Lab Assignment: Multiple Comparisons/Post Hoc Procedures

Christian Acosta

Oct. 9, 2015

Create a Word document from this R Markdown file for the following exercises. Submit the R markdown file and the knitted Word document via D2L Dropbox.

## Preliminaries

Cars were selected at random from among 1993 passenger car models that were listed in both the Consumer Reports issue and the PACE Buying Guide. Pickup trucks and Sport/Utility vehicles were eliminated due to incomplete information in the Consumer Reports source. Duplicate models (e.g., Dodge Shadow and Plymouth Sundance) were listed at most once. Use the data set Cars93 to do the following. (Type ?Cars93 to learn more about the data.)

For the first couple of exercises we are going to use the Cars93 data set from the MASS package. We'll delete the data having to do with vans so that we are only dealing with cars. The code to load and prepare the data is here:

if (!require(MASS)){  
 install.packages('MASS')  
 library(MASS)  
}

## Loading required package: MASS

data(Cars93)  
Cars93 <- Cars93[Cars93$Type != 'Van',]  
Cars93$Type <- factor(Cars93$Type) # recasting Type forces the factor levels to reset  
# shorten level labels to make them fit on boxplots  
# Cm = Compact  
# Lg = Large  
# Md = Midsize  
# Sm = Small  
# Sp = Sporty  
Cars93$Type <- factor(Cars93$Type,labels=c('Cm','Lg','Md','Sm','Sp'))  
head(Cars93)

## Manufacturer Model Type Min.Price Price Max.Price MPG.city MPG.highway  
## 1 Acura Integra Sm 12.9 15.9 18.8 25 31  
## 2 Acura Legend Md 29.2 33.9 38.7 18 25  
## 3 Audi 90 Cm 25.9 29.1 32.3 20 26  
## 4 Audi 100 Md 30.8 37.7 44.6 19 26  
## 5 BMW 535i Md 23.7 30.0 36.2 22 30  
## 6 Buick Century Md 14.2 15.7 17.3 22 31  
## AirBags DriveTrain Cylinders EngineSize Horsepower RPM  
## 1 None Front 4 1.8 140 6300  
## 2 Driver & Passenger Front 6 3.2 200 5500  
## 3 Driver only Front 6 2.8 172 5500  
## 4 Driver & Passenger Front 6 2.8 172 5500  
## 5 Driver only Rear 4 3.5 208 5700  
## 6 Driver only Front 4 2.2 110 5200  
## Rev.per.mile Man.trans.avail Fuel.tank.capacity Passengers Length  
## 1 2890 Yes 13.2 5 177  
## 2 2335 Yes 18.0 5 195  
## 3 2280 Yes 16.9 5 180  
## 4 2535 Yes 21.1 6 193  
## 5 2545 Yes 21.1 4 186  
## 6 2565 No 16.4 6 189  
## Wheelbase Width Turn.circle Rear.seat.room Luggage.room Weight Origin  
## 1 102 68 37 26.5 11 2705 non-USA  
## 2 115 71 38 30.0 15 3560 non-USA  
## 3 102 67 37 28.0 14 3375 non-USA  
## 4 106 70 37 31.0 17 3405 non-USA  
## 5 109 69 39 27.0 13 3640 non-USA  
## 6 105 69 41 28.0 16 2880 USA  
## Make  
## 1 Acura Integra  
## 2 Acura Legend  
## 3 Audi 90  
## 4 Audi 100  
## 5 BMW 535i  
## 6 Buick Century

Here is another trick which will simply your analysis a bit. You can attach a data frame so that it's simple to refer to the variables.

attach(Cars93)  
summary(Price) # Price is one of the variables in the Cars93 data frame, after attaching we don't have to refer to the data frame. Don't forget to detach(Cars93) after you're done.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7.40 11.75 16.40 19.55 24.75 61.90

## Exercise 1

We are going to look for differences in population mean engine revolutions per minute at maximum horsepower (RPM) of the different types of cars (Type). Assume that the RPM distributions are normal and have equal variances for the different types of cars. To use onewayComp() you'll have to load the DS705data package.

### Part 1a

Use a one step procedure to find a family of 95% simultaneous confidence intervals for the difference in population means.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

require(DS705data)

## Loading required package: DS705data

carhoc <- onewayComp(RPM~Type, data = Cars93, var.equal = T)  
carhoc

## $call  
## onewayComp(formula = RPM ~ Type, data = Cars93, var.equal = T)  
##   
## $comp  
## diff lwr upr t p p adj  
## Lg-Cm -689.77273 -1271.0080 -108.5374 -3.3131069 1.393808e-03 1.182831e-02  
## Md-Cm -26.13636 -513.7175 461.4448 -0.1496510 8.814214e-01 9.998848e-01  
## Sm-Cm 270.83333 -221.6116 763.2783 1.5354153 1.286774e-01 5.429018e-01  
## Sp-Cm 30.35714 -512.7219 573.4362 0.1560556 8.763873e-01 9.998639e-01  
## Md-Lg 663.63636 115.6425 1211.6303 3.3809275 1.125310e-03 9.651980e-03  
## Sm-Lg 960.60606 408.2801 1512.9320 4.8554702 5.951613e-06 5.744492e-05  
## Sp-Lg 720.12987 122.2195 1318.0402 3.3624525 1.193175e-03 1.020512e-02  
## Sm-Md 296.96970 -155.7607 749.7001 1.8312763 7.082930e-02 3.632245e-01  
## Sp-Md 56.49351 -450.8503 563.8373 0.3108691 7.567189e-01 9.979305e-01  
## Sp-Sm -240.47619 -752.4960 271.5437 -1.3111932 1.935909e-01 6.851085e-01  
## rej H\_0  
## Lg-Cm 1  
## Md-Cm 0  
## Sm-Cm 0  
## Sp-Cm 0  
## Md-Lg 1  
## Sm-Lg 1  
## Sp-Lg 1  
## Sm-Md 0  
## Sp-Md 0  
## Sp-Sm 0  
##   
## $pairw  
##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: RPM and Type   
##   
## Cm Lg Md Sm   
## Lg 0.0118 - - -   
## Md 0.9999 0.0097 - -   
## Sm 0.5429 5.7e-05 0.3632 -   
## Sp 0.9999 0.0102 0.9979 0.6851  
##   
## P value adjustment method: one.step

### Part 1b

Use the multcompView package to produce a boxplot and letters/T display illustrating the differences in population means. (Review slides 17-19 in Multiple Comparisons Part 2 presentation.)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

#install.packages("multcomp")  
require("multcomp")

## Loading required package: multcomp

## Warning: package 'multcomp' was built under R version 3.1.3

## Loading required package: mvtnorm

## Warning: package 'mvtnorm' was built under R version 3.1.3

## Loading required package: survival

## Loading required package: splines

## Loading required package: TH.data

## Warning: package 'TH.data' was built under R version 3.1.3

##   
## Attaching package: 'TH.data'

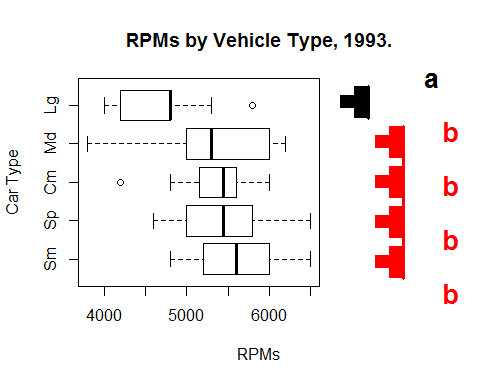
## The following object is masked from 'package:MASS':  
##   
## geyser

#install.packages("multcompView")  
require(multcompView)

## Loading required package: multcompView

## Warning: package 'multcompView' was built under R version 3.1.3

padj\_extract <- function(formula, data){carhoc$comp[,'p adj']}  
multcompBoxplot(RPM~Type, data = Cars93, horizontal = T, compFn="padj\_extract")  
title(main = "RPMs by Vehicle Type, 1993.", ylab = "Car Type", xlab = "RPMs")



?multcompBoxplot

## starting httpd help server ...

## done

### Part 1c

Summarize your findings about the differences in population mean RPM for the different types of cars. Which means are different and by how much? You should start with "We are 95% confident that ..."

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

We are 95% confident that the population mean RPM for large 1993 vehicles is different than all other 1993 vehicle types - confidence intervals provided below. Lg vs. Md [-1211.6, -115.6] Lg vs. Cm [-1271, -108] Lg vs. Sp [-1318, -122.2] Lg vs. Sm [-1512.9, -408.3] ---

### Part 1d

Use the REGWQ multi-step procedure (function regwqComp in DS705data) to test for pairwise differences in population mean RPM at (Don't forget each comparison includes an adjusted P-value and an adjusted significance level. See the presentation for more details.) How do the results compare to the one-step procedure you chose in 1b)?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

carhocReg <- regwqComp(RPM~Type)  
carhocReg

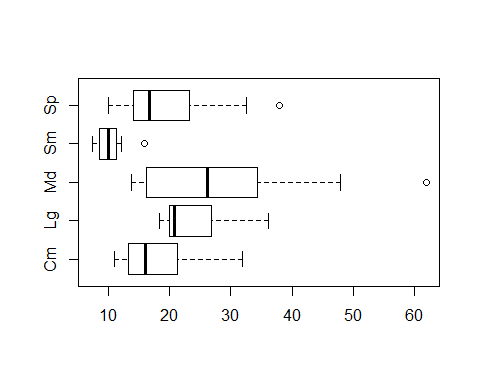
## diff t p p adj alpha adj rej H\_0  
## Lg-Cm -689.77273 -3.3131069 1.393808e-03 3.942242e-03 0.03030722 1  
## Md-Cm -26.13636 -0.1496510 8.814214e-01 8.814214e-01 0.02030827 0  
## Sm-Cm 270.83333 1.5354153 1.286774e-01 2.800479e-01 0.03030722 0  
## Sp-Cm 30.35714 0.1560556 8.763873e-01 8.763873e-01 0.02030827 0  
## Md-Lg 663.63636 3.3809275 1.125310e-03 1.125310e-03 0.02030827 1  
## Sm-Lg 960.60606 4.8554702 5.951613e-06 5.744492e-05 0.05000000 1  
## Sp-Lg 720.12987 3.3624525 1.193175e-03 6.427215e-03 0.05000000 1  
## Sm-Md 296.96970 1.8312763 7.082930e-02 2.665470e-01 0.05000000 0  
## Sp-Md 56.49351 0.3108691 7.567189e-01 9.481589e-01 0.03030722 0  
## Sp-Sm -240.47619 -1.3111932 1.935909e-01 1.935909e-01 0.02030827 0

## The hypothesis test produced similar results (minus the intervals). We were able to reject the null hypothesis of equal means (RPMs) between large cars compared to all other cars.

## Exercise 2

Now we are going to analyze differences in prices for different types of cars in the Cars93 data set. The boxplot below shows that the prices are skewed and variances are different.

boxplot(Price~Type,horizontal=TRUE)



### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

It should be fairly clear that the price data is not from normal distributions, at least for several of the car types, but ignore that for now and use the Games-Howell procedure with confidence level 90% to do simultaneous comparisons (if interpreting the -values use ).

# Use Games-Howell procedure @ 90% confidence.   
priceTypeComp <- onewayComp(Price~Type, data = Cars93, var.equal = F, alpha = 0.1)$comp  
priceTypeComp

## diff lwr upr t p  
## Lg-Cm 6.087500 -0.5685001 12.7435001 2.3977076 2.524630e-02  
## Md-Cm 9.005682 1.0428168 16.9685468 2.9017065 6.489106e-03  
## Sm-Cm -8.045833 -12.6632617 -3.4284050 -4.6636919 2.240512e-04  
## Sp-Cm 1.180357 -5.8653572 8.2260714 0.4357599 6.666721e-01  
## Md-Lg 2.918182 -5.4240153 11.2603789 0.9010497 3.745297e-01  
## Sm-Lg -14.133333 -19.6218454 -8.6448213 -7.2190114 1.705110e-05  
## Sp-Lg -4.907143 -12.3994478 2.5851621 -1.7142915 9.992807e-02  
## Sm-Md -17.051515 -24.0032322 -10.0997981 -6.4360269 1.739174e-06  
## Sp-Md -7.825325 -16.4785915 0.8279421 -2.3196827 2.650442e-02  
## Sp-Sm 9.226190 3.3048300 15.1475510 4.2447828 8.107007e-04  
## p adj rej H\_0  
## Lg-Cm 1.266833e-01 0  
## Md-Cm 3.759910e-02 1  
## Sm-Cm 1.192465e-04 1  
## Sp-Cm 9.923626e-01 0  
## Md-Lg 8.956675e-01 0  
## Sm-Lg 2.789471e-09 1  
## Sp-Lg 4.312457e-01 0  
## Sm-Md 8.678284e-08 1  
## Sp-Md 1.495367e-01 0  
## Sp-Sm 5.542557e-04 1

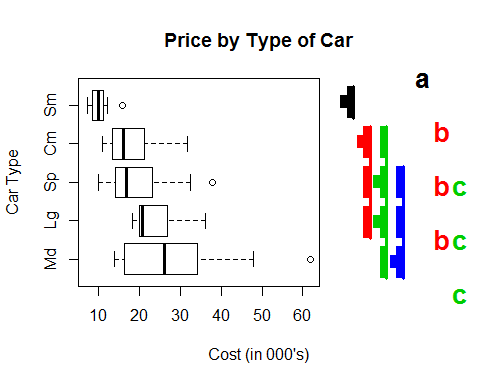
#?onewayComp

### Part 2b

Use the multcompView package to produce a boxplot and letters/T display illustrating the differences in population means. We want to make the comparisons at , but the multcompBoxplot command assumes and that is difficult to change. So instead divide the adjusted p-values by 2 before calling the multcompBoxplot (something like this should work: out$comp[,'p adj'] <- out$comp[,'p adj']/2 ).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

padj\_extract2 <- function(formula, data){priceTypeComp[,'p adj']}  
priceTypeComp[,'p adj'] <- priceTypeComp[,'p adj']/2  
multcompBoxplot(Price~Type, data = Cars93, horizontal = T, compFn="padj\_extract2")  
  
  
#carhoc <- onewayComp(RPM~Type, data = Cars93, var.equal = T)  
title(main = "Price by Type of Car", ylab = "Car Type", xlab = "Cost (in 000's)")



### Part 2c

Summarize the differences in the population mean prices for the different cars at . Since you have confidence intervals you should explain how the mean prices differ and by how much.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

We are 90% confident that the population mean price for small cars is less than all other car types - 90% confidence intervals are below (in thousands of dollars). Sm vs. Cm [-12.7, -3.4] Sm vs. Sp [-15.1, -3.3] Sm vs. Lg [-19.6, -8.6] Sm vs. Md [-24.0, -10.1]

We are 90% confident that the population mean price for compact cars is different from Md(medium?) cars [-17.0, -1.0].

## Exercise 3.

Since the price data is likely not normally distributed, the Games-Howell procedure was not entirely appropriate. However we can use bootstrapping to estimate the P-values and confidence intervals since the theoretical sampling distribution is likely not accurate.

### Part 3a

Repeat part 2a) using bootstrapping, by setting nboot=10000, in the onewayComp() function.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

#install.packages("boot")  
require(boot)

## Loading required package: boot

## Warning: package 'boot' was built under R version 3.1.3

##   
## Attaching package: 'boot'

## The following object is masked from 'package:survival':  
##   
## aml

bootpriceType <- onewayComp(Price~Type, data = Cars93, var.equal = F, alpha = 0.1, nboot=10000)$comp  
bootpriceType

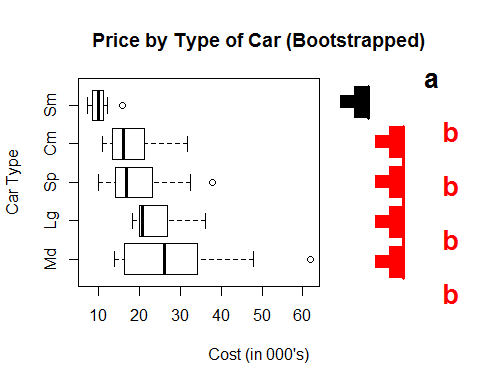
## diff lwr upr t p p adj rej H\_0  
## Lg-Cm 6.087500 -1.8008935 13.975893 2.3977076 0.0281 0.2170 0  
## Md-Cm 9.005682 -0.6372461 18.648610 2.9017065 0.0070 0.1235 0  
## Sm-Cm -8.045833 -13.4061073 -2.685559 -4.6636919 0.0024 0.0261 1  
## Sp-Cm 1.180357 -7.2357647 9.596479 0.4357599 0.6638 0.9932 0  
## Md-Lg 2.918182 -7.1443992 12.980763 0.9010497 0.3900 0.8991 0  
## Sm-Lg -14.133333 -20.2162614 -8.050405 -7.2190114 0.0014 0.0016 1  
## Sp-Lg -4.907143 -13.8009938 3.986708 -1.7142915 0.0976 0.4732 0  
## Sm-Md -17.051515 -25.2832417 -8.819789 -6.4360269 0.0001 0.0048 1  
## Sp-Md -7.825325 -18.3067316 2.656082 -2.3196827 0.0272 0.2358 0  
## Sp-Sm 9.226190 2.4729417 15.979439 4.2447828 0.0108 0.0360 1

### Part 3b

Repeat 2b) using the results produced by bootstrapped Games-Howell. Again use .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

padj\_extract3 <- function(formula, data){bootpriceType[,'p adj']}  
bootpriceType[,'p adj'] <- bootpriceType[,'p adj']/2  
multcompBoxplot(Price~Type, data = Cars93, horizontal = T, compFn="padj\_extract3")  
title(main = "Price by Type of Car (Bootstrapped)", ylab = "Car Type", xlab = "Cost (in 000's)")



### Part 3c

Explain these results in context as you did in 2c).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3c -|-|-|-|-|-|-|-|-|-|-|-

We are 90% confident that the population mean price for small cars is less than all other car types - 90% confidence intervals are below (in thousands of dollars). Sm vs. Cm [-13.3, -2.8] Sm vs. Sp [-15.8, -2.7] Sm vs. Lg [-20.0, -8.2] Sm vs. Md [-25.1, -9.0] ---

## Exercise 4.

One step procedures like Tukey-Kramer and Games-Howell are conservative (lower power) so they can miss significant differences between population means. If you don't need the confidence intervals, then a multi-step procedure such as the Bonferroni-Holm step-down procudere may be used to get more power.

### Part 4a

Repeat 2a, but this time use the Bonferroni-Holm procedure at to compare the population mean prices for the different types of cars. Use bootstrapping and unequal variances.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4a -|-|-|-|-|-|-|-|-|-|-|-

bonfComp <- onewayComp(Price~Type, data = Cars93, adjust = 'holm', nboot = 10000, alpha = 0.1, var.equal = F)$comp  
bonfComp

## diff lwr upr t p p adj rej H\_0  
## Lg-Cm 6.087500 NA NA 2.3977076 0.0245 0.1225 0  
## Md-Cm 9.005682 NA NA 2.9017065 0.0083 0.0581 1  
## Sm-Cm -8.045833 NA NA -4.6636919 0.0020 0.0180 1  
## Sp-Cm 1.180357 NA NA 0.4357599 0.6621 0.7592 0  
## Md-Lg 2.918182 NA NA 0.9010497 0.3796 0.7592 0  
## Sm-Lg -14.133333 NA NA -7.2190114 0.0023 0.0184 1  
## Sp-Lg -4.907143 NA NA -1.7142915 0.1021 0.3063 0  
## Sm-Md -17.051515 NA NA -6.4360269 0.0000 0.0000 1  
## Sp-Md -7.825325 NA NA -2.3196827 0.0254 0.1225 0  
## Sp-Sm 9.226190 NA NA 4.2447828 0.0110 0.0660 1

Ps <- bonfComp[,'p']  
Ps

## Lg-Cm Md-Cm Sm-Cm Sp-Cm Md-Lg Sm-Lg Sp-Lg Sm-Md Sp-Md Sp-Sm   
## 0.0245 0.0083 0.0020 0.6621 0.3796 0.0023 0.1021 0.0000 0.0254 0.0110

p.adjust(p = Ps, method = 'holm') #Appears stepdown p values are different from the p adj in the bonfComp variable (even though results turned out the same).

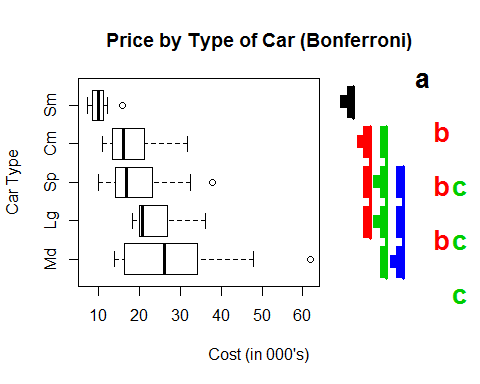
## Lg-Cm Md-Cm Sm-Cm Sp-Cm Md-Lg Sm-Lg Sp-Lg Sm-Md Sp-Md Sp-Sm   
## 0.1225 0.0581 0.0180 0.7592 0.7592 0.0184 0.3063 0.0000 0.1225 0.0660

### Part 4b

Repeat 2b to produce the boxplot with T and letter displays for the output in 4a. Don't forget to manually adjust the P-values to "fool" the plot into use the significance level to produce the plot.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4b -|-|-|-|-|-|-|-|-|-|-|-

padj\_extract4 <- function(formula, data){bonfComp[,'p adj']}  
bonfComp[,'p adj'] <- bonfComp[,'p adj']/2  
multcompBoxplot(Price~Type, data = Cars93, horizontal = T, compFn="padj\_extract4")  
title(main = "Price by Type of Car (Bonferroni)", ylab = "Car Type", xlab = "Cost (in 000's)")



### Part 4c

As in 2c, explain the mean price comparisons in context. Did you find any mean price differences that weren't previously revealed?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4c -|-|-|-|-|-|-|-|-|-|-|-

We are 90% confident that the population mean price for small cars is different from all other car types - 90% confidence intervals are below (in thousands of dollars). Sm vs. Cm [-12.7, -3.3] Sm vs. Sp [-15.0, -3.1] Sm vs. Lg [-19.0, -8.5] Sm vs. Md [-25.1, -10.7]

## In addition, we are 90% confident that the population mean price for compact cars is between 18.2 and 1.4 (-18.2, -1.4) LOWER than the price of Md (midsize? medium?) cars.

### Part 4d

In Exercises 2(games-howell), 3(bootstrapping), and 4(bonferroni stepdown) you used 3 different methods to analyze the differences in population mean prices for the different types of cars. Which analysis do you think is the most reliable? Why?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4d -|-|-|-|-|-|-|-|-|-|-|-

Because the data is assumed to have unequal variance and non-normal, the Games-Howell procedures is not appropriate. While the Bonferroni-Holm stepdown procedure is effective for hypothesis testing, we are more interested in establishing confidence intervals for PRICES between car types. In this case, I would rely on bootstrapping for the intervals and the Bonferroni-Holm procedure to help verify findings. The Bonferroni stepdown procedure found an additional difference in means between Cm and Md vehicles - but given the fact that Sp vehicles can't be separated from any of the other more expensive vehicles (exclusing Sm cars), I am more confident in claiming only the differences between Sm cars and all other types (via bootstrapping procedure with unequal variance).

## Exercise 5

Build a custom contrast matrix that compares the average small and compact prices to the average of the other car types and also compares the mean prices of the midsize and compact cars. (You may have to use levels(Type) to see the ordering of the levels) Use the Bonferroni-Holm procedure at the 10% significance level with bootstrapping and unequal variances to make the comparisons. Summarize your results.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 5 -|-|-|-|-|-|-|-|-|-|-|-

#Unsure if the question is asking to compare small and compact cars individually or together when comparing to others. I assumed this effort was to combine small and compact cars into a singular group.  
levels(Type)

## [1] "Cm" "Lg" "Md" "Sm" "Sp"

CarContrast <- rbind('aveSmCm - aveSpLgMd'=c(-1/2, 1/3, 1/3, -1/2, 1/3), 'Md - Cm'=c(-1, 0, 1, 0, 0))  
onewayComp(Price~Type, var.equal = F, nboot = 10000, con=CarContrast, adjust = 'holm')$comp[,c(1,4,6,7)]

## diff t p adj rej H\_0  
## aveSmCm - aveSpLgMd 9.447430 6.080276 0.0000 1  
## Md - Cm 9.005682 2.901707 0.0076 1

#There are significant differences in both analyses. Small and Compact cars average mean price is different from the mean price of Sporty, Large and Midsize cars (p=0). In addition, the mean price of compact cars is different from the mean price of midsize cars (p = .0078 adjusted). Both tests use bootstrapping with the Bonferroni-Holm procedure.

## Exercise 6

Since the price distributions are skewed it makes more sense to talk about median prices than mean prices.

### Part 6a

The Kruskal-Wallis and Dunn procedures aren't appropritate for comparing population median prices, why? Explain.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 6a -|-|-|-|-|-|-|-|-|-|-|-

The Kruskal-Wallis test and the Dunn test are innappropriate for comparing our car price population medians because the tests require equal variances.

### Part 6b

We're going to make 4 simulataneous confidence intervals for price data (compact - small, sporty-small, midsize-sporty, midsize-compact). If we want familywise confidence level 90%, what confidence level should you use for each individual comparison according to the Bonferroni correction?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 6b -|-|-|-|-|-|-|-|-|-|-|-

We want to take our desired alpha value of 0.1 and divide it by the number of comparisons we plan to make - 4. Therefore we will compare each individual p value to 0.025 (on the flipside, the individual confidence level needs to be 0.975).

### Part 6c

Use the boot package (as in the class presentation) to bootstrap the 4 confidence intervals for the specified differences of population median prices. You'll have to write the helper function and make sure you are referring to the correct columns of the Cars93 data. Don't forget to install and load the 'boot' package.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 6c -|-|-|-|-|-|-|-|-|-|-|-

require(boot)  
bootMedian <- function(x, i){  
 median(x[i])  
}  
  
bootMedianPrint <- function(x, i){  
 print(i)  
 print(x[i])  
 median(x[i])  
}  
  
#Compact, Sporty, Small, Midsize  
Compact <- Price[Type=='Cm']; Compact

## [1] 29.1 13.4 11.4 15.8 13.3 11.3 17.5 16.5 31.9 15.7 13.5 11.1 28.7 19.5  
## [15] 20.0 22.7

Sporty <- Price[Type=='Sp']; Sporty

## [1] 15.1 38.0 25.8 15.9 14.0 12.5 19.8 10.0 32.5 14.1 14.4 17.7 18.4 23.3

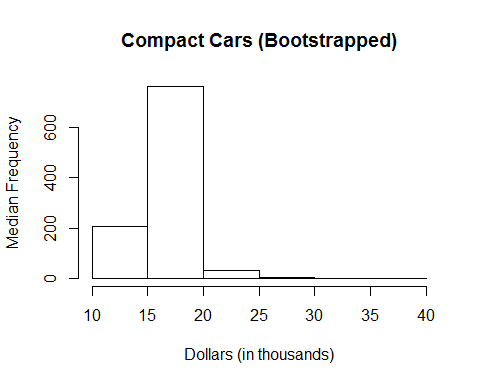
Small <- Price[Type=='Sm']; Small

## [1] 15.9 9.2 11.3 12.2 7.4 10.1 8.4 12.1 8.0 10.0 8.3 11.6 10.3 11.8  
## [15] 9.0 11.1 8.4 10.9 8.6 9.8 9.1

Midsize <- Price[Type=='Md']; Midsize

## [1] 33.9 37.7 30.0 15.7 26.3 40.1 15.9 15.6 20.2 13.9 47.9 28.0 35.2 34.3  
## [15] 61.9 14.9 26.1 21.5 16.3 18.5 18.2 26.7

#Bootstrapping the median for compact cars.  
set.seed(NULL) #set seed to first value   
boot.object.Compact <- boot(Compact, bootMedian, R=1000)  
binsCompact <- seq(10,40,by=5)  
hist(boot.object.Compact$t, breaks = binsCompact, main = "Compact Cars (Bootstrapped)", ylab = "Median Frequency", xlab = "Dollars (in thousands)")



#Build CI for compact cars  
CompactCI <- boot.ci(boot.object.Compact, conf = 0.9)

## Warning in boot.ci(boot.object.Compact, conf = 0.9): bootstrap variances  
## needed for studentized intervals

CompactCI

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 1000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = boot.object.Compact, conf = 0.9)  
##   
## Intervals :   
## Level Normal Basic   
## 90% (12.50, 19.01 ) (12.30, 18.85 )   
##   
## Level Percentile BCa   
## 90% (13.45, 20.00 ) (13.40, 19.50 )   
## Calculations and Intervals on Original Scale

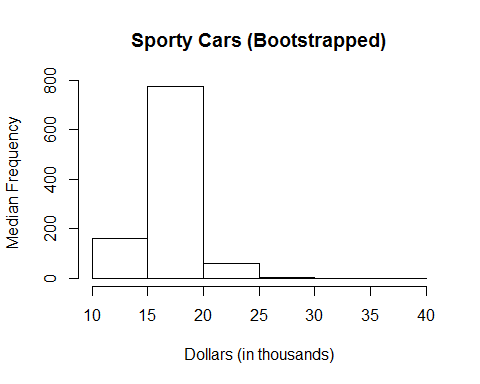
CompactCI$normal

## conf   
## [1,] 0.9 12.49956 19.01024

#Bootstrapping the median for sporty cars.  
set.seed(NULL) #set seed to first value   
Sporty

## [1] 15.1 38.0 25.8 15.9 14.0 12.5 19.8 10.0 32.5 14.1 14.4 17.7 18.4 23.3

boot.object.Sporty <- boot(Sporty, bootMedian, R=1000)  
binsSporty <- seq(10,40,by=5)  
hist(boot.object.Sporty$t, breaks = binsSporty, main = "Sporty Cars (Bootstrapped)", ylab = "Median Frequency", xlab = "Dollars (in thousands)")



#Build CI for compact cars  
SportyCI <- boot.ci(boot.object.Sporty, conf = 0.9)

## Warning in boot.ci(boot.object.Sporty, conf = 0.9): bootstrap variances  
## needed for studentized intervals

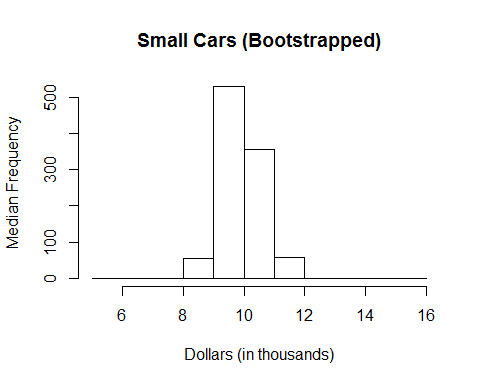
SportyCI$normal

## conf   
## [1,] 0.9 13.09773 20.07287

#Bootstrapping the median for small cars.  
set.seed(NULL) #set seed to first value   
Small #9.0 - 15.9

## [1] 15.9 9.2 11.3 12.2 7.4 10.1 8.4 12.1 8.0 10.0 8.3 11.6 10.3 11.8  
## [15] 9.0 11.1 8.4 10.9 8.6 9.8 9.1

boot.object.Small <- boot(Small, bootMedian, R=1000)  
binsSmall <- seq(5,16,by=1)  
hist(boot.object.Small$t, breaks = binsSmall, main = "Small Cars (Bootstrapped)", ylab = "Median Frequency", xlab = "Dollars (in thousands)")



#Build CI for compact cars  
SmallCI <- boot.ci(boot.object.Small, conf = 0.9)

## Warning in boot.ci(boot.object.Small, conf = 0.9): bootstrap variances  
## needed for studentized intervals

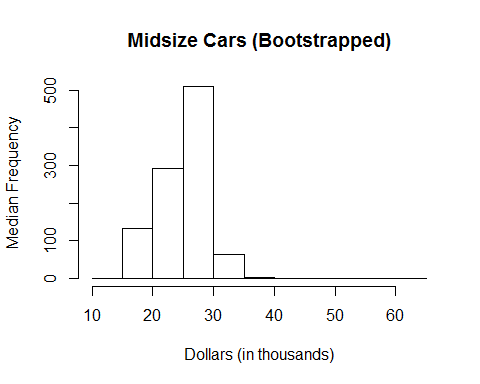
SmallCI$normal

## conf   
## [1,] 0.9 9.040715 11.08908

#Bootstrapping the median for sporty cars.  
set.seed(NULL) #set seed to first value   
Midsize #13.9 - 62

## [1] 33.9 37.7 30.0 15.7 26.3 40.1 15.9 15.6 20.2 13.9 47.9 28.0 35.2 34.3  
## [15] 61.9 14.9 26.1 21.5 16.3 18.5 18.2 26.7

boot.object.Midsize <- boot(Midsize, bootMedian, R=1000)  
binsMidsize <- seq(10,65,by=5)  
hist(boot.object.Midsize$t, breaks = binsMidsize, main = "Midsize Cars (Bootstrapped)", ylab = "Median Frequency", xlab = "Dollars (in thousands)")



#Build CI for compact cars  
MidsizeCI <- boot.ci(boot.object.Midsize, conf = 0.9)

## Warning in boot.ci(boot.object.Midsize, conf = 0.9): bootstrap variances  
## needed for studentized intervals

MidsizeCI$normal

## conf   
## [1,] 0.9 21.3733 33.5737

CompactCI$normal

## conf   
## [1,] 0.9 12.49956 19.01024

SportyCI$normal

## conf   
## [1,] 0.9 13.09773 20.07287

SmallCI$normal

## conf   
## [1,] 0.9 9.040715 11.08908

MidsizeCI$normal

## conf   
## [1,] 0.9 21.3733 33.5737

#median(boot.object.Compact$t)  
#median(boot.object.Sporty$t)  
#median(boot.object.Small$t)  
#median(boot.object.Midsize$t)  
  
  
  
bootMedianDiff <- function(d,i){  
 medians <- tapply(d[i,1], d[,2], median)  
 c(medians[1]-medians[3], medians[2]-medians[3], medians[4]-medians[2], medians[4]-medians[1])  
}  
  
x <- c(boot.object.Compact$t, boot.object.Sporty$t, boot.object.Small$t, boot.object.Midsize$t)  
group <- factor(rep(c('Cm', 'Sp', 'Sm', 'Md'), c(1000, 1000, 1000, 1000)))  
compareFrame <- data.frame(x, group)  
compareFrame

#Set up the contrast matrix to perform individual comparisons. (compact - small, sporty-small, midsize-sporty, midsize-compact)  
levels(compareFrame$group) #Cm, Md, Sm, Sp

## [1] "Cm" "Md" "Sm" "Sp"

contrastMatrix = rbind('Cm-Sm'=c(1, 0, -1, 0),  
 'Sp-Sm'=c(0, 0, -1, 1),  
 'Md-Sp'=c(0, 1, 0, -1),  
 'Md-Cm'=c(-1, 1, 0, 0))  
  
onewayComp(compareFrame$x~compareFrame$group, con = contrastMatrix, adjust = 'bonferroni', alpha=.1)$comp

## diff lwr upr t p p adj rej H\_0  
## Cm-Sm 6.6100 6.371887 6.848113 62.24454 0 0 1  
## Sp-Sm 7.0796 6.841487 7.317713 66.66663 0 0 1  
## Md-Sp 7.9118 7.673687 8.149913 74.50323 0 0 1  
## Md-Cm 8.3814 8.143287 8.619513 78.92532 0 0 1

### Part 6d

Explain the results of your intervals.

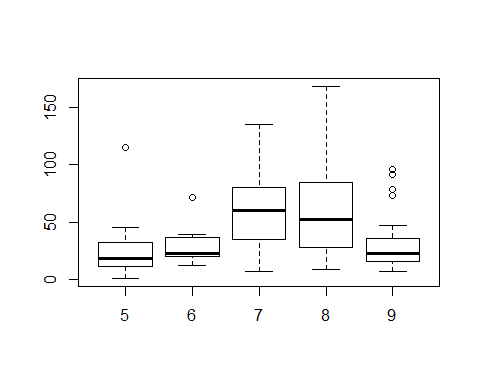
### -|-|-|-|-|-|-|-|-|-|-|- Answer 6d -|-|-|-|-|-|-|-|-|-|-|-

We are 90% confident that the population median price for small cars is 6.24 to 6.70 thousand less than the median price for compact cars. We are 90% confident that the population median price for small cars is 6.92 to 7.39 thousand less than the median price for sporty cars. We are 90% confident that the population median price for sporty cars is 7.55 to 8.02 thousand less than the median price for midsize cars. We are 90% confident that the population median price for compact cars is 8.24 to 8.70 thousand less than the median price for midsize cars.

## Exercise 7

The airquality data set that is built into R looks at air quaility measures in New York City, including ozone levels, for 5 months in 1973. We are going to estimate differences in population median ozone levels for the 5 months using the Dunn procedure which is a traditional follow up to the Kruskal-Wallis test. Here is a boxplot and the Kruskal-Wallis test:

detach(Cars93) # we don't need the Cars93 data now   
data(airquality)  
boxplot(Ozone~Month,data=airquality)



kruskal.test(Ozone ~ Month, data = airquality)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Ozone by Month  
## Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 6.901e-06

The boxplot shows that the distributions of ozone levels are similar, if not identical, for the 5 months. The Kruskal-Wallis test, assuming that the population distributions have the same shapes, shows that there is significant evidence that the population median ozone levels are not the same for all 5 months. Use the Dunn procedure (as shown in the presentations) with Bonferroni-Holm adjusted p-values to see which months have different population median ozone levels. Use Summarize your findings. (Don't forget to install and load the correct package here.)

### -|-|-|-|-|-|-|-|-|-|-|- Answer 7 -|-|-|-|-|-|-|-|-|-|-|-

#install.packages('dunn.test')  
require(dunn.test)

## Loading required package: dunn.test

## Warning: package 'dunn.test' was built under R version 3.1.3

dunn.test(airquality$Ozone, airquality$Month, method='holm', alpha=0.05)

## Kruskal-Wallis rank sum test  
##   
## data: x and group  
## Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 0  
##   
##   
## Comparison of x by group   
## (Holm)   
## Col Mean-|  
## Row Mean | 5 6 7 8  
## ---------+--------------------------------------------  
## 6 | -0.925158  
## | 0.5323  
## |  
## 7 | -4.419470 -2.244208  
## | 0.0000\* 0.0745  
## |  
## 8 | -4.132813 -2.038635 0.286657  
## | 0.0002\* 0.1037 0.7744  
## |  
## 9 | -1.321202 0.002538 3.217199 2.922827  
## | 0.3729 0.4990 0.0052\* 0.0121\*

## There was a significant difference in median ozone levels between several months in 1973. Months 5 and 7 (p=0), months 5 and 8 (p=.0002), months 7 and 9 (p=.0052), and months 8 and 9 (p=.0121) all showed statistically significant differences in ozone using the Dunn procedure.