STOR 455 Class 35 One Way ANOVA (Again)

Homework 8 comments - you have to pick 6 dif models of car and use a random sample selection of 50 cars - choose these cars that are from jap or us companies; they are not part of the sample, but you have the car names - you’ll have to deduce which are jap vs us cars; do based on what they car type name is; think about the company - choose 3 jap cars that are a compact, midsize and suv - you have to find this yourself - Goal: take samples of 50 and find the compact, midsize, and suv for - use rbind like HW 4 to put all in the same dataframe (left join) - make a new variable - a bit of a setup - reference hw 2 and 4

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

Last time talked about anova for a difference in k mean - if we split it up into groups and classify the mean for the group that it is in, does that explain a significant amount more of the variability? Aka: Should we classify things by their group mean or the overall mean?

*Hypo test* Ho: mu1 = mu2=…muk Ha: some mui != muj

We looked to see if there is sig dif in the average grade among the four exams? Or are teh results by chance?

*Partiioning Variability* Data = Model + Error Y = mui + E total varaition in the response, Y = variation explained by MODEL + Unexplained variation in Residuals

Question: Does the model explain a sig amount of the total variability?

We want to know if we should

the error term is what variabiltiy there is grom each group mean

*Partitioning Variability ANOVA for Group Means* Y = muk + E SSTotal = SSGroups + SSE

The SSTotal = the SSdifferences from the overal gran mean SSE = sum of the difference from each indiivdual group mean SSGroup = the variation explained by the groups that we have imposed on this model

If we have sig evi of the diff bt the means, then the model will have SSE = small compared to the SSGroup; b/c the SSGroup tells us if a lot of the variability has been explained by using the group means

*Below is the model we made last class*

Ho: all exam means are equal in population Ha: at least one exam mean is different

The pvalue is big, so there is no evidece for sig dif

**Below:** IF we sum these, then we get the total SS; SSTotal; if we have a good model, then we would want a lot of variability explained by our groups factor(Exam) and we see there is a lot in the residuals, which is below factor(Exam) - this is why we have a large p value

amodG=aov(Grade~factor(Exam),data=Exams4)  
summary(amodG)

## Df Sum Sq Mean Sq F value Pr(>F)  
## factor(Exam) 3 1030 343.3 1.056 0.395  
## Residuals 16 5200 325.0

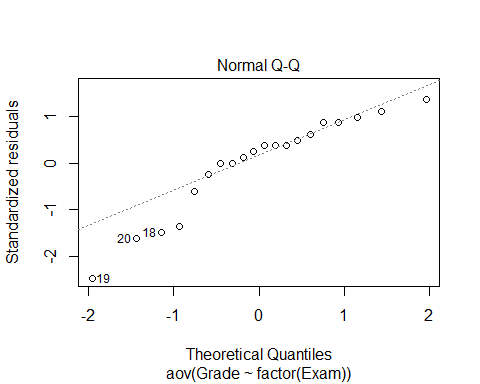
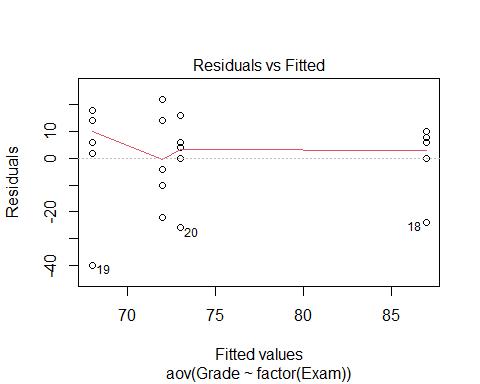
*Checking conditions for ANOVA* *Mostly have to focus on constant variance and normality* -Zero mean: Always holds for sample residuals - Constant Variance = plot residuals vs fits and/or compare std dev.’s of group (check if some groups is is more than twice another) *can be hard to check since we ar looking at groups; and sometimes its just teh mean values; we want to see a similar spread between the groups; can also think that as long as the largest standard deviation is not double the smallest standard deviation, then we ar egood enough for constant variance.* (Think more about the data, because if you have a really small dataset, you might have to throw that idea out the window) - normality: histogram/normal QQplot of residuals - Independence: Pay attention to how the data was collected

**Below: Checking Residuals** This amount of devation is not a problem

**Below: We check to see wht the std are of the functions**

This will split it by categories.

plot(amodG, 1:2)



round(tapply(Exams4$Grade,Exams4$Exam,sd),2)

## 1 2 3 4   
## 17.89 13.93 23.24 15.68

*Question* Is there sig dif in the average grade among the five students? Is it a sig dif?

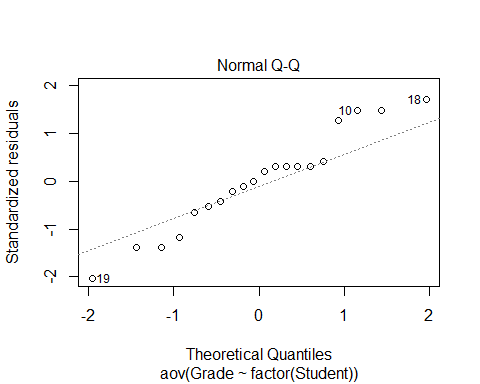
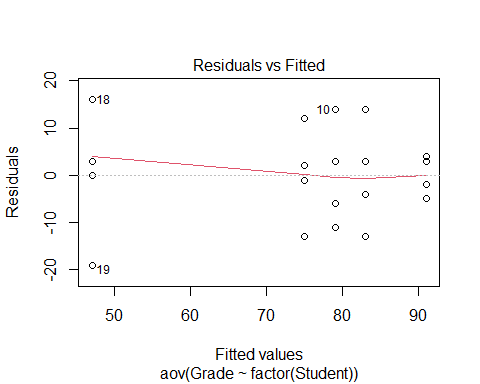
We can build the same model, but just change the group by student level instead of grade. we have evidence to say there is a sig dif.

This plot tells us where each persono’s scores are plotted in relation to each other

modS=aov(Grade~factor(Student),data=Exams4)  
summary(modS)

## Df Sum Sq Mean Sq F value Pr(>F)   
## factor(Student) 4 4480 1120.0 9.6 0.000468 \*\*\*  
## Residuals 15 1750 116.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(modS, 1:2)



round(tapply(Exams4$Grade,Exams4$Student,sd),2)

## Barb Betsy Bill Bob Bud   
## 10.30 4.24 10.98 11.40 14.45

*How many comparisions?* Which one of the indivates caused the sig difference?

There could be a lot of comparisions to do.

*Problem of Multiplicity* When we are doing many pairwaise comparisions we are more likely to make a Tpye I Effor (finding a false difference)

**Possible fixes:** - Do only a few pre planned comparisons – If we know what we want to focus on, then we can do that - Adjust the sig level used for each test

* THe more testes we do, the more like we are th make a type 1 error = null is true, but we find evidence to reject it

*Example:* If we did tests at a .05 sig level, then there is a 5% chance that we will make a type 1 error; so there is a 95% chance that we won’t make a type 1 error.

We have to find out how we can minimize the chance of making a tpye 1 error.

# The probability of not making a type 1 error if you run 10 tests  
(1 - 0.05)^10

## [1] 0.5987369

# Chance that we make at least 1 type 1 error when we do 10 tests  
1 - (1 - 0.05)^10

## [1] 0.4012631

# The probability of not making a type 1 error if you run 100 tests   
(1 - 0.05)^100

## [1] 0.005920529

# Chance that we make at least 1 type 1 error when we do 100 tests  
1 - (1 - 0.05)^100

## [1] 0.9940795

*Pairwise comparision AFTER ANOVA* This is where you do every single test on the data

Compute a C.I for mui - muj

Pairwaise t-test for difference in means Ho: mui = muj Ha: mui != muj

*Modifications* - estimate any std with sqrt(MSE) = Se - Use teh error d.f for any t-values

# We assumed in teh null that these all came from the same population.

Rather than using these ind groups to see the std, you can use the pooled value of all of the groups; we can estim the population std using the std of the residuals in teh model

*Pairwise Inference After ANOVA* C.I for mui - muj = (ybari - ybarj) +/- t-star*sqrt(MSE)*sqrt((1/ni)+(1/nj)) #Note: use the error d.f for the t-star in both the C.I and Test # If 0 is in the C.I then that would give me evidence for a stat sig dif. # The hypothesis testing tells you how unlikely it would be to get that result

Test: Ho: mui = muj Ho: mui != muj

t = (ybari-ybarj)/(sqrt(MSE)\*sqrt((1/ni)+(1/nj)))

**T-Star tells you how many standard errors i need to go in each direction depending on the sampel size; gives you an approximation for what you think the population difference is**

* Bigger sample, fewer in each direction = trend towards a normal distribution
* MSE = standard devation of the residuals and it is weighted based on the size of the groups
* If zero is in the C.I, then we have evidence

**T-Test Stat** **Is what the actual difference is divided by that measure of the standard deviation scales in this way. Tells me how many standard deviations away from the null this data is**

*Fisher’s LSD* The least sig difference: adding or subing to those differences to see how much variability there is

LSD = t-star*sqrt(MSE)*sqrt((1/ni)+(1/nj)) # concludes mui differs from uj if abs(ybari-ybarj)>LSD

*If looking at the t-distribution* The thing is centered at zero and we want to know how far in each direction we need to go to get the middle some percent of the data (whatever level you’re doing the sig test at); 95% = 0.05 sig level

We want to figure out what t-star is, because it will tell me how many i need to go in each direction. We want the value of t-star is 97.5% of the data under that curve to the left of it. **Look below, this is where we do the code for this**

# qt, give it an area under the curve, under the t-dist; and it will tell me the t value that is the bound for the left of it.   
  
#We also need df, and it's coming back from the anova model we made. We want the residuals df. We have 5 groups, because 5 students. This just gives us an idea of how much we trust the variability of the sample and how well it can predict the population   
  
# Big df = big sample, so the sample = more like the population   
# Small sample sizes = variability might not predict population as well   
  
qt(1-0.05/2, 15) #Tells you if you go OUTPUT number of standard deviations in each direction of this distribution, then I will capture the middle 95% of the data.

## [1] 2.13145

t\_LSD = qt(1 - 0.05/2, modS$df.residual)  
  
#WE want to make a prediction for the difference in population of the two groups, so we need to go 2.xxx standard deviations in each direction; we will calc the std by the sqrt(MSE)\*sqrt((1/ni)+(1/nj)) ((This is the last half of the LSD model))  
  
MSE = summary(modS)[[1]][,3][2] # This looks at the anova table output from the summary, that's what teh 1 is; the 3 gives me the third column and all rows; the 2 gives the second element of the 3rd column is what we are pulling out  
  
LSD = t\_LSD \* sqrt(MSE)\*sqrt(1/4 + 1/4) # it is 4 because each student took 4 exams   
LSD

## [1] 16.27921

*Interpreting the results above:* If we do each test basically at the 5% sig level, if any of your group means are more than this amount different, we have sig evidence to say there is a difference there;

we are predicting the mean s bt any two groups are this much or less by chance

*ANOVA for Grades vs Students* - Every compariasion we make; everyone’s score ocmpared with bud’s are more than 16.33 apart; so we have evideicent to say barb and bettey are different and others as well.

* THe multiplicity is an issue here; We are doing these tests at the 0.05 level, there are chances of making a type 1 error. This is a small example, but as it gets bigger it could be an error

*The Problem of Multiplicity* - When doing many pairwise comparisions there is an increased change to make a type 1 error (find a false difference) - Fisher’s LSD may be too lenient

*Possible fizes* - Do only a few preplaed comparsions - Use a smaller alpha for each test

OR

**Bonferroni Adjustment** When doing m tests with a *overall* error rate of alpha-star, use alpha = alpha-star/m for each test

* *This adjustment, basically lowering the levels of each one of the individual tests. So the chance of making a tpye 1 error for all the tests together can be defined and then from there you can figure out how twhat a sig level you need for each individual test*
* Bonferroni is going to let us do everything at the 0.05 level, and not a bigger aggregate of this **That’s really useful**
* IT does this by reducing the smaller alpha for each test. Reduce it by dividing the sig level by whatever number of comparisions you want to do.
* If I want the overall sig to be 0.05 I need to divide that by 10 (because there are 10 comparisons), then each test will be done at the 0.005 level.

#Bonferroni Method  
# No type 1 errors  
(1 - 0.05/10)^10

## [1] 0.9511101

# At least 1 type 1 error  
1 - (1 - 0.05/10)^10

## [1] 0.04888987

All that really changes, are the number of std that you have to go to trap that middle value gets bigger

We have to go much wider to track Bonferroni effect; and by going wider, we ar emaking it less likely that we make a type 1 error, but we are also demanding that we have stronger evidence of a difference to see that sig difference

*ANOVA for Grades vs Students*

*How to calculate the Bonferroni t-value* See below:

We get a value of 25, so using this method, we have to be 25 difference in term of average score to show some sig dif bt groups

The results are the same because of our data, but for other examples there may be a different conclusion

**Things to not do** - DOnt p-value hack; don’t choose what fits your narrative

t\_bf = qt(1 - 0.05/10/2, modS$df.residual) # What changed is the division by 10 ebcause of the total number of comparisions we wnat to do   
  
BSD = t\_bf \* sqrt(MSE)\*sqrt(1/4 + 1/4)  
BSD

## [1] 25.09753

*Tukey’s HSD: Honestly Sig Dif* *Replace t-star with value of q-star from teh studentized range distribution with R*

HSD = (q-star/sqrt(2))*sqrt(MSE)*sqrt((1/ni)+(1/nj))

q-star depends on slpha, # groups = K, and error df

tukey’s test is a bit more in the middle when it comes up with getting the dif values

* trying to get some middle group where we reduce the chance of a type 1 error, but not reducing it as much as the last case where we are inflating the case of making a type 2 error.
* the only change you see here is how many standard deivations you need to go in each difrection; it’s somewhere no quite in the middle; but its not as big as the last value
* Tukey’s is more the middle ground for what we would expect to see by chance

HSD = qtukey(1-0.05, 5, modS$df.residual) \* sqrt(MSE)\*sqrt(1/4 + 1/4)

*How can this be automated better by R?* *Automating the LSD and Bonferroni in R*

* Som eof the pvalues have changed

**LSD Method:**

pairwise.t.test(Exams4$Grade, Exams4$Student, p.adj = 'none') # What this is doing is doing each individual pairs of hypothesis test. It's not calcuing the conf int. It's teh same idea; where if the pvalue is below 0.05, then the dif bt these must be bigger than that 16.3 or something in the previous calculations.

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Exams4$Grade and Exams4$Student   
##   
## Barb Betsy Bill Bob   
## Betsy 0.05357 - - -   
## Bill 0.60812 0.13699 - -   
## Bob 0.31148 0.31148 0.60812 -   
## Bud 0.00229 3.8e-05 0.00079 0.00028  
##   
## P value adjustment method: none

# We see where the differences are in this table   
# We see what we saw with the confidence intervals, where one student has a sig difference, but others dont.   
# The table isn't filled out because where the dashes are, we already have those valeus; we have already compared bob and bill, but bill and bob would give us the same number

**Bonferroni Method in R**

pairwise.t.test(Exams4$Grade, Exams4$Student, p.adj = 'bonf')

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: Exams4$Grade and Exams4$Student   
##   
## Barb Betsy Bill Bob   
## Betsy 0.53567 - - -   
## Bill 1.00000 1.00000 - -   
## Bob 1.00000 1.00000 1.00000 -   
## Bud 0.02293 0.00038 0.00789 0.00277  
##   
## P value adjustment method: bonferroni

# Similar, but the bonferronia adjustment   
# A little differently done by R;   
# When we made the C.I, we adj the sig level of test based on the number of groups we had; we lowered teh sig level, which made me needs a lower pvalue to find a difference  
# R's difference: it's calc the pvalues in the same way for the LSD method, then multiplying that by the number of comparisions; so rather than compariting the orignal pvalue to the smaller sig level, it's comparing an inflated pvalue to that same sig level   
# Mathmatically its teh same thing, and you can se it here   
  
# The 1st for the pvalues show it is mult the last case by 10 and if its 1, that's the upper limit because you cant have a prob higher than that   
  
# THis compares each of the pvalues to the same 0.05 level since they are all being multiplyed by 10  
  
# Bud is the big difference dude here again

*Tukey Function* Does the middle ground of making type one and type 2 errors;

diff = diff in mean scores for each pair - lower and upper = 95% conf int using the Tukey’s distribution - Trying to see if 0 in this distribution or not, if 0 is in it, we are predicting that there could be no difference in teh population, if 0 is not in it, then we’re - if there is no 0 invloved, then there may be a sig dif bt the two

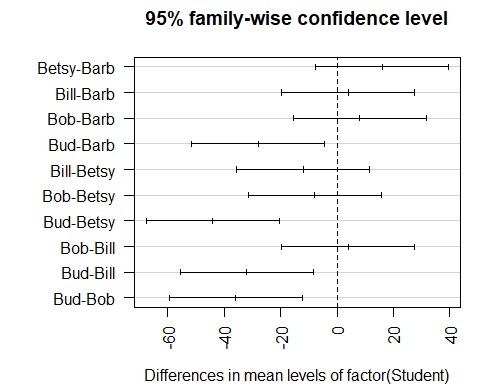
* we could look at the pvalue, which is being inflated; we compare then to the 0.05 level
* the results are the same

TukeyHSD(modS)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Grade ~ factor(Student), data = Exams4)  
##   
## $`factor(Student)`  
## diff lwr upr p adj  
## Betsy-Barb 16 -7.584413 39.584413 0.2720310  
## Bill-Barb 4 -19.584413 27.584413 0.9835150  
## Bob-Barb 8 -15.584413 31.584413 0.8295529  
## Bud-Barb -28 -51.584413 -4.415587 0.0166293  
## Bill-Betsy -12 -35.584413 11.584413 0.5360462  
## Bob-Betsy -8 -31.584413 15.584413 0.8295529  
## Bud-Betsy -44 -67.584413 -20.415587 0.0003116  
## Bob-Bill 4 -19.584413 27.584413 0.9835150  
## Bud-Bill -32 -55.584413 -8.415587 0.0060225  
## Bud-Bob -36 -59.584413 -12.415587 0.0021941

*Added Code* - Visual representation of this - We can plot all the hist on the same axis, and if any cross the vertial 0 line, then there is not sig dif there; - if they dont cross through it, then thats where we see those differences.

origpar = par()  
par(mar=c(4,7,3,1))  
hsd = TukeyHSD(modS)  
plot(hsd, las=2)



par(origpar)