Lab Assignment: Multiple Comparisons/Post Hoc Procedures

Spencer Swartz

March 7, 2017

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'))

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 -|-|-|-|-|-|-|-|-|-|-|-

source('onewayComp.R')  
onewayComp(RPM~Type)

## $call  
## onewayComp(formula = RPM ~ Type)  
##   
## $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 -|-|-|-|-|-|-|-|-|-|-|-

require(multcompView)

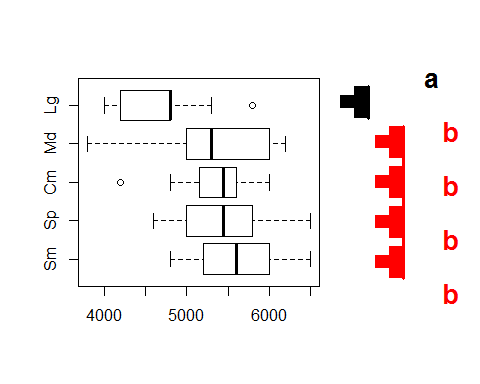
## Loading required package: multcompView

## Warning: package 'multcompView' was built under R version 3.3.2

set.seed(0)  
  
posthoc <- onewayComp(RPM~Type,adjust = 'one.step')  
posthoc$comp[,c(1,6,7)]

## diff p adj rej H\_0  
## Lg-Cm -689.77273 1.182831e-02 1  
## Md-Cm -26.13636 9.998848e-01 0  
## Sm-Cm 270.83333 5.429018e-01 0  
## Sp-Cm 30.35714 9.998639e-01 0  
## Md-Lg 663.63636 9.651980e-03 1  
## Sm-Lg 960.60606 5.744492e-05 1  
## Sp-Lg 720.12987 1.020512e-02 1  
## Sm-Md 296.96970 3.632245e-01 0  
## Sp-Md 56.49351 9.979305e-01 0  
## Sp-Sm -240.47619 6.851085e-01 0

padj\_extract <- function(formula,data){posthoc$comp[,'p adj']}  
multcompBoxplot(RPM~Type,data = Cars93 ,horizontal = TRUE, compFn = "padj\_extract")



### 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% that the means of the car groups are not the same. It would seem that the Large Vehicle mean is the likly the culperet as is difference ranges between 960 and 663 between the other groups means.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

source('regwqComp.R')  
newposthoc <- regwqComp(RPM~Type)  
newposthoc[,c(1,4,5,6)]

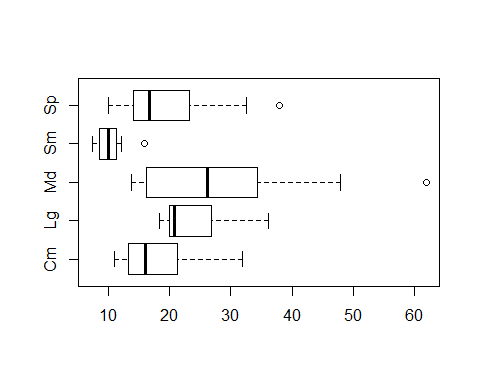
## diff p adj alpha adj rej H\_0  
## Lg-Cm -689.77273 3.942242e-03 0.03030722 1  
## Md-Cm -26.13636 8.814214e-01 0.02030827 0  
## Sm-Cm 270.83333 2.800479e-01 0.03030722 0  
## Sp-Cm 30.35714 8.763873e-01 0.02030827 0  
## Md-Lg 663.63636 1.125310e-03 0.02030827 1  
## Sm-Lg 960.60606 5.744492e-05 0.05000000 1  
## Sp-Lg 720.12987 6.427215e-03 0.05000000 1  
## Sm-Md 296.96970 2.665470e-01 0.05000000 0  
## Sp-Md 56.49351 9.481589e-01 0.03030722 0  
## Sp-Sm -240.47619 1.935909e-01 0.02030827 0

The procedure produced the same result as the one above which conferms that the Large Vehicles mean RPM is different from the rest of the samples.

## 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 ).

posthoc <- TukeyHSD(aov(Price~Type, data = Cars93))$Type  
posthoc

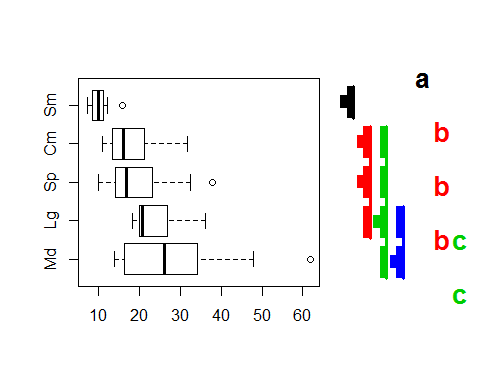
## diff lwr upr p adj  
## Lg-Cm 6.087500 -2.727795 14.9027946 3.113634e-01  
## Md-Cm 9.005682 1.610791 16.4005724 9.112747e-03  
## Sm-Cm -8.045833 -15.514490 -0.5771765 2.829501e-02  
## Sp-Cm 1.180357 -7.056241 9.4169558 9.944945e-01  
## Md-Lg 2.918182 -5.392958 11.2293213 8.633800e-01  
## Sm-Lg -14.133333 -22.510175 -5.7564921 1.000098e-04  
## Sp-Lg -4.907143 -13.975339 4.1610538 5.586891e-01  
## Sm-Md -17.051515 -23.917843 -10.1851875 9.943117e-09  
## Sp-Md -7.825325 -15.519945 -0.1307047 4.427241e-02  
## Sp-Sm 9.226190 1.460651 16.9917301 1.169566e-02

### 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\_extract <- function(formula,data){(posthoc[,'p adj']/2)}  
multcompBoxplot(Price~Type,data = Cars93 ,horizontal = TRUE, compFn = "padj\_extract")



### 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 -|-|-|-|-|-|-|-|-|-|-|-

At Alpha = 0.1 we will reject the the means of the pairs Md-Cm (the price mean is between 1.6 and 16.4 more for Md than Cm),Sm-Cm(the price mean is between 15.5 and 0.6 less for Sm than Cm),Sm-Lg(the price mean is between 22.5 and 5.8 less for Sm than Lg),Sm-Md(the price mean is between 23.9 and 10.2 less for Sm than Md),Sp-Md(the price mean is between 15.5 and 0.13 less for Sp than Md), and Sp-Sm( the price mean is between 1.4 and 17.0 more for Sp than Sm) are the same.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

posthoc <- onewayComp(Price~Type,adjust = 'one.step', nboot = 10000)  
posthoc$comp[,'p adj'] <- posthoc$comp[,'p adj']/2  
posthoc$comp

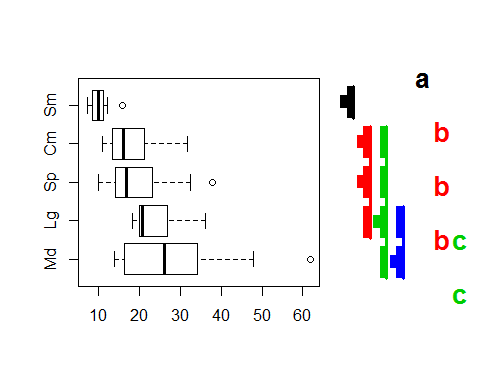
## diff lwr upr t p p adj rej H\_0  
## Lg-Cm 6.087500 -2.950765 15.1257650 1.9278958 0.0202 0.13125 0  
## Md-Cm 9.005682 1.423748 16.5876156 3.3999009 0.0064 0.00860 1  
## Sm-Cm -8.045833 -15.703399 -0.3882674 -3.0075298 0.0000 0.01895 1  
## Sp-Cm 1.180357 -7.264575 9.6252888 0.4000801 0.6740 0.49515 0  
## Md-Lg 2.918182 -5.603176 11.4395398 0.9802417 0.3694 0.40150 0  
## Sm-Lg -14.133333 -22.722055 -5.5446118 -4.7102692 0.0000 0.00065 1  
## Sp-Lg -4.907143 -14.204707 4.3904209 -1.5107381 0.0916 0.23815 0  
## Sm-Md -17.051515 -24.091517 -10.0115135 -6.9329786 0.0000 0.00000 1  
## Sp-Md -7.825325 -15.714569 0.0639198 -2.8392040 0.0252 0.02650 0  
## Sp-Sm 9.226190 1.264233 17.1881484 3.3168985 0.0000 0.01030 1

### Part 3b

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

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

padj\_extract <- function(formula,data){(posthoc$comp[,'p adj']/2)}  
multcompBoxplot(Price~Type,data = Cars93 ,horizontal = TRUE, compFn = "padj\_extract")



### Part 3c

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

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

At Alpha = 0.1 we will reject the the means of the pairs Md-Cm (the price mean is between 1.3 and 16.7 more for Md than Cm),Sm-Cm(the price mean is between 15.7 and 0.3 less for Sm than Cm),Sm-Lg(the price mean is between 22.8 and 5.4 less for Sm than Lg),Sm-Md(the price mean is between 24.2 and 10.0 less for Sm than Md),and Sp-Sm( the price mean is between 1.2 and 17.3 more for Sp than Sm) are the same.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

posthoc <- onewayComp(Price~Type,adjust = 'holm')  
posthoc$comp[,'p adj'] <- posthoc$comp[,'p adj']/2  
posthoc$comp

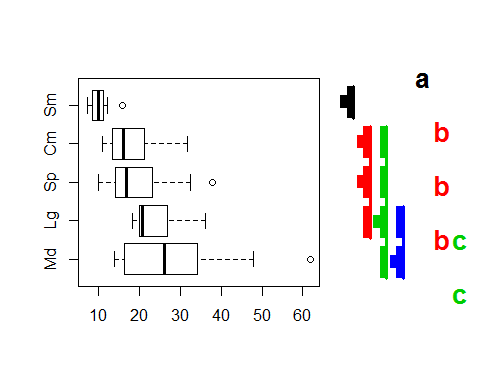
## diff lwr upr t p p adj rej H\_0  
## Lg-Cm 6.087500 NA NA 1.9278958 5.746173e-02 1.149235e-01 0  
## Md-Cm 9.005682 NA NA 3.3999009 1.059411e-03 4.237643e-03 1  
## Sm-Cm -8.045833 NA NA -3.0075298 3.531443e-03 1.059433e-02 1  
## Sp-Cm 1.180357 NA NA 0.4000801 6.901773e-01 3.450886e-01 0  
## Md-Lg 2.918182 NA NA 0.9802417 3.299592e-01 3.299592e-01 0  
## Sm-Lg -14.133333 NA NA -4.7102692 1.041816e-05 4.688173e-05 1  
## Sp-Lg -4.907143 NA NA -1.5107381 1.348450e-01 2.022674e-01 0  
## Sm-Md -17.051515 NA NA -6.9329786 1.003500e-09 5.017498e-09 1  
## Sp-Md -7.825325 NA NA -2.8392040 5.747628e-03 1.436907e-02 1  
## Sp-Sm 9.226190 NA NA 3.3168985 1.377332e-03 4.820663e-03 1

### 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\_extract <- function(formula,data){(posthoc$comp[,'p adj']/2)}  
multcompBoxplot(Price~Type,data = Cars93 ,horizontal = TRUE, compFn = "padj\_extract")



### 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 -|-|-|-|-|-|-|-|-|-|-|-

At Alpha = 0.1 we will reject that the means of the same pairs as in the previous test are the same, One new differece was identifed between the sports and mids (Sp,Md).

### Part 4d

In Exercises 2, 3, and 4 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 -|-|-|-|-|-|-|-|-|-|-|-

All three test retured very similar results making me fairly confident in them. It is known that the Bonferroni-Holm procedure has the highest power so as long as confidence intervals are not required I would go with the results of this test.

## Exercise 5

Build a custom contrast matrix that compares the the average of 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 -|-|-|-|-|-|-|-|-|-|-|-

levels(Type)

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

K = rbind('ave SmCm - ave SpLgMd'=c(1/2,-1/3,-1/3, 1/2,-1/3),  
 ' Md - Cm'=c(-1,0,1,0,0))  
posthoc <- onewayComp(Price~Type,adjust = 'holm', con=K,nboot = 10000, var.equal = FALSE)  
posthoc$comp[,'p adj'] <- posthoc$comp[,'p adj']/2  
posthoc$comp

## diff lwr upr t p p adj rej H\_0  
## ave SmCm - ave SpLgMd -9.447430 NA NA -6.080276 0.0000 0.00000 1  
## Md - Cm 9.005682 NA NA 2.901707 0.0077 0.00385 1

#padj\_extract <- function(formula,data){(posthoc$comp[,'p adj']/2)}  
#multcompBoxplot(Price~Type,data = Cars93 ,horizontal = TRUE, compFn = "padj\_extract")

At Alpha = 0.1 we will reject that the combined means of Sm and Cm are the same as the combined means of Sp,Lg, and Md. In addtion we will also reject that the means of Md are the same as the mean for Cm.

## 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 -|-|-|-|-|-|-|-|-|-|-|-

For these test when testing for the median the null hypothesis chages from test each group to the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different from the population median of at least one other group.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

4c2 = 6 .1/6 = .0166

We should use an individual comparison of 0.0166.

### 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')

## Loading required package: boot

## Warning: package 'boot' was built under R version 3.3.2

bootCars <- data.frame("Price" = Cars93$Price,   
 "Type" = Cars93$Type)  
#head(bootCars)  
  
bootMedDiff <- function(d,i){  
 # d is a dataframe with   
 # quantitative variable in column 1  
 # factor variable in column 2  
 meds <- tapply(d[i,1],d[,2],median)  
 c( meds[1]-meds[4], meds[5]-meds[4], meds[3]-meds[5], meds[3]-meds[1])  
 }  
  
set.seed(1)  
  
boot.object <- boot(bootCars, bootMedDiff, R = 5000,   
 strata = bootCars$Type)  
# Cm - Sm   
boot.ci(boot.object,conf = 1 - .1/6, type='bca', index=1)$bca[4:5]

## [1] 2.2 11.9

# Sp- Sm   
boot.ci(boot.object,conf = 1 - .1/6, type='bca', index=2)$bca[4:5]

## [1] 3.302009 14.450000

# Md - Sp  
boot.ci(boot.object,conf = 1 - .1/6, type='bca', index=3)$bca[4:5]

## [1] -0.45 19.65

# Md - Cm  
boot.ci(boot.object,conf = 1 - .1/6, type='bca', index=4)$bca[4:5]

## [1] 0.35000 20.65491

### Part 6d

Explain the results of your intervals.

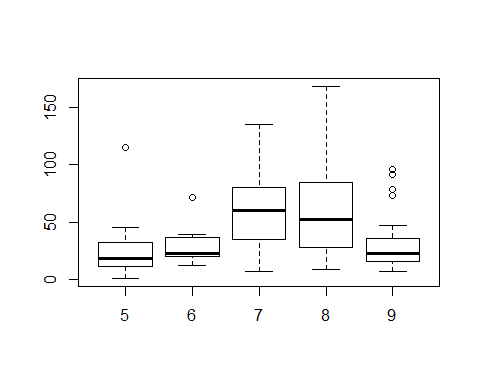
### -|-|-|-|-|-|-|-|-|-|-|- Answer 6d -|-|-|-|-|-|-|-|-|-|-|-

Pop. median Cm is 2.2 to 11.9 larger than pop. median Sm Pop. median Sp is 3.3 to 14.5 larger than pop. median Sm Pop. median Md is 0.5 less to 19.7 larger than pop. median Sp Pop. median Md is 0.4 to 20.7 larger than pop. median Cm

## 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 = 29.267, 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 -|-|-|-|-|-|-|-|-|-|-|-

require("dunn.test")

## Loading required package: dunn.test

## Warning: package 'dunn.test' was built under R version 3.3.2

attach(airquality)  
dunn.test(Ozone,Month, method = 'holm', alpha = .05)

## Kruskal-Wallis rank sum test  
##   
## data: Ozone and Month  
## Kruskal-Wallis chi-squared = 29.2666, df = 4, p-value = 0  
##   
##   
## Comparison of Ozone by Month   
## (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\*

According to this test month 5 and 6 have a statiscal different mean in ozone. in addtion so do the month pairs, 5 and 8, 7 and 9, and 8 and 9.