# Statistical Anova Analysis

Patrick Asztabski

4/7/2022

## 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)
```

# 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 112942064 28235516    4386 <2e-16 ***
## Residuals                      10614  68322556     6437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

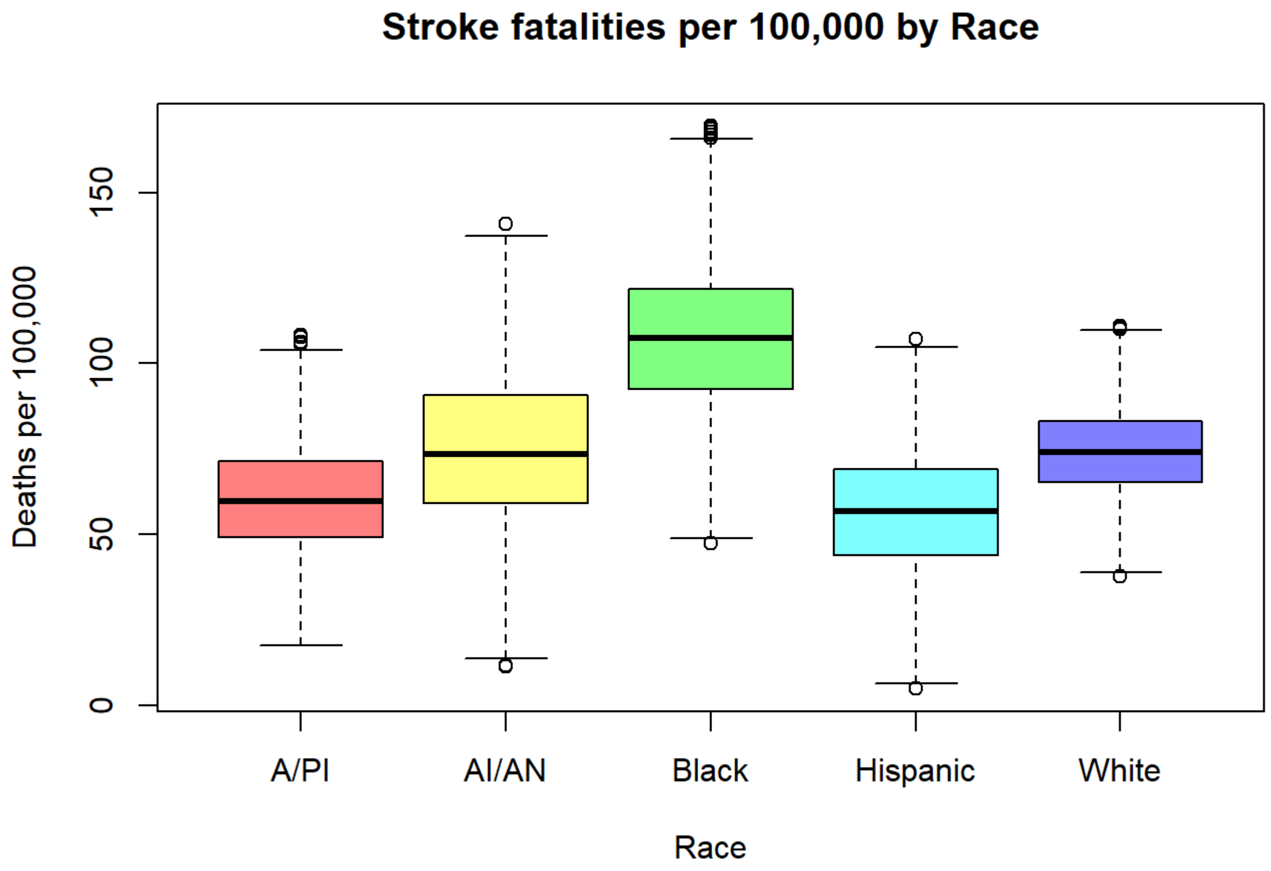
```
summary(dataAnova2)
```

```
##                                    Df  Sum Sq Mean Sq F value Pr(>F)
## mergedstrokeData$Race.Ethnicity    4 2137609  534402    1695 <2e-16 ***
## Residuals                       6584 2075619     315
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

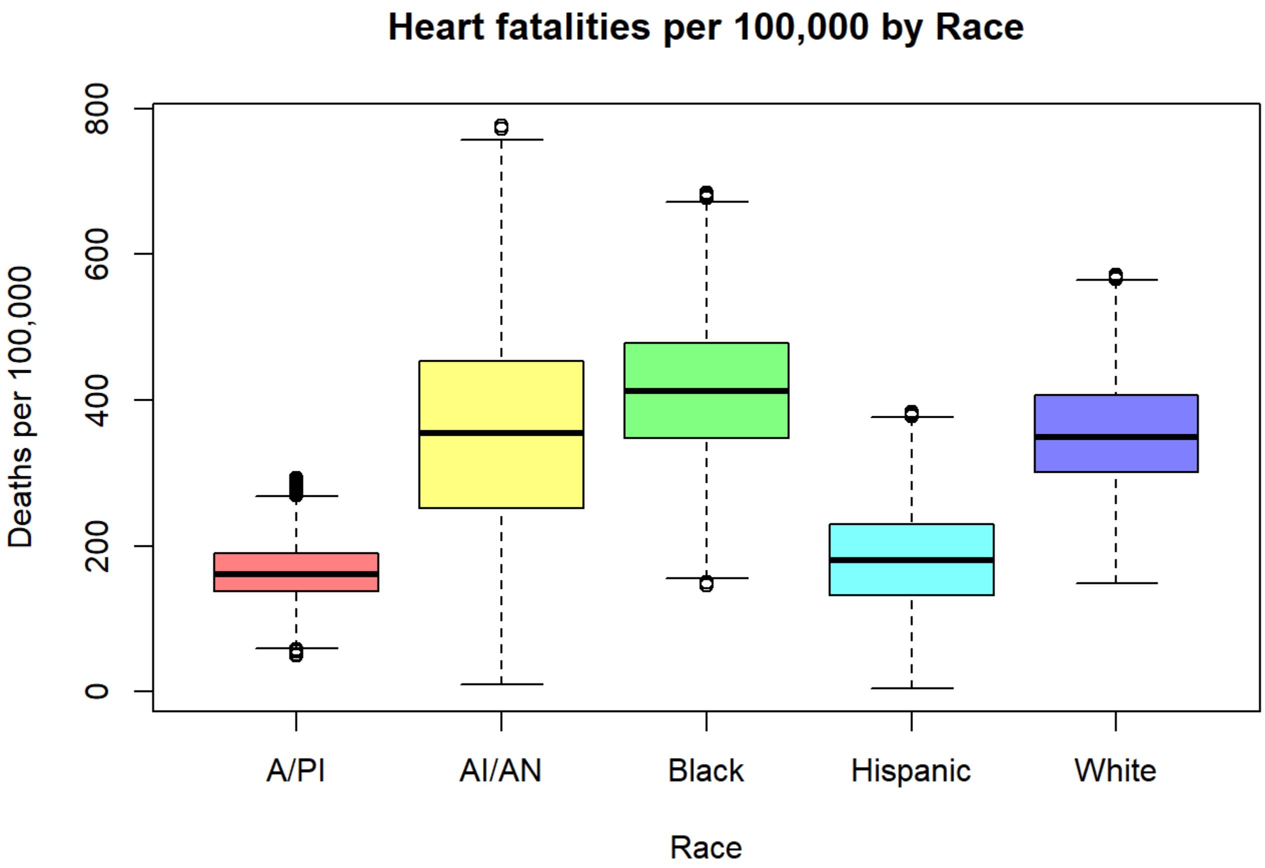
```
# p-value is represented by Pr(>F). It is very small for both tests, reject H_0.
# Conclude at least one mean is different statistically. There is a significant
# difference in the mortality trend between races.
```

# Additional 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
##  489847.8  276762.2  864993.3  308981.1 1122818.0
```

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

```
##     A/PI    AI/AN    Black Hispanic    White
##  38853.7  16646.0 169423.0  58519.5 232077.7
```

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

```
##    Group.1        x
## 1     A/PI 165.7691
## 2    AI/AN 358.0365
## 3    Black 412.8846
## 4 Hispanic 186.4702
## 5    White 357.6993
```

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

```
##    Group.1         x
## 1     A/PI  60.89922
## 2    AI/AN  74.31250
## 3    Black 107.98152
## 4 Hispanic  56.10690
## 5    White  74.50327
```