STOR 455 Class 38 Homogeneity of Variances and Contrasts

library(readr)  
library(car)  
library(Stat2Data)  
  
Exams4 <- read\_csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/Exams4.csv")

*Goal: Trying to see if there is a dif in means bt groups* - Look how we can test variability between means

**ANOVA for Difference in K Means** Data: Samples from K different groups Summary statistics: n1, ybar1, s1 for each group Combine all

Test: Ho: mu1 = mu2 = muk Ha: Some mui != muj

**Checking Conditions for ANOVA** E~N(0, Std of Errors) Check with residuals

* Zero Mean: Always holds for sample residuals
* COnstant Variance: Plot residuals vs. fits and/or compare std. dev.’s of groups (Check if some group si is more than twice another)
* Normaility: Histogram/normal plot of residuals
* Independence:Pay attention to data collection

**Example: Five Students** *Question of intereset: Is there a significant difference in average grade among the five students?* - Test tells us if there is sig, but it doesn’t always mean useful - Found there are sig dif by student, and the variability are suspect to assume there weas constant variance here (Betsy’s scores were all sim to each other while Bud was all close) - Laready groups so that the first student is grouped with their scores, etc, etc

**ANOVA for Grades vs. Students** *Can we assume equal variances?*

tapply(Exams4$Grade, Exams4$Student, median)

## Barb Betsy Bill Bob Bud   
## 75.5 91.5 77.5 82.5 48.5

medians = rep(tapply(Exams4$Grade, Exams4$Student, median), each=4)  
medians # Median scores for each student; some skew, which makes it different than means

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 75.5 75.5 75.5 75.5 91.5 91.5 91.5 91.5 77.5 77.5 77.5 77.5 82.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 82.5 82.5 82.5 48.5 48.5 48.5 48.5

# Want to keep track of not comparing this with an anova analysis   
# Want to see how far away each score is from the median value   
#Analysis of the variances of these absolute devations

* Need to take the abso score - media to make a new array with all the values
* It’s going to work nicely here, but it’s not going to look unless we make the data look like this data

**Levene’s Test for Equality of Variances**

Test: Ho: variance1 = variance2 =…=variancek H1: some variancei!=variancej

(y-ybar) = (ybark-ybar) + (y-ybark) where (y-ybark) conducts analysis of variance on teh collection of absolute deviations abs(y-ytildek); where ytildek = median of group k

LeveneData = abs(Exams4$Grade - medians)  
LeveneData

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 13.5 11.5 1.5 1.5 2.5 3.5 5.5 2.5 9.5 15.5 4.5 4.5 3.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 14.5 12.5 3.5 1.5 14.5 20.5 1.5

Levene\_aov = aov(LeveneData~Exams4$Student)  
summary(Levene\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Exams4$Student 4 88.8 22.20 0.567 0.69  
## Residuals 15 587.0 39.13

leveneTest(Exams4$Grade, factor(Exams4$Student))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.5673 0.6903  
## 15

**Levene’s Test for Grades versus Students** Ho: variance(barb) = variance2 = variance3 = etc Ha: variancei != variancej; for at least one pair of students (i,j)

*Conclusion*: There is not statistically significant evidence to suggest that students’ grades have difference variances.

medians=rep(tapply(Exams4$Grade, Exams4$Student,median),each=4) # Each = 4 because we want barbs 4 times, betsy 4 times, etc.   
LeveneData=abs(Exams4$Grade-medians) #They are lined up perfectly right in this dataset; so all the grades will be subed by each student's median score.   
LeveneData

## Barb Barb Barb Barb Betsy Betsy Betsy Betsy Bill Bill Bill Bill Bob   
## 13.5 11.5 1.5 1.5 2.5 3.5 5.5 2.5 9.5 15.5 4.5 4.5 3.5   
## Bob Bob Bob Bud Bud Bud Bud   
## 14.5 12.5 3.5 1.5 14.5 20.5 1.5

Levene\_aov = aov(LeveneData~Exams4$Student) # Do we see any sig dif in the subbed median by actual score?   
#Testing: Build model assume the var are equal in each group and trying to see if we have edv that there are some dif here; same as like anova model, but at var   
# most of the sum of squares are in resdu, which assum there are equal variance; little are in sumsquares of the model itself (grand variance)   
# Go in with teh same ideas, but with variances and assume that htey are equal, then the p value is big, we dont have evidence to support Ha; even though the rule of variance not being double another, but the test it didnt show sign  
# Comes down to sample size; we would have a lot of dif to have a sig dif   
# tehere is dif here, but its not big enough to reject Ho  
summary(Levene\_aov)

## Df Sum Sq Mean Sq F value Pr(>F)  
## Exams4$Student 4 88.8 22.20 0.567 0.69  
## Residuals 15 587.0 39.13

library(car)   
leveneTest(Exams4$Grade, Exams4$Student) # It wants grade and student (what we are splitting grade up with); it shouldnt really matter if you facotr it

## Warning in leveneTest.default(Exams4$Grade, Exams4$Student): Exams4$Student  
## coerced to factor.

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.5673 0.6903  
## 15

**Cancer Survival with Ascorbate Supplement**  - Use this when you have issues with constant variance and how to use transofmraitons to help with that

**INformation about the data:** *In the 1970’s doctors wondered if giving terminal cancer patients a supplement of ascorbate would prolong their lives. They designed an experiment to compare cancer patients who received ascorbate to cancer patients who did not receive the supplement. The result of that experiment was that, in fact, ascorbate did seem to prolong the lives of these patients. But then a second question arose. Was the effect of the ascorbate different when different organs were affected by the cancer? The researchers took a second look at the data. This time they concentrated only on those patients who received the ascorbate and divided the data up by which organ was affected by the cancer. They had 5 different organs represented among the patients (all of whom only had one organ affected): Stomach, bronchus, colon, ovary, and breast.*

* Looking at treatmetn for cancer; is the outcome dife for the type of cancer the people have?

data("CancerSurvival")  
head(CancerSurvival)

## Survival Organ  
## 1 124 Stomach  
## 2 42 Stomach  
## 3 25 Stomach  
## 4 45 Stomach  
## 5 412 Stomach  
## 6 51 Stomach

**Cancer Survival with Ascorbate Supplement** Test:

Ho: mu1 = mu2 = …= muk Ha: Some mui != muj

* Want to see if the avg # days are equal depn on type of cancer or if no dif in treatment vs cancer type

Cancermod = aov(Survival~Organ, data=CancerSurvival)  
summary(Cancermod) # Sig evid to say surivial time changes depending on the organ that has cancer

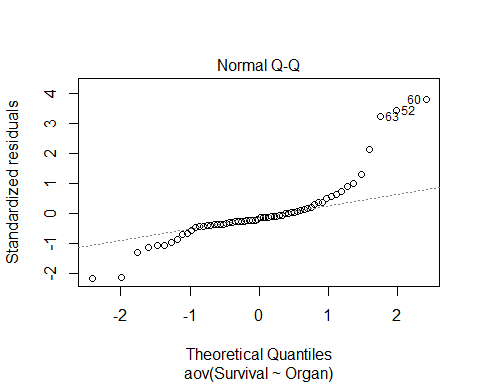
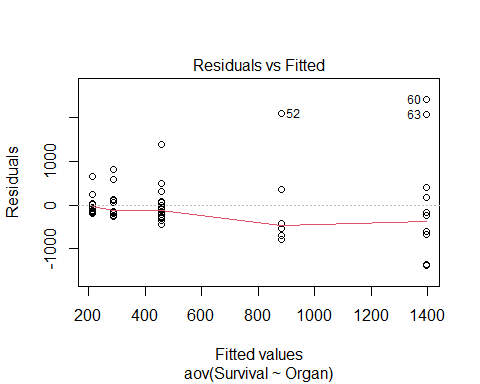
## Df Sum Sq Mean Sq F value Pr(>F)   
## Organ 4 11535761 2883940 6.433 0.000229 \*\*\*  
## Residuals 59 26448144 448274   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# THIS IS ASSUME CONDITIONS ARE MET

**Cancer Survival with Ascorbate Supplement**  *Check conditions* - See there is a trend in variability increases - The organ that has the lowest survial rate = more compact - as survival rate for organ increase as - clear pattern in plot

* Normal resiudal s
* off because the tails are wonky

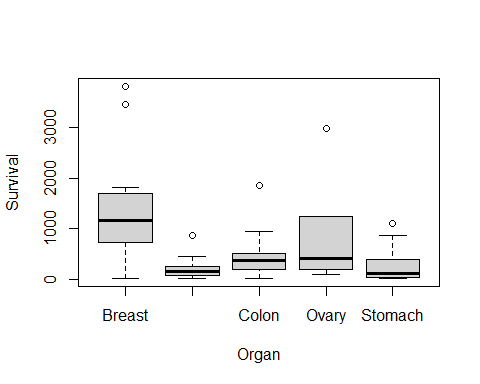
plot(Cancermod, 1:2)



**Cancer Survival with Ascorbate Supplement**  Ho: variance1 = variance2=…=variancek Ha: Some variancei != variancej

* Looking at visuals of survival by organ

boxplot(Survival~Organ, data=CancerSurvival)



tapply(CancerSurvival$Survival, CancerSurvival$Organ, sd)

## Breast Bronchus Colon Ovary Stomach   
## 1238.9667 209.8586 427.1686 1098.5788 346.3096

leveneTest(CancerSurvival$Survival, CancerSurvival$Organ)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 4 4.4524 0.003271 \*\*  
## 59   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

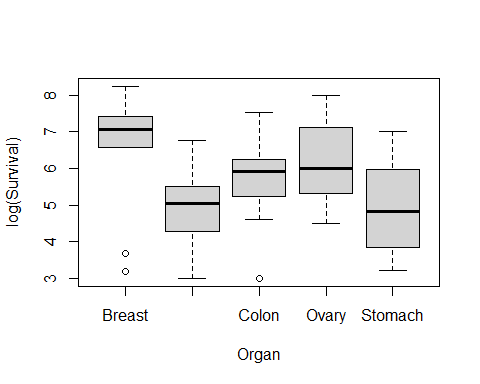
# Constant variance is a huge issue here   
  
#Levenetest gives a really small pvalue

* are these sim by chance or
* we have evid to say that the var is dif than the others
* we cant really use the results because the conditions aren’t good

**Would a transformation help?** - Yes, if you can make it look more normal then you will be good

LeveneTest Ho: variance1=varaince2=…varicnek Ha: some variancei != variancej

boxplot(log(Survival)~Organ, data=CancerSurvival)



tapply(log(CancerSurvival$Survival), CancerSurvival$Organ, sd)

## Breast Bronchus Colon Ovary Stomach   
## 1.6477550 0.9534041 0.9974766 1.2569313 1.2502073

leveneTest(log(CancerSurvival$Survival), CancerSurvival$Organ)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 4 0.6685 0.6164  
## 59

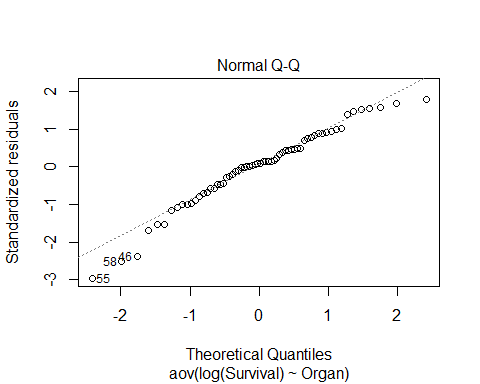
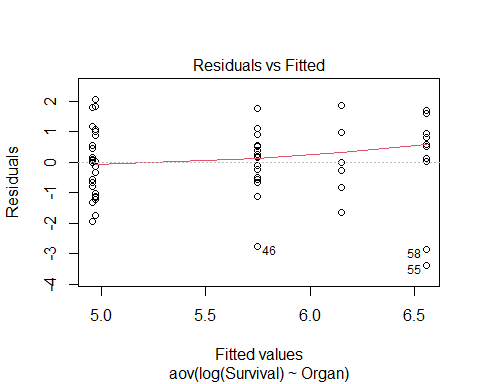
* above looks a lot better
* the sd are much closer and they are squished better
* the levenetest is also better with teh pvalue is high
* now we dont see there is dif
* we like higher pvalue in this

Cancermod\_log = aov(log(Survival)~Organ, data=CancerSurvival)  
summary(Cancermod\_log)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Organ 4 24.49 6.122 4.286 0.00412 \*\*  
## Residuals 59 84.27 1.428   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

*Cehck conditions* - not as big of an issue - not perfect for QQNorm, but its beter

plot(Cancermod\_log, 1:2)



TukeyHSD(Cancermod\_log) # shows us that Bronchus-Breast and stomach-breast are sig

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = log(Survival) ~ Organ, data = CancerSurvival)  
##   
## $Organ  
## diff lwr upr p adj  
## Bronchus-Breast -1.60543320 -2.906741 -0.3041254 0.0083352  
## Colon-Breast -0.80948110 -2.110789 0.4918267 0.4119156  
## Ovary-Breast -0.40798703 -2.114754 1.2987803 0.9615409  
## Stomach-Breast -1.59068365 -2.968399 -0.2129685 0.0158132  
## Colon-Bronchus 0.79595210 -0.357534 1.9494382 0.3072938  
## Ovary-Bronchus 1.19744617 -0.399483 2.7943753 0.2296079  
## Stomach-Bronchus 0.01474955 -1.224293 1.2537924 0.9999997  
## Ovary-Colon 0.40149407 -1.195435 1.9984232 0.9540004  
## Stomach-Colon -0.78120255 -2.020245 0.4578403 0.3981146  
## Stomach-Ovary -1.18269662 -2.842480 0.4770864 0.2763506

*Below shows the difference between small level what is “sig” and big sample level what is “sig”* - As sample size increases, if its sig doesn’t mean its pracitcal to us - we want ot make sure its true roughly, but we might have practical difference - we dont always use it because it can be sus

n=10 # 10 groups, and taking 10 random groups; of 4 groups; and binding them todaya dn making a df;   
  
#Name the data you randomly call   
x1 <- rnorm(n)  
x2 <- rnorm(n)  
x3 <- rnorm(n)  
x4 <- rnorm(n)  
  
x = c(rbind(x1, x2, x3, x4))  
data <- data.frame(x)  
  
data$group[ 1:n ] = "A"  
data$group[( n + 1):(2\*n)] = "B"  
data$group[(2\*n + 1):(3\*n)] = "C"  
data$group[(3\*n + 1):(4\*n)] = "D"  
  
#Lavene test  
# Should all have the same sd from population; by chance what the vlav test would look like   
leveneTest(data$x, factor(data$group))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.673 0.5742  
## 36

# Get sds  
tapply(data$x, data$group, sd)

## A B C D   
## 1.2604865 0.9596752 1.1035292 0.6933545

# D is almost double and levne pval = 0.2313  
# If we run a few times, then we eventually get something closer to 0.05  
# Gives idea of the dif that we would need to see for levene test to not like the dat aor say its sig

n=400 # 400 groups, and taking 10 random groups; of 4 groups; and binding them todaya dn making a df;   
  
#Name the data you randomly call   
x1 <- rnorm(n)  
x2 <- rnorm(n)  
x3 <- rnorm(n)  
x4 <- rnorm(n)  
  
x = c(rbind(x1, x2, x3, x4))  
data <- data.frame(x)  
  
data$group[ 1:n ] = "A"  
data$group[( n + 1):(2\*n)] = "B"  
data$group[(2\*n + 1):(3\*n)] = "C"  
data$group[(3\*n + 1):(4\*n)] = "D"  
  
#Lavene test  
# Should all have the same sd from population; by chance what the vlav test would look like   
leveneTest(data$x, factor(data$group))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 3 0.6395 0.5896  
## 1596

# Get sds  
tapply(data$x, data$group, sd)

## A B C D   
## 0.9938886 1.0174135 1.0623579 1.0265391