Jordan Dever

Xfd461

# HW1

---

title: "STA6443 HW1"

author: "Jordan Dever"

format: docx

editor: visual

---

setwd() # set your own path

######################################################

# Exercise 1

########################################################

cars=read.csv("Cars.csv", header = TRUE) # read dataset

View(cars)

library(ggplot2)

library(BSDA)

### (a)

MPG\_Combo <- 0.6\*cars$MPG\_City+0.4\*cars$MPG\_Highway # combined mpg variable

cars=data.frame(cars, MPG\_Combo) # data frame with MPG\_Combo

boxplot(cars$MPG\_Combo)

qqnorm(cars$MPG\_Combo);qqline(cars$MPG\_Combo)

summary(cars$MPG\_Combo)

#This boxplot shows us that most of our cars in this data set are around

#our mean (22.54) and median (22.00) with a skew to the right with a lot of

# outliers to the right as well.

### (b)

boxplot(MPG\_Combo ~ Type, data = cars)

# In this boxplot, we can see that Sedans differ the most compared to the other

# 4 types of vehicles. The mean and medians of the SUV and Truck types looks to

# be outside the interquartile range of our first box box plot with both being

# at about ~16 MPG. In the first boxplot there were a few outliers to the left

# but in this plot there are none. There are still a few outliers to the right

# for the Sedan group and then just one outlier for the Sports and Truck groups.

# By looking at the plots for each, it looks like Sedan, Sports, Truck,

# and Wagon are skewed right and probably not normally distributed.

### (c)

summary(cars$Horsepower)

hist(cars$Horsepower)

boxplot(cars$Horsepower)

qqnorm(cars$Horsepower);qqline(cars$Horsepower, col ='red')

shapiro.test(cars$Horsepower)

# Horsepower seems to be skewed right and is not normally distributed. Just by

# looking at the histogram and QQplot I can see that the right skew seemed to

# be enough to make Horsepower not normal, but our Shapiro-Wilk test came back

# with a p-value < 0.0001 so we reject the null hypothesis of the data being

# normally distributed.

### (d)

boxplot(Horsepower ~ Type, data = cars)

shapiro.test(cars$Horsepower [cars$Type == 'Sedan'])

shapiro.test(cars$Horsepower [cars$Type == 'Sports'])

shapiro.test(cars$Horsepower [cars$Type == 'SUV'])

shapiro.test(cars$Horsepower [cars$Type == 'Truck'])

shapiro.test(cars$Horsepower [cars$Type == 'Wagon'])

# Using the Shapiro-Wilk test, Sedan came back not normal (Sedanp= 1.2e-07),

# Sports came back not normal (Sportsp = 0.01898), SUV came back not normal but

# was close to our alpha level of 0.05 (SUVp = 0.04423), Truck came back not

# normal(Truckp = 0.01697), and Wagon surprisingly came back normal even though

# its boxplot looks skewed right because its median is close to Q1 and not near

# the mean (Wagonp = 0.09525). Visually, all of the vehicle types look not

# normal in horsepower with the exception of SUV but with our p value being

# less than 0.05 we must reject the null hypothesis and state that SUV is not

# normally distributed.

######################################################

# Exercise 2

########################################################

#Perform a hypothesis test of whether SUV has a different Horsepower than Truck,

#and state your conclusions

#a) Which test should we perform, and why?

#Justify your answer based on findings on Exercise 1 (d).

# We should use the Wilcoxon Ranked Sum Test because we are comparing two

# populations and at least one is not normally distributed (but both aren't).

#b) Specify null and alternative hypotheses.

# H0 = SUV and Truck are from the same distribution - muSUV = muTruck (medians)

# HA = SUV and Truck are NOT from the same distribution - muSUV != muTruck

# One group has a higher median than the other

#c) State the conclusion based on the test result.

wilcox.test(cars$Horsepower [cars$Type == 'SUV'],

cars$Horsepower [cars$Type == 'Truck'], exact = F)

# In the Wilcoxon rank sum test, w = 806.5 and p = 0.3942. With this we can

# fail to reject the null hypothesis because p > 0.05 (alpha). Both medians

# of SUV and Truck come from the same distribution.

#######################################################

# Exercise 3

#######################################################

# Perform a hypothesis test -whether Wind in July has a different speed (mph)

# than Wind in August.

# a) Which test should we perform, and why? See QQ-plot and perform the

# Shapiro-Wilk test for normality check.

airqual = airquality

View(airqual)

boxplot(Wind ~ Month, data = airqual)

boxplot(airqual$Wind [airqual$Month==7], airqual$Wind [airqual$Month==8])

qqnorm(airqual$Wind [airqual$Month==7]); qqline(airqual$Wind [airqual$Month==7],

col = 'red')

qqnorm(airqual$Wind [airqual$Month==8]); qqline(airqual$Wind [airqual$Month==8],

col = 'blue')

shapiro.test(airqual$Wind [airqual$Month==7])

#alpha = 0.05

# w = 0.95003 p = 0.1564 > alpha

# fail to reject null hypothesis: July is normal

shapiro.test(airqual$Wind [airqual$Month==8])

# w = 0.98533 p = 0.937 > alpha

# fail to reject null: August is normal

# Since both are normally distributed visually and quantitatively, we can use a

# 2 sample t-test, but we will need to check is variance is equal to check if we

# should use a Pooled t-test or Satterwaithe t-test

#b) Specify null and alternative hypotheses

# t-test

# H0: mu July = mu August (mean), both means are equal to each other

# HA: mu July != mu August, the means are not equal

# f-test

# H0 July variance = August variance, both have equal variance

# HA July variance != August variance, unequal variance between groups

#c) State the conclusion based on the test result.

var.test(airqual$Wind [airqual$Month==7], airqual$Wind [airqual$Month==8],

alternative = 'two.sided')

# F = 0.8857, p variance = 0.7418

# We fail to reject the null hypothesis and state that variance is equal

# between groups. We will use pooled variance since variance is equal.

t.test(airqual$Wind [airqual$Month==7], airqual$Wind [airqual$Month==8],

alternative = 'two.sided',var.equal = T)

# t = 0.1865 p = 0.8527

# We fail to reject the null hypothesis and state that both groups have equal

# means. July and August have similar wind speeds (mph) throughout the month

# based on our analysis of the data.