Unit 6 HW Solutions

## Question 1 (25 points total)

Handicap Study. Use the Bonferroni method to construct simultaneous confidence intervals for , , and (to see whether there are differences in attitude toward the mobility type of handicaps).

, , , , and are the mean scores in the none, amputee, crutches, hearing, and wheelchair groups. Be careful when identifying “k” here. This study is mentioned throughout Chapter 6 of Statistical Sleuth.

*Note: the identification of the correct k is worth 4 points, and each of the three 95% CI’s are worth 7 points. If you use the wrong k but follow the procedure correctly, only 4 points should be deducted (even though all your answers will not match the key).*

**DF = 65, I = 5, k = 3, (from several displays in Chapter 6)**

**Using Bonferroni Method, , Margin of Error**

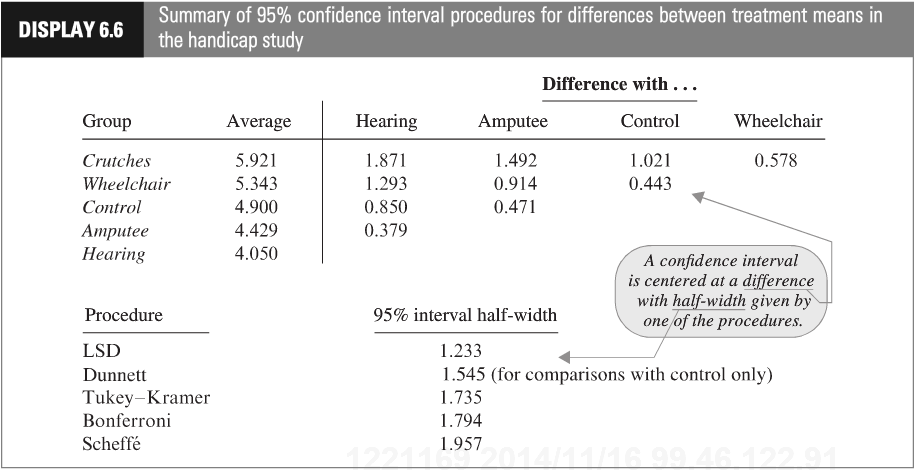
**95% CI for : .**

**95% CI for : .**

**95% CI for : .**

## Question 2 (25 points total)

Handicap Study. See what multiple comparison procedures are available within the one-way analysis of variance procedure. Verify the 95% confidence interval half-widths in Display 6.6.



Show your work for this problem by simply copying the code and relevant output for each comparison (cut and paste your code and relevant output). The half-width might be found directly from your output. If so, note where it is found. If not, show how you would use the output to find it. Do this for both R and SAS.

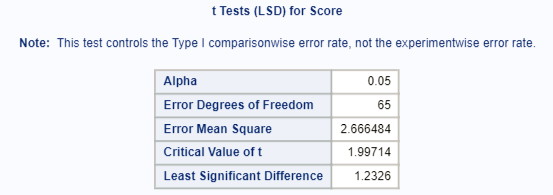
**SAS:**

\*To run several types of pairwise comparisons–not all of these may have been covered during live session but they are all provided to match the textbook output;  
proc glm data=handicapdata;  
class handicap;  
model score=handicap;  
means handicap /t;  
means handicap /dunnett (“None”);  
means handicap /tukey;  
means handicap /bon;  
means handicap /scheffe;  
run;

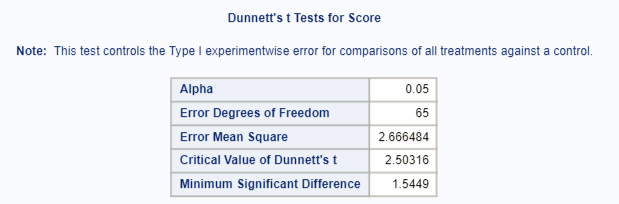
**We are fortunate that SAS provides the 95% CI half-width in the “least significant difference” box. For each confidence interval, we could also find the half-width using the critical value and standard error. The standard error for the difference in means of two of any two of these groups:**

**Every half-width is determined by a multiplier SE.**

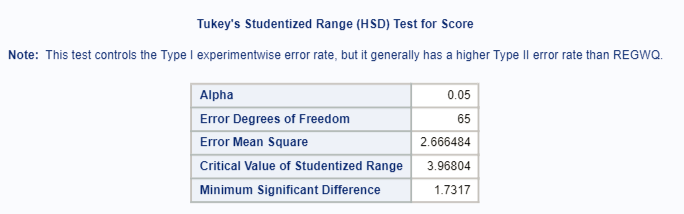
**LSD (t): Multiplier = 1.99714, 95% CI half-width =**



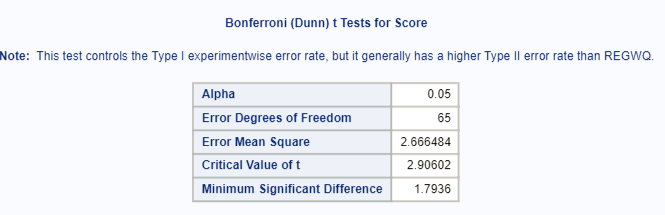
**Dunnett: Multiplier = 2.50316, 95% CI half-width =**



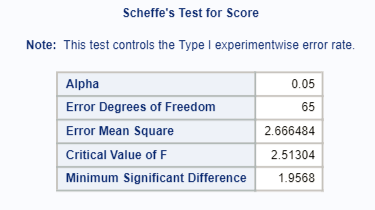
**Tukey: Multiplier = , 95% CI half-width =**



**Bonferroni: Multiplier = 2.90602, 95% CI half-width =**



**Scheffe: Multiplier = , 95% CI half-width =**



**R:**

# Use the agricolae package  
#References for agricolae  
#https://cran.r-project.org/web/packages/agricolae/agricolae.pdf  
  
library(agricolae)  
handicap <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 6/HW/Unit 6 Handicap Data.csv')  
  
aovHandi <- aov(Score ~ Handicap, data = handicap)  
  
handi.bonf <- LSD.test(aovHandi, 'Handicap', p.adj='bonferroni')  
handi.lsd <- LSD.test(aovHandi, 'Handicap', p.adj='none')  
handi.scheffe <- scheffe.test(aovHandi, 'Handicap')  
handi.tukey <- HSD.test(aovHandi, 'Handicap')  
  
handi.bonf$statistics[6]

## MSD  
## 1.79357

handi.tukey$statistics[5]

## MSD  
## 1.731733

handi.lsd$statistics[6]

## LSD  
## 1.232618

handi.scheffe$statistics[7]

## CriticalDifference  
## 1.956817

#For Dunnett's we need the multcomp package  
#References for multcomp  
#https://cran.r-project.org/web/packages/multcomp/multcomp.pdf  
library(multcomp)

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

gout <- glht(aovHandi, mcp(Handicap = 'Dunnett'))  
confint(gout)

##   
## Simultaneous Confidence Intervals  
##   
## Multiple Comparisons of Means: Dunnett Contrasts  
##   
##   
## Fit: aov(formula = Score ~ Handicap, data = handicap)  
##   
## Quantile = 2.5024  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate lwr upr   
## Crutches - Amputee == 0 1.4929 -0.0516 3.0373  
## Hearing - Amputee == 0 -0.3786 -1.9230 1.1659  
## None - Amputee == 0 0.4714 -1.0730 2.0159  
## Wheelchair - Amputee == 0 0.9143 -0.6302 2.4587

#Note that multcomp can do some of the other methods as well  
gout <- glht(aovHandi, mcp(Handicap = 'Tukey'))  
#confint(gout)

## Question 3 (50 points total)

Education and Future Income. Reconsider the data problem of Exercises 5.25 concerning the distributions of annual incomes in 2005 for Americans in each of five education categories. (a) Use the Tukey-Kramer procedure to compare every group to every other group. Which pairs of means differ and by how many dollars (or by what percent)? (Use p-values and confidence intervals in your answer.) (b) Use the Dunnett procedure to compare every other group to the group with 12 years of education. Which group means apparently differ from the mean for those with 12 years of education and by how many dollars (or by what percent)? (Use p-values and confidence intervals in your answer.)

This question is obviously from the book, but assume you are starting this problem from scratch. Show all parts: (1) Discussion of Assumptions (This could result in the inferences no longer being about the means. IF that happens, you should still compare the groups, just use the appropriate parameters when making inferences. Remember that you already did the work for addressing assumptions in prior homeworks.) (2) Selection and Execution of Test (3) Interpretation and Conclusion.

In short, perform a complete analysis like you usually do. Provide and interpret all the confidence intervals that suggest a significant difference in incomes; provide your SAS and R code as well. (Generate your statistics using both softwares.)

Finally, you should first test to see if any of the groups are different before you consider pairwise comparisons.

*Note: you have two options for this problem - working with logged data or working with the original data. Either is acceptable, and solutions for both approaches are outlined below. 20 points will be assigned to choosing an approach and implementing it correctly, and 15 points each will be assigned for the multiple comparisons in parts A and B.*

**Problem (2 points): How strong is the evidence that at least one of the five population distributions of education level has a different mean (median) income than any of the others?**

**Assumptions: The assumptions of the ANOVA are: the incomes in each educational group come from a normal distribution, the variances of these normal distributions are equal, the data are independent within each group, and the data are independent between each group.**

**Normality (3 points): The histograms and QQ plots below (of original data) appear to each show strong evidence of right skew, and thus provide evidence against coming from a normal distribution. This is not unexpected, as income data is often right skewed. However, each group has a sample size greater than 130, thus allowing the CLT to enable the ANOVA to be robust to this assumption. The log transformed data appears to be slightly less skewed (in the other direction), but only slightly.**

\*To address ANOVA assumptions on original data with histograms and QQ plots;

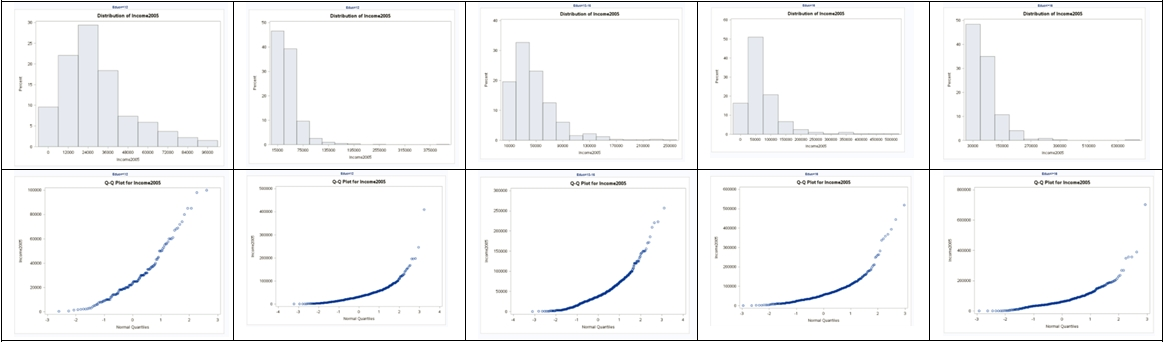
proc univariate data = incomedata;

by educ;

histogram income2005;

qqplot income2005;

run;



\*Perform a log transform;

data incomedata;

set incomedata;

logincome2005 = log(income2005);

run;

\*To address ANOVA assumptions on log transformed data with histograms and QQ plots;

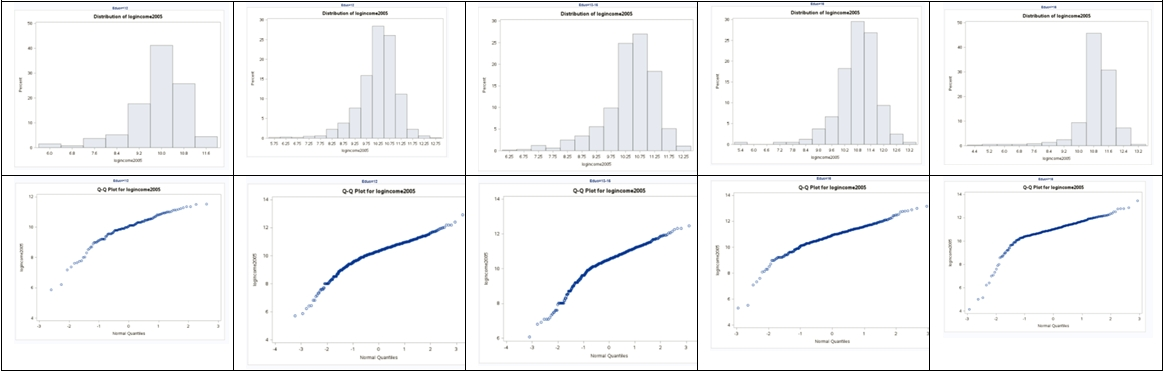
proc univariate data = incomedata;

by educ;

histogram logincome2005;

qqplot logincome2005;

run;



**Equal Standard Deviations (3 points): It appears that the original data shows evidence against equal standard deviations in the scatter plot, and this is further supported by the p-value of <0.0001 in a Brown-Forsythe test for equal variances. The log transformed data appears to show equal standard deviations via the scatter plot, and the p-value=.2377 > 0.05 in a Brown-Forsythe test also supports the equal standard deviation assumption for the logged data.**

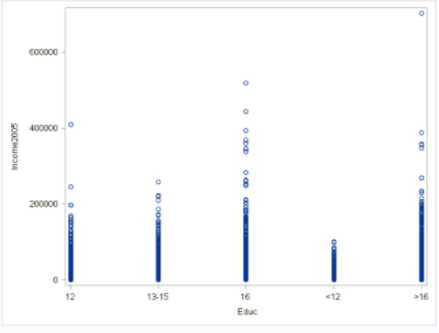
\*To address ANOVA assumptions on original data with scatter plots;

proc sgplot data = incomedata;

scatter x= educ

y = income2005;

run;



\*To test for equal variances in original data using tests (that do not require normality);

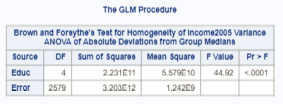
proc glm data = incomedata;

class educ;

model income2005 = educ;

means educ / hovtest = bf;

run;



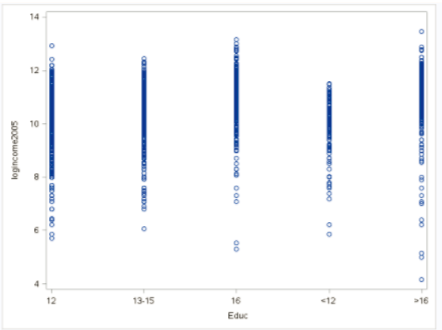
\*To address ANOVA assumptions on log transformed data with scatter plots;

proc sgplot data = incomedata;

scatter x= educ

y = logincome2005;

run;



\*To test for equal variances in logged data using tests (that do not require normality);

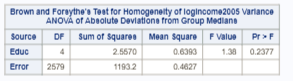
proc glm data = incomedata;

class educ;

model logincome2005 = educ;

means educ / hovtest = bf;

run;



**Independence (3 points): We will assume the data are independent, both between and within groups.**

**So, we can test the question of whether any of the groups are different from each other by:**

**1) Pure ANOVA on logged data. (Normality is somewhat fixed by logging, and we have the CLT; logging fixes the equal standard deviation assumption violation.) Note that this will change our inferences to be on the medians.**

**2) Welch’s ANOVA on original data. (While the underlying data does not appear to be normal, the CLT kicks in with the large sample size; Welch’s can handle unequal standard deviations.)**

*Note: hypotheses are worth 2 points, the test statistic is worth 1 point, the p-value is worth 1 point, the decision is worth 1 point, and the conclusion is worth 4 points.*

**Using pure ANOVA on the logged data**

**Step 1 - Hypotheses:**

**: All median incomes are the same across education levels.**

**: At least one pair of income medians are different between education levels.**

**Step 2 - Identification of Critical Value: You may skip step 2 (critical value) in ANOVA settings, although one could be found (and the comparison to the F statistic should match the p-value’s comparison to alpha).**

**Step 3 - Value of Test Statistic:**

**Step 4 - Give p-value:**

**Step 5 - Decision: Reject**

**Step 6 - Conclusion: There is strong evidence to suggest that at least one of the median incomes (median, not mean, because we used a log transform) for a particular education level is different from the others ( from a pure ANOVA).**

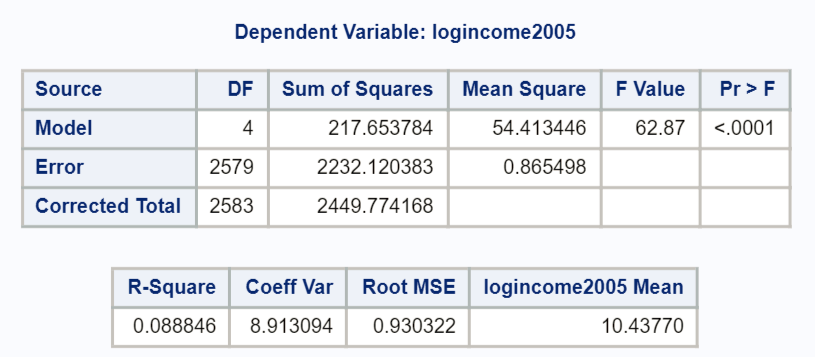
\*To perform ANOVA on log transformed data;

proc glm data = incomedata;

class educ;

model logincome2005 = educ;

run;



##Here is how to answer the problem using R  
##Read in the data, note your directory will be different  
  
edu <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 5/HW/ex0525.csv')  
  
edu$log.income <- log(edu$Income2005)  
  
edu.anova <- aov(log.income ~ Educ, data=edu)  
summary(edu.anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Educ 4 217.7 54.41 62.87 <2e-16 \*\*\*  
## Residuals 2579 2232.1 0.87   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Using Welch’s ANOVA on original data**

**Step 1 - Hypotheses:**

**: All mean incomes are the same across education levels.**

**: At least one pair of income means are different between education levels.**

**Step 2 - Identification of Critical Value: You may skip step 2 (critical value) in ANOVA settings, although one could be found (and the comparison to the F statistic should match the p-value’s comparison to alpha).**

**Step 3 - Value of Test Statistic:**

**Step 4 - Give p-value:**

**Step 5 - Decision: Reject**

**Step 6 - Conclusion: There is strong evidence to suggest that at least one of the mean incomes for a particular education level is different from the others ( from a pure ANOVA).**

\*To perform Welch’s ANOVA on original data;

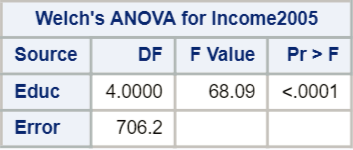
proc glm data = incomedata;

class educ;

model income2005 = educ;

means educ/ Welch;

run;



##In R:  
oneway.test(Income2005 ~ Educ, data=edu, var.equal=F)

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: Income2005 and Educ  
## F = 68.089, num df = 4.00, denom df = 706.18, p-value < 2.2e-16

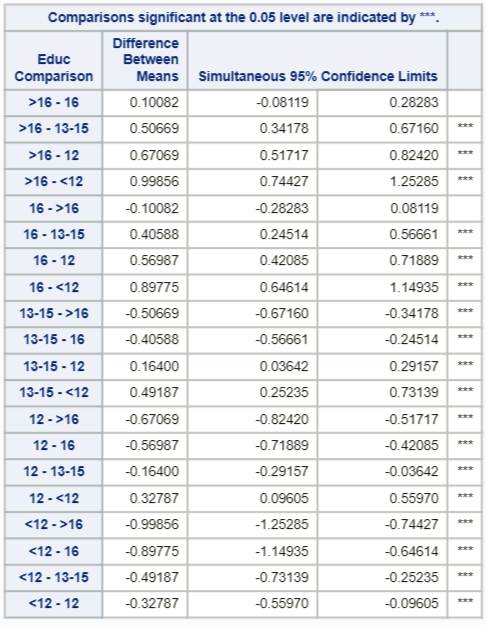
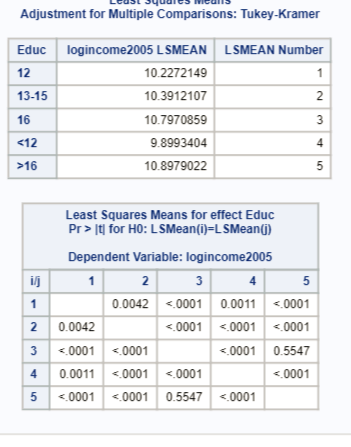
**Fortunately, Welch’s ANOVA on original data and pure ANOVA on log-transformed data are in agreement that at least one group is different from the others.**

**Now, onto multiple comparisons.**

### Part A (15 points)

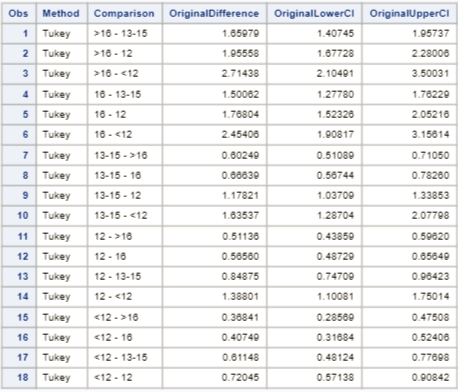
**Note that Tukey (Tukey-Kramer) requires normality and equal standard deviations, so we must perform the analysis on the logged data and make inferences accordingly. Confidence intervals and p-values for differences in means of logged data.**

*For part A, 5 points are given for correctly identifying the need to use the logged data, 5 points are given for implementing in either SAS or R, and 5 points are given for discussing the results. If you used the original data but implemented and interpreted those results, 5 points should be subtracted for using the wrong data (even though all your answers will be incorrect).*

**The only comparison that is not significant is group 3 vs. group 5, as evidenced by a p-value of 0.5547, and the confidence intervals on the logged means (the only interval that contains 0).**

**Confidence intervals for the multiplicative difference in medians for SIGNIFICANT differences.**



**An interpretation would be that when going from the “only some college” group (13-15) to the more than college graduate (>16), the median increases by an estimated factor of 1.66, or by 66%. A 95% confidence interval for this factor increase of the medians is (1.407, 1.957), or (40.7%, 95.7%). Other confidence intervals should be interpreted accordingly. Beware that if a decrease in incomes is discussed, caution must be used. For example, going from the more than college graduate group (>16) to the “only some college” group (13-15), the median of the latter (13-15) is only .6025 times the former (>16), with a 95% confidence interval for the factor decrease in medians of (.5109, .7105). That is, going from more than college graduate group (>16) to the “only some college” group (13-15), the median decreases by 1-.6025 = 39.75%, with a confidence interval for this percent decrease of (1-.5109, 1-.7105), or (28.95%, 48.91%). Note that all pairs appear twice in the table.**

\*To put output of confidence intervals into its own dataset, knowing the CIs will need to be unlogged;

ods output CLDiffs= ConfLevels;

ods trace on;

\*To perform Tukey comparisons on logged data;

proc glm data = incomedata;

class educ;

model logincome2005 = educ;

means educ/ tukey cldiff;

lsmeans educ / adjust = tukey;

run;

ods trace off;

\*To back transform the CI limits and report on significant comparisons;

data conflevels;

set conflevels;

OriginalDifference = exp(difference);

OriginalLowerCI= exp(LowerCL);

OriginalUpperCI= exp(UpperCL);

where significance = 1;

keep Method Comparison OriginalDifference OriginalLowerCI OriginalUpperCI;

run;

proc print data = ConfLevels;

run;

**Note: SAS proc glm provides confidence intervals with the “means” statement and adjusted p-values with the “lsmeans” statement. What is the difference between “means” and “lsmeans?” The latter “lsmeans” are least squares means while “means” are the simple arithmetic means that we commonly think of. These two are actually the same when there is no missing data (when sample sizes are equal for each group). But when one or more values are missing, the “lsmeans” (least squares means) calculate the mean as the average of the marginal (group) averages. An easy example of this can be found by following this url**

\*\*[http://webpages.uidaho.edu/cals-statprog/sas/workshops/glm/lsmeans.htm\*\*](http://webpages.uidaho.edu/cals-statprog/sas/workshops/glm/lsmeans.htm**)

##In R:  
edu.anova <- aov(log.income ~ Educ, data=edu)  
  
edu.tukey <- HSD.test(edu.anova, 'Educ')  
##To see the LSMeans  
edu.tukey$means

## log.income std r Min Max Q25 Q50  
## <12 9.89934 0.9988809 136 5.857933 11.51293 9.546813 10.06453  
## >16 10.89790 1.0665910 374 4.143135 13.46402 10.596635 11.01036  
## 12 10.22721 0.8539854 1020 5.703782 12.92393 9.902234 10.34174  
## 13-15 10.39121 0.9288173 648 6.061457 12.45794 10.085809 10.54534  
## 16 10.79709 0.9581051 406 5.298317 13.16031 10.373491 10.94196  
## Q75  
## <12 10.51867  
## >16 11.47210  
## 12 10.77896  
## 13-15 10.96820  
## 16 11.39639

##To see the pairwise differences, p-values and CI's  
TukeyHSD(edu.anova)$Educ

## diff lwr upr p adj  
## >16-<12 0.9985617 0.74427197 1.25285151 0.000000e+00  
## 12-<12 0.3278745 0.09605048 0.55969853 1.095421e-03  
## 13-15-<12 0.4918703 0.25234568 0.73139489 2.285656e-07  
## 16-<12 0.8977455 0.64614228 1.14934876 0.000000e+00  
## 12->16 -0.6706872 -0.82420002 -0.51717445 0.000000e+00  
## 13-15->16 -0.5066915 -0.67160306 -0.34177986 0.000000e+00  
## 16->16 -0.1008162 -0.28282718 0.08119474 5.547062e-01  
## 13-15-12 0.1639958 0.03642256 0.29156899 4.173877e-03  
## 16-12 0.5698710 0.42085061 0.71889141 0.000000e+00  
## 16-13-15 0.4058752 0.24513713 0.56661334 0.000000e+00

##To get the multiplicative differences on the raw scale  
exp(TukeyHSD(edu.anova)$Educ)

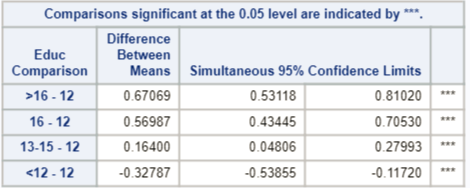
## diff lwr upr p adj  
## >16-<12 2.7143750 2.1049084 3.5003099 1.000000  
## 12-<12 1.3880148 1.1008146 1.7501448 1.001096  
## 13-15-<12 1.6353720 1.2870409 2.0779771 1.000000  
## 16-<12 2.4540642 1.9081654 3.1561368 1.000000  
## 12->16 0.5113570 0.4385857 0.5962028 1.000000  
## 13-15->16 0.6024856 0.5108889 0.7105046 1.000000  
## 16->16 0.9040992 0.7536500 1.0845821 1.741429  
## 13-15-12 1.1782093 1.0370940 1.3385260 1.004183  
## 16-12 1.7680390 1.5232567 2.0521569 1.000000  
## 16-13-15 1.5006153 1.2777965 1.7622887 1.000000

### Part B (15 points)

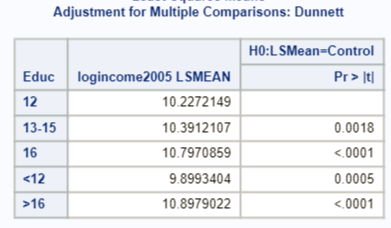
**Dunnett’s requires normality, equal standard deviations, and independence. So, we will proceed with Dunnett’s comparisons on the log transformed data.**

*Similar to part A, 5 points are given for correctly identifying the need to use the logged data, 5 points are given for implementing in either SAS or R, and 5 points are given for discussing the results. If you used the original data but implemented and interpreted those results, only 5 points should be subtracted for using the wrong data (even though all your answers will be incorrect).*

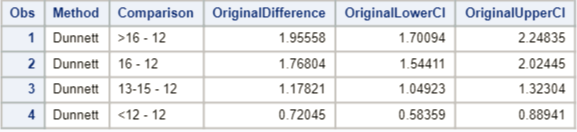
**95% confidence intervals on differences in means of logged data**



**P-values for Dunnett’s comparisons on log transformed data**



**95% confidence intervals on the multiplicative difference in medians of original data.**



**According to the confidence intervals and the multiple hypothesis tests, all means of logged data are significantly different from the high school degree only (12) group at a 0.05 familywise Type I error rate.**

**An interpretation would be that when going from the only high school diploma group (12) to the more than college graduate (>16), the median increases by an estimated factor of 1.956, or by 95.6%. A 95% confidence interval for this factor increase of the medians is (1.7001, 2.24835), or (70.0%, 125.8%). Other confidence intervals should be interpreted accordingly. Beware that if a decrease in incomes is discussed, caution must be used. For example, going from the only high school diploma group (12) to the high school dropout group (<12), the median of the latter (<12) is only .7205 times the former (12), with a 95% confidence interval for the factor decrease in medians of (.5836, .8894). That is, going from high school graduate only group (12) to the high school dropout group (<12), the median decreases by 1-.5836 = 41.64%, with a confidence interval for this percent decrease of (1-.8894, 1-.5836) or (11.06%, 41.64%).**

##In R:  
library(multcomp)  
edu <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 5/HW/ex0525.csv')  
edu$log.income <- log(edu$Income2005)  
edu.anova <- aov(log.income ~ Educ, data=edu)  
  
##We have to set the 'control' group to 12  
##It is not intuitive at all, but we first need to construct a table  
##of counts for each group, then define a contrast matrix with  
##12 as the reference level  
edu.count <- table(edu$Educ)  
base\_12 <- contrMat(edu.count, base=3, type="Dunnett")  
dunnett.edu <- glht(edu.anova, linfct=mcp(Educ=base\_12))  
  
##To view the confidence intervals on the log scale  
confint(dunnett.edu)$confint

## Estimate lwr upr  
## <12 - 12 -0.3278745 -0.5385372 -0.1172119  
## >16 - 12 0.6706872 0.5311874 0.8101871  
## 13-15 - 12 0.1639958 0.0480677 0.2799238  
## 16 - 12 0.5698710 0.4344535 0.7052885  
## attr(,"conf.level")  
## [1] 0.95  
## attr(,"calpha")  
## [1] 2.480533

##To get multiplicative intervals on original scale  
exp(confint(dunnett.edu)$confint)

## Estimate lwr upr  
## <12 - 12 0.7204534 0.5836053 0.8893907  
## >16 - 12 1.9555808 1.7009585 2.2483184  
## 13-15 - 12 1.1782093 1.0492456 1.3230241  
## 16 - 12 1.7680390 1.5441257 2.0244219  
## attr(,"conf.level")  
## [1] 0.95  
## attr(,"calpha")  
## [1] 2.480453

**Just in case you didn’t use log transformed data, here are the results you should have gotten, although the assumptions for these tests are not met!**

**Tukey**

\*Find Tukey comparisons on original data (erroneously) despite assumptions not being met;

proc glm data = incomedata;

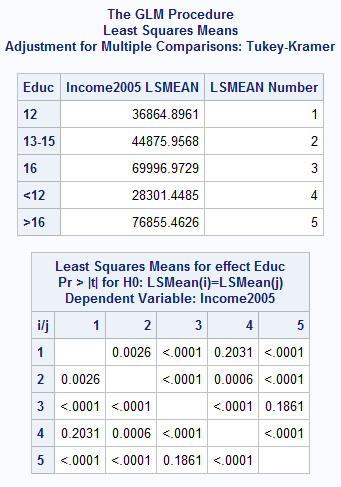
class educ;

model income2005 = educ;

means educ/ tukey cldiff;

lsmeans educ / adjust = tukey;

run;

**Every pairing besides more than college (>16) vs. only college (16) (p-value = 0.1861) and high school only (12) vs. high school dropout (<12) (p-value = 0.2031) have significantly different means.**

##In R:  
aovEduc <- aov(Income2005 ~ Educ, data = edu)  
edu.count <- table(edu$Educ)  
base\_12 <- contrMat(edu.count, base=3, type="Tukey")  
tukey.edu <- glht(aovEduc, linfct=mcp(Educ=base\_12))  
  
##To view the confidence intervals on the log scale  
confint(tukey.edu)$confint

## Estimate lwr upr  
## >16 - <12 48554.014 36684.669 60423.359  
## 12 - <12 8563.448 -2257.276 19384.171  
## 13-15 - <12 16574.508 5394.349 27754.668  
## 16 - <12 41695.524 29951.577 53439.472  
## 12 - >16 -39990.566 -47155.999 -32825.134  
## 13-15 - >16 -31979.506 -39676.994 -24282.017  
## 16 - >16 -6858.490 -15354.116 1637.137  
## 13-15 - 12 8011.061 2056.395 13965.726  
## 16 - 12 33132.077 26176.333 40087.821  
## 16 - 13-15 25121.016 17618.331 32623.701  
## attr(,"conf.level")  
## [1] 0.95  
## attr(,"calpha")  
## [1] 2.705164

**Limits are slightly different because of some differences in calculations when the sample sizes are different (Tukey-Kramer v. Tukey).**

**The interesting thing so see here is that different results are produced through analysis of the logged versus the unlogged data. With the logged data, only the more than college degree (>16) versus the college degree only (16) levels are found not to be significantly different with Tukey comparisons, whereas when Tukey comparisons are performed on original (unlogged) data, the comparison of high school dropout (<12) to high school diploma only (12) are also significantly different. This implies that the medians of the latter comparison are not significantly different. Perhaps outliers or heavily skewed data is at play! You may also infer that the assumptions for the logged data were better met and that the analysis on the logged data is more dependable. Finally, one may infer that since they both found significant difference in the upper education levels that there is stronger evidence of a difference. All of these have merit but must be supported by the assumptions being met for the individual test.**

**Similar analysis was done for Dunnett’s multiple comparison of each group against the high school level (12). Note that the intervals are narrower for Dunnett’s than the Tukey-Kramer counterparts above.**

\*Find Dunnett’s comparisons on original data (erroneously) despite assumptions not being met;

proc glm data = incomedata;

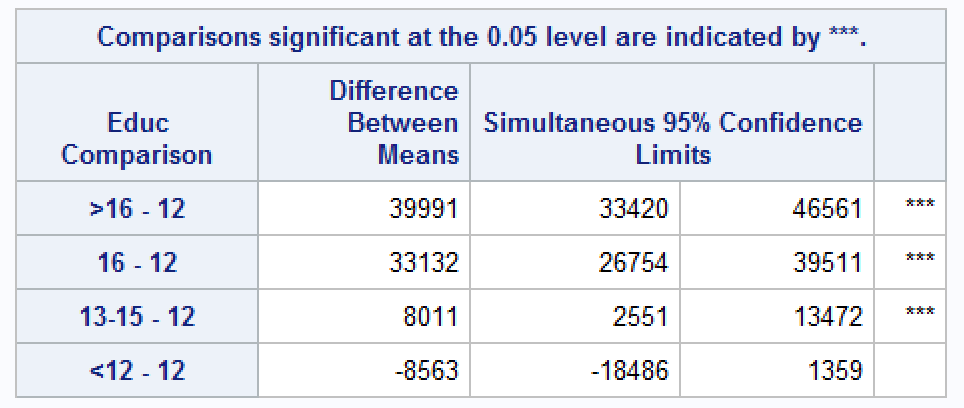
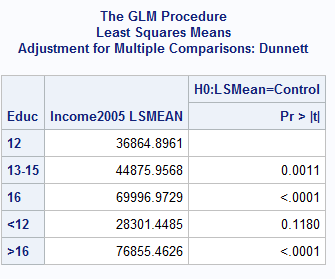
class educ;

model income2005 = educ;

means educ/ dunnett cldiff;

lsmeans educ / adjust = dunnett;

run;

**The table above shows that for the unlogged data, the only difference of means that is not statistically different is between less than high school (<12) and high school diploma only (12) levels. For the logged data, recall that all the levels have medians that are statistically different from the high school diploma only (12) median. Again, this can be interpreted as means vs. medians on the original scale and / or an argument can be made about which test’s assumptions are better met.**

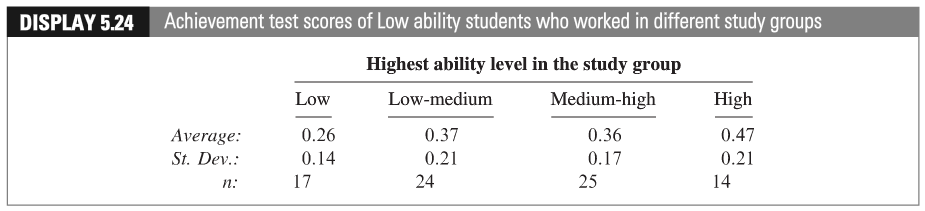
**We would certainly hope that a high school education is significant! But we shouldn’t let this idea drive our interpretation of the data.**

##In R:  
aovEduc <- aov(Income2005 ~ Educ, data = edu)  
edu.count <- table(edu$Educ)  
base\_12 <- contrMat(edu.count, base=3, type="Dunnett")  
dunnett.edu2 <- glht(aovEduc, linfct=mcp(Educ=base\_12))  
confint(dunnett.edu2)

##   
## Simultaneous Confidence Intervals  
##   
## Multiple Comparisons of Means: User-defined Contrasts  
##   
##   
## Fit: aov(formula = Income2005 ~ Educ, data = edu)  
##   
## Quantile = 2.4802  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate lwr upr   
## <12 - 12 == 0 -8563.4475 -18484.1671 1357.2720  
## >16 - 12 == 0 39990.5665 33421.1122 46560.0208  
## 13-15 - 12 == 0 8011.0607 2551.6693 13470.4521  
## 16 - 12 == 0 33132.0768 26754.8707 39509.2830

## Bonus (+5 points total)

Equity in Group Learning. [Continuation of Exercise 5.22.] (a) To see if the performance of low-ability students increases steadily with the ability of the best student in the group, form a linear contrast with increasing weights: -3 = Low, -1 = Low-Medium, +1 = Medium-High, and +3 = High. Estimate the contrast and construct a 95% confidence interval. (b) For the High-ability students, use multiple comparisons to determine which group composition differences are associated with different levels of test performance.

 (c) Give the levels of ability a quantitative representation (Low = 1, Low-Medium = 2, etc.). After completing the questions above, conduct a linear regression (we haven’t studied this yet!) of the AVERAGE performance against the level variable you just created. Be sure and address the assumptions. Defend the ones you can and assume the others are met. Include a scatterplot and residual plot. Is there evidence of linear trend? Is this inferred from the contrast? Assume the levels are equidistant in ability from each other.

Note: the data for Part b above is in Display 5.25 in your textbook.

### Part A (+2 points)

\*Bonus;  
data mycritval;  
cv = quantile(“t”, 1-(.05/2), 80-4);  
run;  
proc print data = mycritval;  
run;



**95% CI for the contrast is equal to**

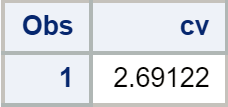
**The CI doesn’t include 0, which means low-ability students’ performances increase with the ability of the best students in the group (contrast is significantly different from 0).**

### Part B (+2 points)

**For the high ability students, , respectively ().**

**Bonferroni Method,**

*Bonferroni critical value;*  
*data mycritval2;*  
*cv = quantile(“t”, 1-(.05/(2*6)), 105-4);  
run;  
proc print data = mycritval2;  
run;



**95% CI centers for difference between study group means**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Average | Low | Low-Medium | Medium-High |
| High | 0.85 | 0.1 | 0.08 | 0.13 |
| Medium-High | 0.72 | -0.03 | -0.05 |  |
| Low-Medium | 0.77 | 0.02 |  |  |
| Low | 0.75 |  |  |  |

**95% CI for difference between study group means**

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Low | Low-Medium | Medium-High |
| High | [-0.0082, 0.208] | [-0.012, 0.172] | [0.052, 0.209] |
| Medium-High | [-0.133, 0.163] | [-0.135, 0.035] |  |
| Low-Medium | [-0.093, 0.133] |  |  |
| Low |  |  |  |

**Tukey-Kramer Method,**

\*\*Half-width of 95% CI:

**95% CI centers for difference between study group means**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Average | Low | Low-Medium | Medium-High |
| High | 0.85 | 0.1 | 0.08 | 0.13 |
| Medium-High | 0.72 | -0.03 | -0.05 |  |
| Low-Medium | 0.77 | 0.02 |  |  |
| Low | 0.75 |  |  |  |

**95% CI for difference between study group means**

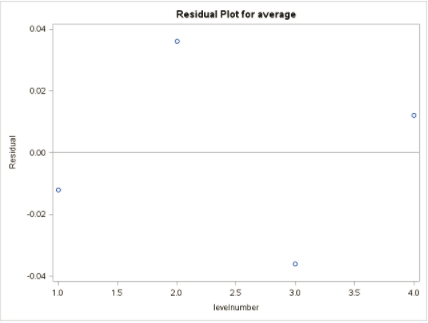
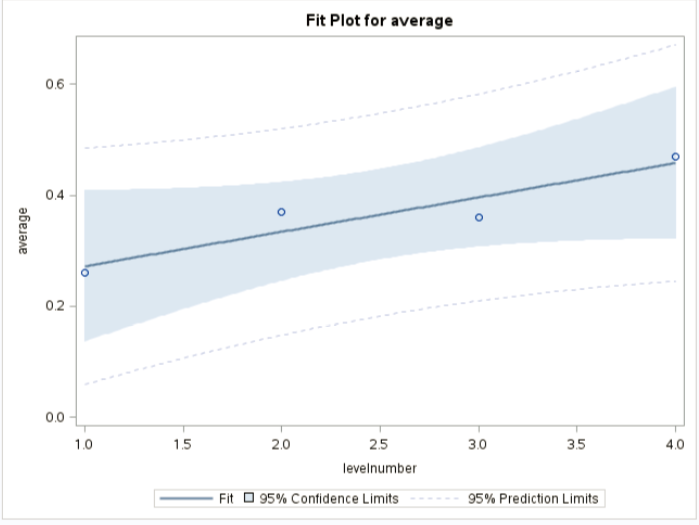
|  |  |  |  |
| --- | --- | --- | --- |
| Group | Low | Low-Medium | Medium-High |
| High | [-0.004, 0.204] | [-0.008, 0.168] | [0.055, 0.205] |
| Medium-High | [-0.128, 0.068] | [-0.131, 0.031] |  |
| Low-Medium | [-0.088, 0.128] |  |  |
| Low |  |  |  |

**It can be seen that the composition of Group High and Group Medium-High are associated with different levels of test performances.**

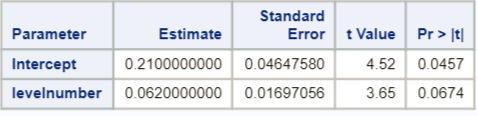
### Part C (+1 point)

\*To input average data;  
data studygroups;  
input level $ levelnumber average;  
datalines;  
Low 1 0.26  
Low-medium 2 0.37  
Medium-high 3 0.36  
High 4 0.47  
;  
run;

\*To perform regression on the averages (will also produce scatter plot);  
proc glm data=studygroups plots = all;  
model average = levelnumber;  
run;

**Assumptions: Since there’re too little data to check the assumptions, we granted that all the assumptions for simple linear regression are satisfied, including normality, independence, and constant variance across values of the independent variable.**



**The p-value of the independent variable is 0.0647, not significant at the level but significant at 0.1. So, there is likely a slight linear trend in the data.**