Lab Assignment: Multiple Comparisons/Post Hoc Procedures

Maggie Schweihs

Oct. 2017

## 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 simplify 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 -|-|-|-|-|-|-|-|-|-|-|-

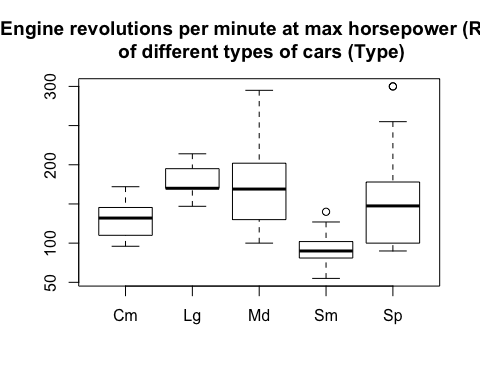
source('./onewayComp.R')  
source('./regwqComp.R')

summary(Cars93$Type)

## Cm Lg Md Sm Sp   
## 16 11 22 21 14

Sample Sizes are not balanced.

plot.new()  
boxplot(Horsepower~Type, main = "Engine revolutions per minute at max horsepower (RPM)\n of different types of cars (Type)")



From the boxplots, we can see some possible skewness, a few potential outliers. Each group could have a normal distribution, but the variances appear unequal.

rpm\_ghtest <- onewayComp(Horsepower~Type, data = Cars93,  
 #Games-Howell Test, var.equal = F  
 var.equal = FALSE,   
 adjust = "one.step")$comp[,c(1,2,3,6)]  
rpm\_ghtest

## diff lwr upr p adj  
## Lg-Cm 48.454545 22.651139 74.25795 3.390582e-06  
## Md-Cm 42.090909 5.698391 78.48343 1.053348e-02  
## Sm-Cm -40.000000 -61.212745 -18.78726 5.289306e-06  
## Sp-Cm 29.142857 -34.660692 92.94641 6.238263e-01  
## Md-Lg -6.363636 -44.001060 31.27379 9.880785e-01  
## Sm-Lg -88.454545 -112.538601 -64.37049 0.000000e+00  
## Sp-Lg -19.311688 -83.604445 44.98107 8.876327e-01  
## Sm-Md -82.090909 -117.373890 -46.80793 1.945301e-08  
## Sp-Md -12.948052 -80.864244 54.96814 9.793898e-01  
## Sp-Sm 69.142857 5.777603 132.50811 9.479101e-03

### Part 1b

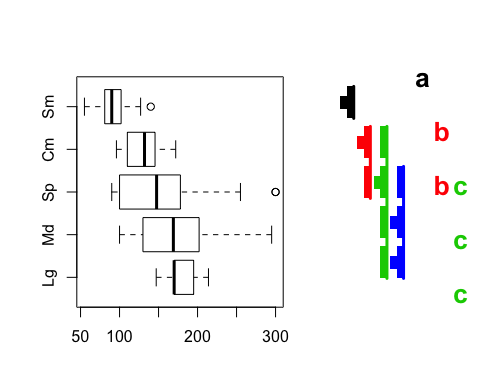
Use the multcompView package to produce a boxplot and letters/T display illustrating the differences in population means. (Review slides 17-19 in Multiple Comparisons Part 2 presentation.)

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

require(multcompView)

## Loading required package: multcompView

padj\_extract <- function(formula, data){  
 rpm\_ghtest[,'p adj']  
}  
par(mar=c(4,4,4,4))  
multcompBoxplot(Horsepower~Type, data = Cars93, horizontal = TRUE, compFn = "padj\_extract", sortFn = "mean", decreasing = TRUE)



* Cm = Compact
* Lg = Large
* Md = Midsize
* Sm = Small
* Sp = Sporty

### 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 of the following: The population mean RPM for small cars differs from the others. The mean RPM for compact and sporty cars are similar. The mean RPM for sporty cars cannot be separated from any type except small. The means for large, medium, and sporty are all similar to each other and different from the mean of the compact and sporty cars and different from the mean of the small cars.

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

posthoc <- regwqComp(Horsepower~Type, alpha = 0.05)  
posthoc[,c(1,4,5,6)]

## diff p adj alpha adj rej H\_0  
## Lg-Cm 48.454545 2.977841e-02 0.05000000 1  
## Md-Cm 42.090909 1.229987e-02 0.03030722 1  
## Sm-Cm -40.000000 7.304917e-03 0.02030827 1  
## Sp-Cm 29.142857 7.264597e-02 0.02030827 0  
## Md-Lg -6.363636 6.948547e-01 0.02030827 0  
## Sm-Lg -88.454545 5.918820e-06 0.05000000 1  
## Sp-Lg -19.311688 5.199588e-01 0.03030722 0  
## Sm-Md -82.090909 1.796958e-07 0.05000000 1  
## Sp-Md -12.948052 3.895119e-01 0.02030827 0  
## Sp-Sm 69.142857 5.076917e-05 0.03030722 1

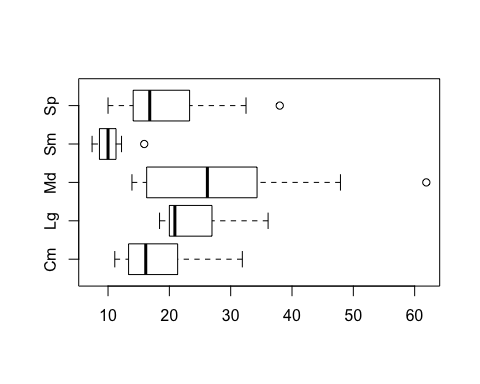
|  |  |  |
| --- | --- | --- |
|  | reject $\H\_0$ onewayComp | reject $\H\_0$ regwqComp |
| Lg-Cm | **yes** | **yes** |
| Md-Cm | **yes** | **yes** |
| Sm-Cm | **yes** | **yes** |
| Sp-Cm | no | no |
| Md-Lg | no | no |
| Sm-Lg | **yes** | **yes** |
| Sp-Lg | no | no |
| Sm-Md | **yes** | **yes** |
| Sp-Md | no | no |
| Sp-Sm | **yes** | **yes** |

The results are the same.

## 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 ).

price\_ghtest <- onewayComp(Price~Type, data = Cars93,  
 #Games-Howell Test, var.equal = F  
 var.equal = FALSE,   
 alpha = 0.10)$comp[,c(1,2,3,6)]  
price\_ghtest

## diff lwr upr p adj  
## Lg-Cm 6.087500 -0.5685001 12.7435001 1.266833e-01  
## Md-Cm 9.005682 1.0428168 16.9685468 3.759910e-02  
## Sm-Cm -8.045833 -12.6632617 -3.4284050 1.192465e-04  
## Sp-Cm 1.180357 -5.8653572 8.2260714 9.923626e-01  
## Md-Lg 2.918182 -5.4240153 11.2603789 8.956675e-01  
## Sm-Lg -14.133333 -19.6218454 -8.6448213 2.789471e-09  
## Sp-Lg -4.907143 -12.3994478 2.5851621 4.312457e-01  
## Sm-Md -17.051515 -24.0032322 -10.0997981 8.678284e-08  
## Sp-Md -7.825325 -16.4785915 0.8279421 1.495367e-01  
## Sp-Sm 9.226190 3.3048300 15.1475510 5.542557e-04

We are 90% confident, , that there is a difference in mean price between

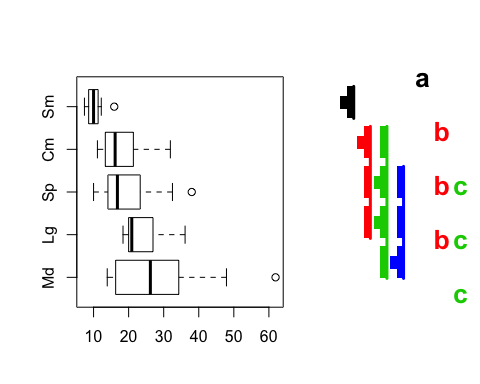
* Medium and Compact cars ()
* Small and Compact cars ()
* Small and Large cars ()
* Small and Medium cars ()
* Sports and Small cars ()

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

price\_ghtest[,'p adj'] <- price\_ghtest[,'p adj']/2  
padj\_extract <- function(formula, data){  
 price\_ghtest[,'p adj']  
}  
par(mar=c(4,4,4,4))  
multcompBoxplot(Price~Type, data = Cars93, horizontal = TRUE, compFn = "padj\_extract", sortFn = "mean", decreasing = TRUE)



### 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 there is a difference in mean price between

* Small and Compact cars ()
* Population mean cost of **Small** cars is between $3428 and $12,663 less than the population mean price of **Compact** cars
* Small and Large cars ()
* Population mean cost of **Small** cars is between $8644 and $19,621 less than the population mean price of **Large** cars
* Small and Medium cars ()
* Population mean cost of **Small** cars is between $10,099 and $24,003 less than the population mean price of **Medium** cars
* Sports and Small cars ()
* Population mean cost of **Small** cars is between $3304 and $15,147 less than the population mean price of **Sports** cars

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

price\_boot <- onewayComp(Price~Type, data = Cars93,  
 var.equal = FALSE,   
 alpha = 0.10,   
 adjust = "one.step",  
 nboot=10000)$comp[,c(1,2,3,5:7)]  
price\_boot

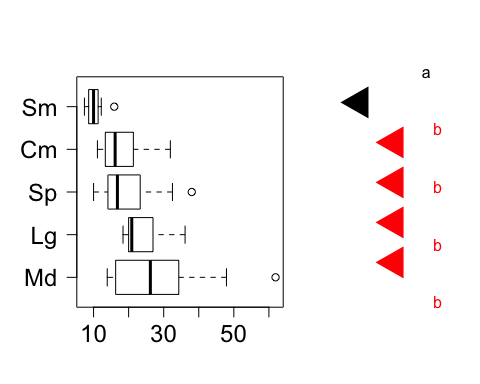
## diff lwr upr p p adj rej H\_0  
## Lg-Cm 6.087500 -1.6691350 13.844135 0.0271 0.2088 0  
## Md-Cm 9.005682 -0.4761819 18.487546 0.0083 0.1194 0  
## Sm-Cm -8.045833 -13.3165756 -2.775091 0.0031 0.0230 1  
## Sp-Cm 1.180357 -7.0951916 9.455906 0.6664 0.9916 0  
## Md-Lg 2.918182 -6.9763255 12.812689 0.3756 0.8947 0  
## Sm-Lg -14.133333 -20.1146593 -8.152007 0.0014 0.0021 1  
## Sp-Lg -4.907143 -13.6524412 3.838156 0.1011 0.4692 0  
## Sm-Md -17.051515 -25.1457485 -8.957282 0.0001 0.0048 1  
## Sp-Md -7.825325 -18.1316624 2.481013 0.0273 0.2305 0  
## Sp-Sm 9.226190 2.5857401 15.866641 0.0095 0.0337 1

### Part 3b

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

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

price\_boot[,'p adj'] <- price\_boot[,'p adj']/2  
padj\_extract <- function(formula, data){  
 price\_boot[,'p adj']  
}  
plot.new()  
par(mar=c(4,4,4,4))  
multcompBoxplot(Price~Type, data = Cars93, horizontal = TRUE, compFn = "padj\_extract", sortFn = "mean", decreasing = TRUE,   
 plotList=list(  
 boxplot=list(  
 fig=c(0, 0.75, 0, 1), las=1,  
 cex.axis=1.5),  
 multcompTs=list(  
 fig=c(0.7, 0.85, 0, 1),  
 type='boxes'),  
 multcompLetters=list(  
 fig=c(0.80, 1, 0, 1),  
 type='Letters') ) )



### Part 3c

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

After using the bootstrapping procedure at level of significance, we are now detecting different population means for groups Small cars (represented by group a), and all other types of cars (represented by group b).

With 90% confidence, we can construct the following intervals for the difference between the population mean of small cars and each other car type:

* Small and Compact cars ( 0.0115)
* Population mean cost of **Small** cars is between 13.317 and 2.775 $K less than the population mean price of **Compact** cars
* Small and Large cars ( 0.00105)
* Population mean cost of **Small** cars is between 20.115 and 8.152 $K less than the population mean price of **Large** cars
* Small and Medium cars ( 0.0024)
* Population mean cost of **Small** cars is between 25.146 and 8.957 $K less than the population mean price of **Medium** cars
* Sports and Small cars ( 0.01685)
* Population mean cost of **Small** cars is between 15.867 and 2.586 $K less than the population mean price of **Sports** cars

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

price\_bhtest <- onewayComp(Price~Type, data = Cars93,  
 #Games-Howell Test, var.equal = F  
 var.equal = FALSE,   
 #bootstrap  
 nboot = 10000,  
 #Bonferroni-Holm Correction  
 adjust = 'holm',  
 alpha = 0.10)$comp[,c(1,4,5:7)]  
price\_bhtest

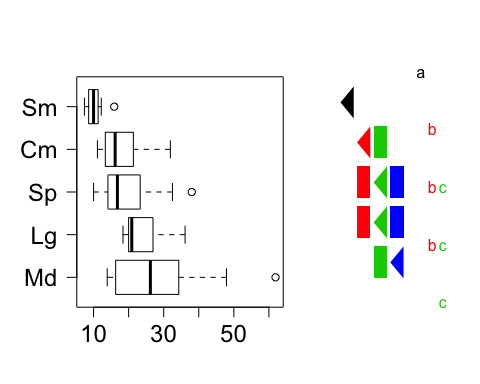
## diff t p p adj rej H\_0  
## Lg-Cm 6.087500 2.3977076 0.0287 0.1205 0  
## Md-Cm 9.005682 2.9017065 0.0065 0.0455 1  
## Sm-Cm -8.045833 -4.6636919 0.0031 0.0248 1  
## Sp-Cm 1.180357 0.4357599 0.6620 0.7528 0  
## Md-Lg 2.918182 0.9010497 0.3764 0.7528 0  
## Sm-Lg -14.133333 -7.2190114 0.0018 0.0162 1  
## Sp-Lg -4.907143 -1.7142915 0.1002 0.3006 0  
## Sm-Md -17.051515 -6.4360269 0.0001 0.0010 1  
## Sp-Md -7.825325 -2.3196827 0.0241 0.1205 0  
## Sp-Sm 9.226190 4.2447828 0.0093 0.0558 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 -|-|-|-|-|-|-|-|-|-|-|-

price\_bhtest[,'p adj'] <- price\_bhtest[,'p adj']/2  
padj\_extract <- function(formula, data){  
 price\_bhtest[,'p adj']  
}  
plot.new()  
par(mar=c(4,4,4,4))  
multcompBoxplot(Price~Type, data = Cars93, horizontal = TRUE, compFn = "padj\_extract", sortFn = "mean", decreasing = TRUE,   
 plotList=list(  
 boxplot=list(  
 fig=c(0, 0.75, 0, 1), las=1,  
 cex.axis=1.5),  
 multcompTs=list(  
 fig=c(0.7, 0.85, 0, 1),  
 type='boxes'),  
 multcompLetters=list(  
 fig=c(0.80, 1, 0, 1),  
 type='Letters') ) )



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

The Bonferroni-Holm boostraped procedure revealed more groupings than the Games-Howell bootstrapped procedure, yet it's results are nearly identical to the Games-Howell procedure without bootstrapping.

We are 90% confident, , that there is a difference in mean price between

* Small and Compact cars ( 0.0124)
* Population mean cost of **Small** cars is less than the population mean price of **Compact** cars
* Small and Large cars ( 0.0081)
* Population mean cost of **Small** cars is less than the population mean price of **Large** cars
* Small and Medium cars ( 510^{-4})
* Population mean cost of **Small** cars is less than the population mean price of **Medium** cars
* Sports and Small cars ( 0.0279)
* Population mean cost of **Small** cars is less than the population mean price of **Sports** cars

**The Bonferroni-Holm Procedure reveals an additional contrast**

* Medium and Compact cars ( 0.02275)
* Population mean cost of **Compact** cars is less than the population mean price of **Medium** cars

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

require(car)

## Loading required package: car

## Warning: package 'car' was built under R version 3.2.5

p\_value <- leveneTest(Price~Type)[1,3]

The Games Howell test is good for unequal variances. We can use Levene's Test to conclude ( 5.6310^{-4}) that the variances of our groups are unequal at 0.05 level of significance. Games-Howell also takes into acount unequal group sizes, which we have in this data set.

Another assumption that was not met was the assumption of normal distribution. The boxplots of the prices show skewed distributions. Bootstrapping helps us deal with the issue of small samples and non-normal distributions. [Bootstrapping](https://en.wikipedia.org/wiki/Bootstrapping_(statistics)) also reduces the effects of random sampling errors and allows us to make better inferences about our data given the relatively small sample size (n < 30) in each group. For these reasons the Game-Howell Bootstrapped method is an improvement over the Game-Howell method.

In the third test, we used a Bonferroni-Holm correction to our bootstrapped, Games-Howell method. This added the advantage of correcting for false discoveries in the multiple comparisons.

|  |  |  |  |
| --- | --- | --- | --- |
| reject null with method: | Game-Howell | Game-Howell Boot | Bonferroni-Holm |
| Lg-Cm | no | no | no |
| Md-Cm | no | no | **yes** |
| Sm-Cm | **yes** | **yes** | **yes** |
| Sp-Cm | no | no | no |
| Md-Lg | no | no | no |
| Sm-Lg | **yes** | **yes** | **yes** |
| Sp-Lg | no | no | no |
| Sm-Md | **yes** | **yes** | **yes** |
| Sp-Md | no | no | no |
| Sp-Sm | **yes** | **yes** | **yes** |

The results of the three tests are almost identical, however, the final method, the Bootstrapped Games-Howell method, is the least conservative and reveals one additional difference than the other two. For the reasons listed above, I believe the Bootstrapped Games-Howell with Bonferroni-Holm correction to be the most reliable.

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

K = rbind( 'aveSM/CM - aveOther' = c(1/2,-1/3,-1/3,1/2,-1/3),  
 'aveMd - aveCm' = c(-1,0,1,0,0)  
)  
price\_custom <- onewayComp( Price~Type, data = Cars93, con = K,  
 adjust = 'bonferroni')$comp[,c(1:4,6)]  
price\_custom

## diff lwr upr t p adj  
## aveSM/CM - aveOther -9.447430 -13.591752 -5.303108 -5.208735 2.951544e-06  
## aveMd - aveCm 9.005682 2.953355 15.058009 3.399901 2.118821e-03

At a 0.05 level of significance, we can say that there is a difference between the mean of average small and compact prices compared to the mean price of the other car types ( 2.9510^{-6}). We are 95% confident that the true mean prices of average Small and Compact cars is between $13,592 and $5303 lower than the mean prices of other cars.

We can also say, at a 0.05 level of significance, that there is a difference between the mean prices of the midsize and compact cars ( 0.00212). We are 95% confident that the true mean prices of average Midsize cars is between $2953 and $15,058 higher than the mean prices of compact cars.

## 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 assumes that groups have the same distribution/shape. "While Kruskal-Wallis does not assume that the data are normal, it does assume that the different groups have the same distribution, and groups with different standard deviations have different distributions." [McDonald](http://www.biostathandbook.com/kruskalwallis.html) From the boxplots, we can see that the group variances are not equal, So Kruskall-Wallis and Dun procedures are not appropriate in this case.

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

Bonferroni Inequality:

We want and , since we have 4 contrasts. So, Then, And .

Each individual comparison should have a confidence level of 97.5% to achieve a familywise confidence level of 90%.

### 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 3.2.5

##   
## Attaching package: 'boot'

## The following object is masked from 'package:car':  
##   
## logit

bootMedianDiff <- function(d, i){  
 medians <- tapply(d[i,1],d[,2], median)  
 c(medians[1]-medians[2], ##Group 1: compact - small  
 medians[3]-medians[2], ##Group 2: sporty - small  
 medians[4]-medians[3], ##Group 3: midsize - sporty  
 medians[4]-medians[1]) ##Group 4: midsize - compact  
}  
xSm <- Cars93$Price[Cars93$Type == "Sm"]  
xCm <- Cars93$Price[Cars93$Type == "Cm"]  
xSp <- Cars93$Price[Cars93$Type == "Sp"]  
xMd <- Cars93$Price[Cars93$Type == "Md"]  
  
gSm <- Cars93$Type[Cars93$Type == "Sm"]  
gCm <- Cars93$Type[Cars93$Type == "Cm"]  
gSp <- Cars93$Type[Cars93$Type == "Sp"]  
gMd <- Cars93$Type[Cars93$Type == "Md"]  
  
#Index 1: compact - small  
#Index 2: sporty - small  
#Index 3: midsize - sporty  
#Index 4: midsize - compact  
x <- (c(xCm,xSm,xSp,xMd))  
group <- factor(c(as.character(gCm),  
 c(as.character(gSm)),  
 c(as.character(gSp)),  
 c(as.character(gMd))))  
d <- data.frame(x,group)  
boot.object <- boot(d, bootMedianDiff, R=5000, strata = d$group)   
  
print("Confidence Interval for Diff in Median Price: compact - small")

## [1] "Confidence Interval for Diff in Median Price: compact - small"

boot.ci(boot.object,conf = .975, type='bca',index=1)$bca[4:5]

## [1] -20.4363544 -0.8927586

print("Confidence Interval for Diff in Median Price: sporty - small")

## [1] "Confidence Interval for Diff in Median Price: sporty - small"

boot.ci(boot.object,conf = .975, type='bca',index=2)$bca[4:5]

## [1] -24.74243 -7.90000

print("Confidence Interval for Diff in Median Price: midsize - sporty")

## [1] "Confidence Interval for Diff in Median Price: midsize - sporty"

boot.ci(boot.object,conf = .975, type='bca',index=3)$bca[4:5]

## [1] 3.382036 13.323269

print("Confidence Interval for Diff in Median Price: midsize - compact")

## [1] "Confidence Interval for Diff in Median Price: midsize - compact"

boot.ci(boot.object,conf = .975, type='bca',index=4)$bca[4:5]

## [1] -5.5 8.3

### Part 6d

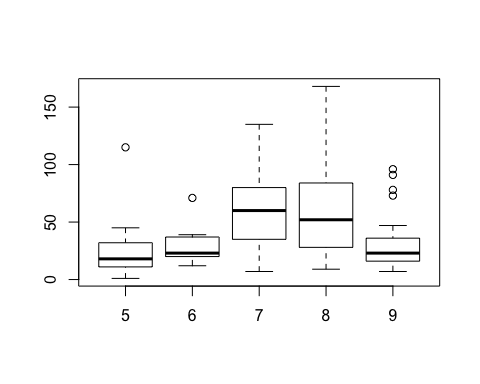
### -|-|-|-|-|-|-|-|-|-|-|- Answer 6d -|-|-|-|-|-|-|-|-|-|-|-

We can say, with 90% confidence, that the median price of small cars is between $20,104 and $850 less than the median price of compact cars; the median price of small cars is between $24,104 and $785 less than the median price of sporty cars; and the median price of sporty cars is between$3450 and $13,921 more than the median price of midsize cars. Since 0 is in the interval for the difference between the median price of midsize and compact cars, there may be no difference between the two.

## 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 (note, we are allowing considerble latitude here to say these distributions are identical). 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 3.2.5

dunn.test(airquality$Ozone,airquality$Month, method='holm',alpha=0.5)

## 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\*

Using the Dunn test to examine which months have different population median ozone levels, we can say, with a 0.05 level of significance, that there is a significant difference between the median ozone levels in May and July, June and July, May and August, June and August, July and September, and August and September (assuming 1= January, etc.).

The significance implies those distributions are shifted and have different medians from each other.