#two way anova

install.packages("corrplot")

library("corrplot")

install.packages("dplyr")

library("dplyr")

install.packages("plyr") #alternative installation of the %>%

library(plyr)

setwd("/Users/jemy/desktop/advanced statistics") #work computer

mortality <- read.csv("USRegionalMortality\_ANOVA.csv", sep=",", header=T)

attach(mortality)

head(mortality)

#lets now first test for normality to see if its normal or not.

#let's get the sd and mean of the number of deaths

sd(mortality$death.rate.per.100.000.population) #63.161

mean(mortality$death.rate.per.100.000.population) #56.90

median(mortality$death.rate.per.100.000.population) #29.2

hist(mortality$death.rate.per.100.000.population) #really does not look normal

shapiro.test(mortality$death.rate.per.100.000.population) # < 2.2e-16 p value is less than our alpha 0.05

#this means, anova baby

#ho1: that the means for all deaths causes are the same

#ho2: that gender deaths are the same means

#lets now verify the factors

levels(mortality$Cause) # returns: 9 values

levels(mortality$Sex) # returns: male, female

#verify as factors

mortality$Cause <- as.factor(mortality$Cause)

mortality$Sex <- as.factor(mortality$Sex)

#lets

boxplot(mortality$death.rate.per.100.000.population ~ mortality$Cause)

boxplot(mortality$death.rate.per.100.000.population ~ mortality$Sex)

# Two way factorial design

m1 = aov(mortality$death.rate.per.100.000.population ~ mortality$Cause + mortality$Sex + mortality$Cause:mortality$Sex)

summary(m1)

#here we can see that all the means are different including for both gender and causes of deahth.

#for a more realistic test, lets assume alpha is 0.0001

#we can see that cause of mortality does indeed have a big effect on the number of deaths.

#gender also has a big effect on the number of deaths

ph1 = TukeyHSD(m1, "mortality$Cause")

ph1

#there is a 99 per cent chance that suicide and pneumonia will have the same mean number of deaths

plot(ph1)

ph2 = TukeyHSD(m1, "mortality$Sex")

ph2

#the means difference is 20.49 between male and female.

#lets illustrate this.

gender <- mortality %>% select(death.rate.per.100.000.population, Sex)

male\_deaths <- gender$death.rate.per.100.000.population[gender$Sex == 'Male']

print(male\_deaths)

female\_deaths <- gender$death.rate.per.100.000.population[gender$Sex == 'Female']

print(female\_deaths)

boxplot(female\_deaths, male\_deaths, names=c("Female", "Male")) #this shows visually that there this a differene between male and female deaths.

#lets verify this by adding up all deaths by cancer

sum(mortality$death.rate.per.100.000.population[mortality$Cause=="Cancer"]) #6898.2

mean(mortality$death.rate.per.100.000.population[mortality$Cause=="Cancer"]) #172.455

boxplot(mortality\_rates\_cancer) #boxplot of mortality rates caused by cancer

#lets give this a name

mortality\_rates\_cancer <- (mortality$death.rate.per.100.000.population[mortality$Cause=="Cancer"])

#all deaths by heart disease

sum(mortality$death.rate.per.100.000.population[mortality$Cause=="Heart disease"]) #7068.6

mean(mortality$death.rate.per.100.000.population[mortality$Cause=="Heart disease"]) #176.715

mortality\_rates\_heartdisease <- (mortality$death.rate.per.100.000.population[mortality$Cause=="Heart disease"]) #7068.6

sum(mortality$death.rate.per.100.000.population[mortality$Cause=="Heart disease"]) #7068.6

sum(mortality$death.rate.per.100.000.population[mortality$Cause=="Cancer"]) #6898.2

boxplot(mortality\_rates\_heartdisease) #boxplot of mortality rates caused by heart dieseas

boxplot(mortality\_rates\_heartdisease,mortality\_rates\_cancer)

plot(mortality\_rates\_heartdisease,mortality\_rates\_cancer, pch = 15, col = c("red", "blue")) #red is heart diseas and cancer is red

cor(mortality\_rates\_heartdisease,mortality\_rates\_cancer) #very highn positive correlation

#diabetes and alzheimers have similar means of deaths rates with a difference of only 1.6300. The test has conluded this with a high certainty.

#lets do some power calculation to find out more. The problem is, after conducting the tukey test, we have seen some groups are similiar and some are very

#significantly different from each other.

#doing some extra testing on suicide and flu as anova gave 99% chance of these having similar means

mean(mortality$death.rate.per.100.000.population[mortality$Cause=="Suicide"]) #14.965

mean(mortality$death.rate.per.100.000.population[mortality$Cause=="Flu and pneumonia"]) #16.18

sum(mortality$death.rate.per.100.000.population[mortality$Cause=="Suicide"]) #598.6

sum(mortality$death.rate.per.100.000.population[mortality$Cause=="Flu and pneumonia"]) #647.2