ANOVA etc.

Josh Jarvey

07/02/2020

Create a Word docx from this R Markdown file for the following exercise. Submit the R markdown file and resulting Word docx file.

## Exercise 1

In the Lesson 3 presentation you saw how to use the Wilcoxon Rank Sum test to compare the difference in median repair times for Macs and PCs. You’ll find the repair dataset in the DS705data package. In this problem we’ll test the hypothesis that the population mean repair times are different for Macs and PCs at the 5% significance level using three different approaches.

### Part 1a

Even though repair times for both computer types are skewed right go ahead and use t.test to test if the population mean times are significantly different. Include your R code below and write a conclusion to the test for practice.

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

#loading the data from the library  
library(DS705data)  
data("repair")  
 #t-test to determine if there is a difference between the population means.   
t.test(time~type, data = repair, alternative = "two.sided")

##   
## Welch Two Sample t-test  
##   
## data: time by type  
## t = -1.9442, df = 89.019, p-value = 0.05502  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -3.29705347 0.03583632  
## sample estimates:  
## mean in group Mac mean in group PC   
## 3.150659 4.781267

At a 5% significance level, there is not enough evidence to suggest that there is a true difference in mean repair times between PC and Mac computers at this company (p-value = 0.05502).

### Part 1b

Now use the boot package to construct a 95% BCa confidence interval for the difference in population mean repair times. Use at least 5000 resamples. Use that confidence interval to write a hypothesis test conclusion to this hypothesis test. (Review: you made similar bootstrap CI’s in Lesson 3.)

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

#import required packages  
library(boot)  
  
 #create an auxiliary function to calculate the mean of the re-sampled data.   
bootMean <- function(x,i){  
 mean(x[i])  
}  
  
 #complete the bootstrapping to create a sampling distribution of 5000 means from the re-sampled data. This is split by the factor "type" variable.   
boot.object = boot(repair$time, bootMean, R = 5000, strata = repair$type)  
  
 #create a 95% bca confidence interval using the bootstrapped data.  
boot.ci(boot.object, conf = 0.95, type = 'bca')

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
## Based on 5000 bootstrap replicates  
##   
## CALL :   
## boot.ci(boot.out = boot.object, conf = 0.95, type = "bca")  
##   
## Intervals :   
## Level BCa   
## 95% ( 3.236, 4.872 )   
## Calculations and Intervals on Original Scale

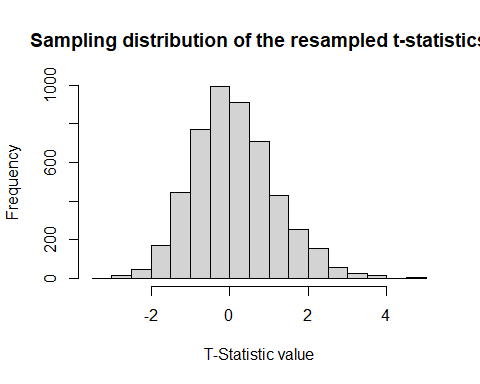
At a 5% significance level, there is enough evidence to suggest that there is a true difference in mean repair times between PC and Mac computers at this company.

### Part 1c

Follow along with with Two Means example in the Bootstrap Hypothesis Testing presentation to bootstrap the two means t test to see if there is a significant difference in population mean repair times. Include a histrogram of the boostrapped t-distribution and write a conclusion to the hypothesis test. (NOTE: in the P value computation slide the last part got cut off, the full code is P <- 2\*min( sum( bootdist < toriginal), sum( bootdist > toriginal ) )/5000.)

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

#setting seed for reproducability  
set.seed(123)  
 #determining the indices of the 2 populations within the original sorted dataframe.   
first\_mac\_index = 1  
last\_mac\_index = nrow(repair[repair$type == "Mac",])  
first\_pc\_index = last\_mac\_index+1  
last\_pc\_index = nrow(repair)  
  
 #re-centering the repair times by subtracting the population's sample mean from the population's sample repair times.  
xnull = c(repair$time[first\_mac\_index:last\_mac\_index] - mean(repair$time[first\_mac\_index:last\_mac\_index]),  
 repair$time[first\_pc\_index:last\_pc\_index] - mean(repair$time[first\_pc\_index:last\_pc\_index]))  
  
  
 #now i want to "re"sample with replacement from each of the population's sample independently.   
 #because we are not pooling the populations, we need to have two sample() functions, each pulling from their respective populations  
 #we want to replicate each of these sampling processes 5000 times.  
 #and finally, for each sampling iteration, we will row-bind the "re"samples together  
 #again, each population's "re"sample is constrained by these indices per each column.  
 #side note: creates a matrix of 122 original samples by 5000 iterations = 610,000 overall "re"samples. This is the pseudo-samples.  
 #each column in the matrix is the "re"sampled dataset.   
rs = rbind(replicate(5000, sample(xnull[first\_mac\_index:last\_mac\_index], replace = TRUE)),  
 replicate(5000, sample(xnull[first\_pc\_index:last\_pc\_index], replace = TRUE)))  
  
 #now we apply the t.test() function to the 5000 "re"sampled datasets from the "rs" matrix.  
 #each column of the rs matrix is a "re"sampled dataset, so "c" is a placeholder for that column (per the apply function(c))  
 #we still are comparing two populations, so split it by the repair$type when doing the t.test  
 #\*\*\*I dont understand how this works since rs isnt a dataframe and doesnt contain type??? How does it know?  
 #ultimately, we want the t-statistic from this t.test, so add the $statistic to pull that out.   
 #repeat this for each column that exists in the rs dataframe (i.e. 5000 times)  
bootdist = apply(rs,2,function(c) t.test(c~repair$type)$statistic)  
  
 #displaying a histogram of the resampled t-statistics.  
 #note: you can also get the t-statistics directly from a boot.object by accessing the "$t" attribute.  
hist(bootdist, main = "Sampling distribution of the resampled t-statistics", xlab = "T-Statistic value")



#calculating and storing the t-statistic from the original t.test  
 #this will be used as a baseline to compare against each bootstrapped t-statistic, which is used to calculate the p-value.  
toriginal = t.test(time~type, data = repair)$statistic  
  
  
 #calculating the pvalue based on the bootstrapped hypothesis test.   
 #Remember this is a two.sided test, so the next step is finding the minimum of the sums, and then multiplying by 2.   
 #first we need to find the sum of all bootstrapped t-statistics that are less than the baseline original t-statistic  
 #next we need to find the sum of all bootstrapped t-statistics that are greater than the baseline original t-statistic  
 #whichever is the minimum of these two values is what we'll use  
 #now we multiply that value by 2 (because its a two.sided test)  
 #then we divide that value by 5000 to get the final pvalue.  
P = 2\*min( sum(bootdist < toriginal),  
 sum(bootdist > toriginal)  
 )/5000  
P

## [1] 0.0292

At a 5% significance level, there is enough evidence to suggest that there is a true difference in mean repair times between PC and Mac computers (p-value = 0.0292)

### Part 1d

The bootstrap and theoretical t-distributions give different results here. Which do you trust? Why?

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

I trust the bootstrapped version a bit more since the original data was skewed. A condition that the t-test requires is the data must be normally distributed (or else it loses its power), so using the bootstrap technique provided a more accurate p-value, which ended up showing enough evidence that there is a difference in mean repair times between Mac and PC computers at this business.

## Exercise 2

This exercise is based on the data and experimental design from exercises 8.42 & 8.43 in the Ott textbook.

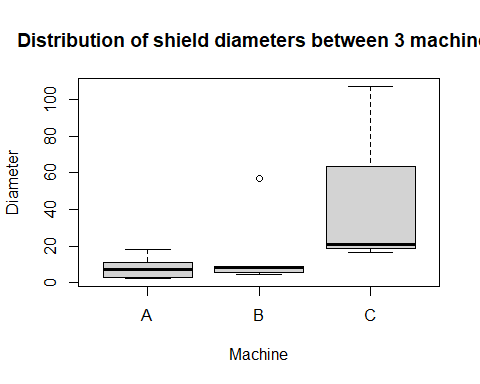
A small corporation makes insulation shields for electrical wires using three different types of machines. The corporation wants to evaluate the variation in the inside diameter dimension of the shields produced by the machines. A quality engineer at the corporation randomly selects shields produced by each of the machines and records the inside diameters of each shield (in millimeters). The goal is to determine whether the location parameters (i.e. mean or median) of the three machines differ. The data set shields is in the DS705data package. The R code to load it is already below.

### Part 2a

Construct side-by-side boxplots for the inside diameters of the insulation shields for the three machines.

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

require(DS705data)  
data(shields)  
 #create boxplots of the distribution of shield diameter from the 3 different machines  
boxplot(Diameter~Machine, data = shields, main = "Distribution of shield diameters between 3 machines")



### Part 2b

Comment on what you see in the boxplots. How do the medians compare visually? Do the samples look like they have roughly the same variability? Is there severe skewness or outliers in any of the samples? Be specific.

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

Machine A & B seem pretty similar overall (B does have quite a large outlier in it’s group), however machine C appears very different from A & B, with little or no overlap against A & B. The variability between all groups seems to be drastically different, and even the variability within the groups is different. Machine A doesn’t really have a bottom tail, and therefore appears to be slightly skewed to the right. Machine B doesnt really have a top tail, and therefore appears slightly skewed to the left. Machine C appears to have the highest variability, with relatively extreme skeweness to the right.

### Part 2c

Which data conditions for ANOVA appear not to be met, if any? Provide reasoning for your answer.

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

Data conditions for ANOVA are as follows:

1. The samples are randomly selected - the opening scenario states the engineer randomly selected these observations, so this is met.
2. The samples are selected independently - we dont know if the observations were selected independently, so this may be a violation.
3. The populations are approximately normally distributed - this assumption does appear to be violated as there is evidence of skewness
4. The population variances are equal - this also appears to be violated since the boxplot sizes are drastically different.

### Part 2d

Conduct an analysis of variance test (the standard one that assumes normality and equal variance). (i) State the null and alternative hypotheses, (ii) use R to compute the test statistic and p-value, and (iii) write a conclusion in context at .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2d -|-|-|-|-|-|-|-|-|-|-|-

mu1 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine A at this company. mu2 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine B at this company. mu3 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine C at this company.

H\_null: mu1 = mu2 = mu3 (all the population means are the same). H\_Alternative: At least one population mean is different than the others.

#first build a linear model regressing the diameter against machine.  
 #then conduct the analysis of variance on the data.  
anova(lm(Diameter~Machine,data = shields))

## Analysis of Variance Table  
##   
## Response: Diameter  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Machine 2 4141 2070.52 2.727 0.09393 .  
## Residuals 17 12907 759.26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#F-statistic = 2.727  
#P-value = 0.09393

At a 5% level of significance, there is not enough evidence to suggest that there is a true difference in population mean diameter (in millimeters) of insulation shields for electrical wires from either of the machines at this company (p-value = 0.09393).

### Part 2e

Conduct an analysis of variance test with the Welch correction. (i) State the null and alternative hypotheses, (ii) use R to compute the test statistic and p-value, and (iii) write a conclusion in context at .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e -|-|-|-|-|-|-|-|-|-|-|-

mu1 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine A at this company. mu2 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine B at this company. mu3 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine C at this company.

H\_null: mu1 = mu2 = mu3 (all the population means are the same). H\_Alternative: At least one population mean is different than the others.

#this is completing the Welch-Corrected ANOVA. Its the same thing as the anova() (you can use oneway.test()), but setting var.equal= FALSE makes it welch-corrected.   
oneway.test(Diameter~Machine, data = shields, var.equal = FALSE)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Diameter and Machine  
## F = 3.974, num df = 2.0000, denom df = 8.4087, p-value = 0.06096

#F\_prime-statistic = 3.971  
#P-value = 0.06096  
  
#the F\_prime-statistic is getting larger when using the welch corrected ANOVA and accounting for unequal variances, but still does not detect a difference in population means.

At a 5% level of significance, there is not enough evidence to suggest that there is a true difference in population mean diameter (in millimeters) of insulation shields for electrical wires from either of the machines at this company (p-value = 0.06096).

### Part 2f

Which data conditions for Welch ANOVA are not met, if any? Provide reasoning for your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2f -|-|-|-|-|-|-|-|-|-|-|-

While the Welch-corrected can do away with the “equal variances” condition of the standard ANOVA, they still both require that the sample be normally distributed, which we discovered in the box plots does not appear to be the case. This is why they both still appear to be failing.

### Part 2g

Conduct a Kruskal-Wallis test. (i) State the null and alternative hypotheses, (ii) use R to compute the test statistic and p-value, and (iii) write a conclusion in context using .

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2g -|-|-|-|-|-|-|-|-|-|-|-

mu1 = the population distribution of the diameter (in millimeters) of insulation shields for electrical wires from Machine A at this company. mu2 = the population distribution of the diameter (in millimeters) of insulation shields for electrical wires from Machine B at this company. mu3 = the population distribution of the diameter (in millimeters) of insulation shields for electrical wires from Machine C at this company.

H\_null: mu1 = mu2 = mu3 (all the population distributions are equal to each other, that is the median is located in the same position). H\_Alternative: At least one population distribution (the median’s location) is different than the others.

#perform the kruskal-wallis test to check if the medians of the populations are different from one another. Looks like this test is detecting a shift in medians from at least one of the populations.  
kruskal.test(Diameter~Machine, data = shields)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Diameter by Machine  
## Kruskal-Wallis chi-squared = 9.8914, df = 2, p-value = 0.007114

#Kruskal-Wallis X^2-Statistic = 9.8914  
#P-value = 0.007114

At a 5% level of significance, there is enough evidence to suggest that there is a true difference in population median diameter (in millimeters) of insulation shields for electrical wires from at least one of the machines at this company (p-value = 0.007114).

### Part 2h

Which data conditions for the Kruskal-Wallis test are not met, if any? Provide reasoning for your answer.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2h -|-|-|-|-|-|-|-|-|-|-|-

The data conditions for Kruskal-Wallis test are:

1. The samples are randomly selected - the opening scenario states the engineer randomly selected these observations, so this is met.
2. The samples are selected independently - we dont know if the observations were selected independently, so this may be a violation.
3. The populations have a similar shape (which also means similar variance) - this assumption does appear to be violated as the variation within groups seems to be drastically different.

### Part 2i

Conduct a bootstrapped ANOVA test using pooled residuals and unequal variances as in the notes. (i) State the null and alternative hypotheses, (ii) use R to compute the test statistic and p-value, and (iii) write a conclusion in context . Do not use a helper function, instead mimic the code in the notes using a for loop to construct the boostrapped sampling distribution.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2i -|-|-|-|-|-|-|-|-|-|-|-

mu1 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine A at this company. mu2 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine B at this company. mu3 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine C at this company.

H\_null: mu1 = mu2 = mu3 (all the population means are the same). H\_Alternative: At least one population mean is different than the others.

#set seed for reproducibility.  
set.seed(123)  
 #this is the F\_prime-statistic from the welch ANOVA (unequal variances - same value as in the above step) - 3.97  
F\_observed = oneway.test(Diameter~Machine, data = shields, var.equal = FALSE)$statistic  
  
 #now we center the data by subtracting the populations sample mean from each observation within that population's sample.   
res\_machine\_A = shields$Diameter[shields$Machine == "A"] - mean(shields$Diameter[shields$Machine == "A"])  
res\_machine\_B = shields$Diameter[shields$Machine == "B"] - mean(shields$Diameter[shields$Machine == "B"])  
res\_machine\_C = shields$Diameter[shields$Machine == "C"] - mean(shields$Diameter[shields$Machine == "C"])  
  
 #we build a new dataframe that is a copy of the old, but contains the residuals (or what was left over after subtracting the mean)  
pop\_null = data.frame(resids=c(res\_machine\_A,res\_machine\_B,res\_machine\_C), shields$Machine)  
  
 #for each machine, calculate the mean of the residuals within this new dataframe  
 #why display this recentering Mean of 0? Well it shows that our populations have equal means (which is the null hypo here).  
with(pop\_null, tapply(resids,shields$Machine,mean))

## A B C   
## 5.329071e-16 -3.552714e-16 2.131975e-15

#set the number of bootstrap resamples we would like to take  
num\_bootstraps = 10000  
 #create an empty vector that will store the F\_star-statistics from the bootstrapping process. Size = num\_bootstraps.  
Fstar1 = numeric(num\_bootstraps)  
  
 #complete the number of bootstrap iterations as specified  
for (i in 1:num\_bootstraps){  
 #for each iteration, create a new dataframe that contains randomly selected residuals from the "pooled" data of Machine A/B/C  
 #this builds a "re"sampled dataset with replacement and has both the residual and the machine letter.  
 pop\_null = data.frame(resids = sample(c(res\_machine\_A,res\_machine\_B,res\_machine\_C), replace = TRUE), shields$Machine)  
   
 #now calculate the F\_star-statistic (i.e. the bootstrapped pseudo statistic of ANOVA) from the current dataframe.  
 #this calculated F\_star-statistic is stored in the Fstar1 vector  
 Fstar1[i] = oneway.test(resids~shields.Machine, data = pop\_null, var.equal = FALSE)$statistic  
}  
  
 #for each pseudo F\_star-statistic that was calculated, if it is NA, then replace it with 100\*the F=statistic from above (3.97)  
Fstar1[is.na(Fstar1)] = 100\*F\_observed  
  
 #first, if the value of the pseudo f\_star-statistic is greater than the original F, then keep it and sum it up to numerator  
 #take the summed values where Fstar is greater, and divide that by the number of bootstrap iterations - this is pvalue.  
P = sum(Fstar1 > F\_observed) / num\_bootstraps  
  
P

## [1] 0.0578

1. At a 5% level of significance, there is not enough evidence to suggest that there is a true difference in population mean diameter (in millimeters) of insulation shields for electrical wires from either of the machines at this company (p-value = 0.0578).

### Part 2j

Repeat the bootstrapped ANOVA test using unpooled residuals and unequal variances as in the notes. (i) State the null and alternative hypotheses, (ii) use R to compute the test statistic and p-value, and (iii) write a conclusion in context . Go ahead and use the helper function or t1waybt do do this problem.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2j -|-|-|-|-|-|-|-|-|-|-|-

1. mu1 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine A at this company. mu2 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine B at this company. mu3 = the population mean diameter (in millimeters) of insulation shields for electrical wires from Machine C at this company.

H\_null: mu1 = mu2 = mu3 (all the population means are the same). H\_Alternative: At least one population mean is different than the others.

#loading the necessary library  
library(WRS2)

## Warning: package 'WRS2' was built under R version 4.0.2

#t1waybt() completes a bootstrapped ANOVA using unpooled residuals (i.e. keeping the populations seperate) and unequal variances  
 #in fact, it assumes both of these things (unpooled resids and unequal variance)  
 #this is useful in cases where the data has outliers or is skewed, because it does a "trimming" process, which removes tr=% of the values from either end.  
  
 #This doesnt solve a small sample size problem though!!!  
t1waybt(Diameter~Machine, data = shields, tr=0.1, nboot = 10000)

## Warning in t1waybt(Diameter ~ Machine, data = shields, tr = 0.1, nboot = 10000):  
## Some bootstrap estimates of the test statistic could not be computed.

## Call:  
## t1waybt(formula = Diameter ~ Machine, data = shields, tr = 0.1,   
## nboot = 10000)  
##   
## Effective number of bootstrap samples was 9963.  
##   
## Test statistic: 2.2086   
## p-value: 0.31095   
## Variance explained: 0.463   
## Effect size: 0.68

1. At a 5% level of significance, there is not enough evidence to suggest that there is a true difference in population mean diameter (in millimeters) of insulation shields for electrical wires from either of the machines at this company (p-value = 0.31064).

### Part 2k

Which seems better and why, the bootstrap procedure with the pooled or unpooled residuals?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2k -|-|-|-|-|-|-|-|-|-|-|-

The bootstrap technique with the pooled residuals seems like a better choice because the sample size within each group (i.e. Machine) seems a bit small to use the unpooled version. Yet there still seems to be an issue with the pooled version since the distributions are not similarly shaped in their distribution. Verdict is still out…

### Part 2l

Do any of the four statistical inference procedures used here provide a clear answer to the question of whether or not the three machines produce the same average inside diameter for the insulation shields?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2l -|-|-|-|-|-|-|-|-|-|-|-

As the narrative continues, none of these tests seem to meet their specific requirements for conditions (some come close, but not all the way), therefore I don’t feel we have a clear answer yet on if the three machines produce the same average inside diameter for the insulation shields.

### Part 2m

If you were responsible for conducting the statistical analysis here, what would you report to the engineer?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2m -|-|-|-|-|-|-|-|-|-|-|-

I would suggest getting more data! If the sample sizes for each machine increase, maybe we can use the unpooled bootstrapped ANOVA with a fair amount of confidence. However, at that point one of the more simpler tests might meet conditions, so it would be wise to save all the work we did in this notebook and rerun the tests once more data was made available. :)

### Part 2n

What impact do you think samples of sizes 5, 5, and 10 play here?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2n -|-|-|-|-|-|-|-|-|-|-|-

The sample sizes are just too small to extract the necessary information for these tests to work properly. I guess it really hits home the idea that bootstrapping cannot save you from small sample sizes!

### Part 2o

Often the Kruskall Wallis test is presented as a test of

the population distributions are all the same

the population distributions are not all the same,

but this is not what KW tests as this example shows. Take 3 random samples of size 100 from normal distributions having mean 0 and standard deviations 1, 10, and 50. If KW were testing the hypotheses above, then we should reject since these three distributions are clearly different. Run the KW test. You should get a large -value. Why did you get a large -value when the distributions are so different?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2o -|-|-|-|-|-|-|-|-|-|-|-

#set the seed value for reproducability  
set.seed(321)  
 #create 3 sets of 100 random numbers from the normal distribution with same means but different std  
x <- c( rnorm(100,0,1), rnorm(100,0,10), rnorm(100,0,50))  
 #generate randomly factored elements to the data above.  
groups <- factor( (rep( c('A','B','C'), each=100 ) ) )  
 #perform the kruskal-wallis test  
kruskal.test(x~groups)

##   
## Kruskal-Wallis rank sum test  
##   
## data: x by groups  
## Kruskal-Wallis chi-squared = 0.84306, df = 2, p-value = 0.656

The Kruskal-Wallis test is testing that the distributions (and therefore medians) of those populations are shifted from each other - that is to say, moved up or down relative to each other. Kruskal-Wallis does not test that the distributions are equal to each other. Having similar shape distributions is different from distributions that are shifted. Finally, the pvalue = 0.656 on this particular Kruskal-Wallis test confirms that the median from each group are equal, which is true because we used a normal distribution with the same mean (which in theory should all generally have the same center point)