.Ethnicity, xlab = "Race"
        , ylab = "Deaths per 100,000", col=c(rainbowcols), main = "Stroke fatalities per 100,000
by Race")
```




```
boxplot(mergedheartData$Deaths.per.100.000 ~ mergedheartData$Race.Ethnicity, xlab = "Race"
        , ylab = "Deaths per 100,000", col=c(rainbowcols), main = "Heart fatalities per 100,000
 by Race")
```




```
#Boxplots to distinguish outliers and present visualization on significant
#differences between multiple race/ethnicities


case.vector = tapply(mergedheartData$Deaths.per.100.000, mergedheartData$Race.Ethnicity, sum)
case.vector2 = tapply(mergedstrokeData$Deaths.per.100.000, mergedstrokeData$Race.Ethnicity, sum)

case.vector #Total number of deaths for this year from heart disease
```

```
##      A/PI     AI/AN     Black  Hispanic     White
##  489314.1  276464.0  864577.4  308747.5 1122484.0
```

```
case.vector2 #Total number of deaths for this year from stroke
```

```
##     A/PI    AI/AN    Black Hispanic    White
##  38794.3  16584.7 169321.3  58457.1 232007.3
```

```
aggregate(mergedheartData$Deaths.per.100.000, list(mergedheartData$Race.Ethnicity), FUN=mean) #A
verage number of deaths
```

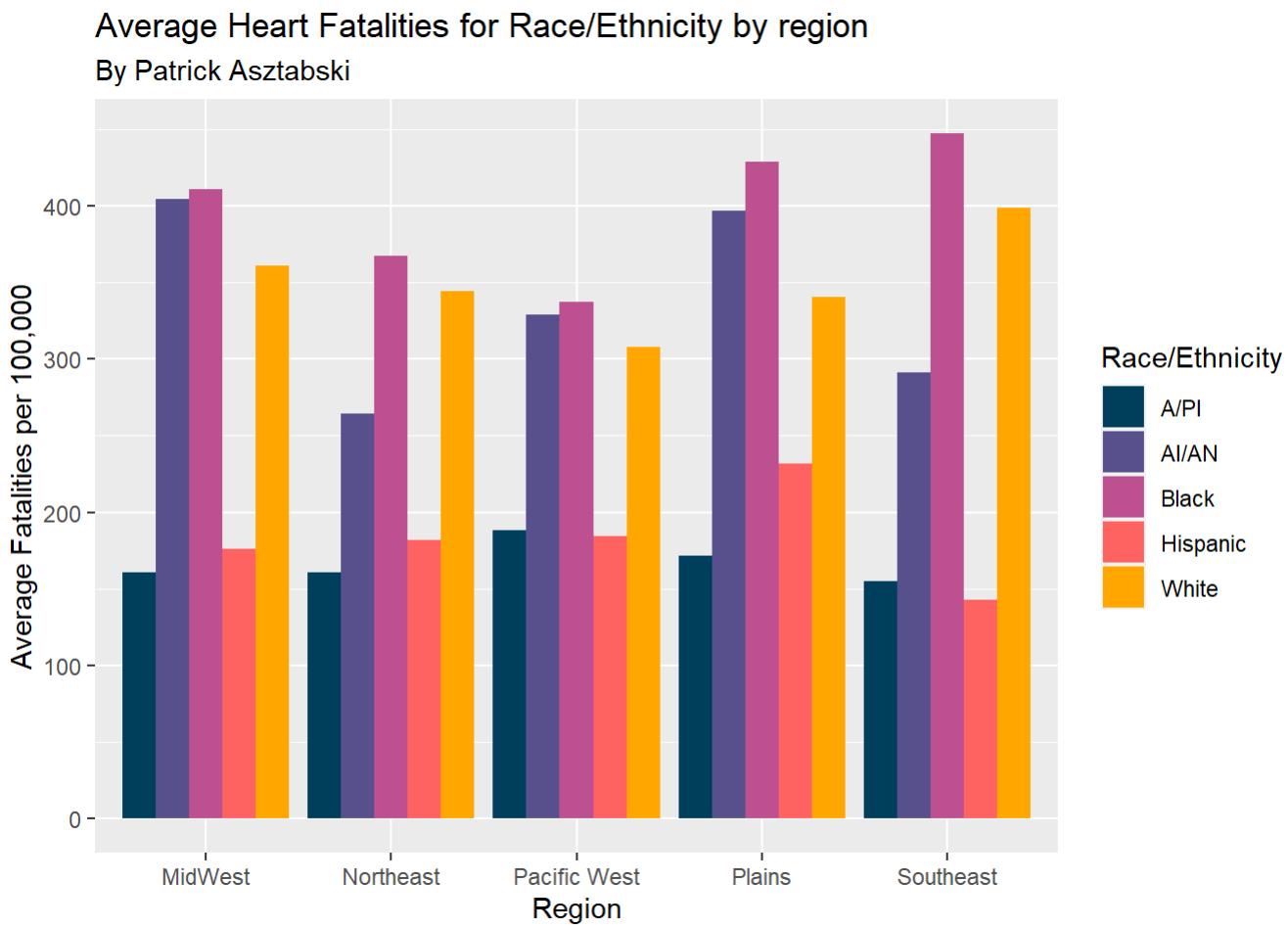
```
##    Group.1        x
## 1     A/PI 165.7568
## 2    AI/AN 358.1140
## 3    Black 412.8832
## 4 Hispanic 186.4417
## 5    White 357.7068
```

```
aggregate(mergedstrokeData$Deaths.per.100.000, list(mergedstrokeData$Race.Ethnicity), FUN=mean)
#Average number of deaths
```

```
##    Group.1         x
## 1     A/PI  60.90157
## 2    AI/AN  74.37085
## 3    Black 107.98552
## 4 Hispanic  56.10086
## 5    White  74.50459
```

##Additional Data Visualization

```
cbp1 <- c("#003f5c","#58508d","#bc5090","#ff6361","#ffa600")
x <- ggplot(data = mergedheartData) +
  geom_bar(mapping = aes(x=region, y = Deaths.per.100.000, fill=Race.Ethnicity), position = "dod
ge", stat = "summary", fun = "mean")
x + scale_fill_manual(values = cbp1) + labs(title = "Average Heart Fatalities for Race/Ethnicity
by region",
                                            x = "Region", y = "Average Fatalities per 100,000",
                                            subtitle = "By Patrick Asztabski", fill = "Race/Ethn
icity")
```




```
x <- ggplot(data = mergedstrokeData) +
  geom_bar(mapping = aes(x=region, y = Deaths.per.100.000, fill=Race.Ethnicity), position = "dod
ge", stat = "summary", fun = "mean")
x + scale_fill_manual(values = cbp1) + labs(title = "Average Stroke Fatalities for Race/Ethnicit
y by region",
                                            x = "Region", y = "Average Fatalities per 100,000",
                                            subtitle = "By Patrick Asztabski", fill = "Race/Ethn
icity")
```

Average Stroke Fatalities for Race/Ethnicity by region

By Patrick Asztabski