Contents

[Homework #1 Solutions 3](#_Toc507070097)

[Question 1 (20 points total) 3](#_Toc507070098)

[Question 2 (10 points total) 3](#_Toc507070099)

[Question 3 (32 points total) 4](#_Toc507070100)

[Part A 4](#_Toc507070101)

[Part B 4](#_Toc507070102)

[Part C 4](#_Toc507070103)

[Part D 5](#_Toc507070104)

[Question 4 (30 points total) 5](#_Toc507070105)

[Part A (10 points) 5](#_Toc507070106)

[Part B (5 points) 6](#_Toc507070107)

[Homework #2 Solutions 10](#_Toc507070108)

[Question 1 (15 points total) 10](#_Toc507070109)

[Part A (14 points total) 10](#_Toc507070110)

[Part B (1 point) 11](#_Toc507070111)

[Question 2 (40 points total) 11](#_Toc507070112)

[Part A (3 points for running the test, 1 point for the p-value, and 1 point for the correct conclusion) 11](#_Toc507070113)

[Part B (15 points total, scored exactly as Problem 1, part A) 13](#_Toc507070114)

[Part C (5 points total) 14](#_Toc507070115)

[Part D (5 points total) 15](#_Toc507070116)

[Part E (5 points total) 15](#_Toc507070117)

[Part F (5 points total) 15](#_Toc507070118)

[Question 3 (25 points total) 15](#_Toc507070119)

[Part A (15 points, score like the other hypothesis tests) 16](#_Toc507070120)

[Part B (10 points) 17](#_Toc507070121)

[Question 4 (15 points total) 17](#_Toc507070122)

[Part A (5 points) 17](#_Toc507070123)

[Part B (5 points) 18](#_Toc507070124)

[Part C (5 points) 19](#_Toc507070125)

[Unit 3 HW Solutions 21](#_Toc507070126)

[Homework #4 Solutions 62](#_Toc507070127)

[Question 1 (0 points total) 62](#_Toc507070128)

[Question 2 (20 points total) 62](#_Toc507070129)

[Part A (15 points total) 62](#_Toc507070130)

[Part B (5 points total) 64](#_Toc507070131)

[Question 3 (25 points total) 65](#_Toc507070132)

[Part A (5 points) 65](#_Toc507070133)

[Part B (5 points) 67](#_Toc507070134)

[Part C (5 points) 69](#_Toc507070135)

[Part D (5 points) 69](#_Toc507070136)

[Part E (5 points) 70](#_Toc507070137)

[Question 4 (25 points total) 70](#_Toc507070138)

[Part A (10 points) 70](#_Toc507070139)

[Part B (5 points) 71](#_Toc507070140)

[Part C (10 points) 73](#_Toc507070141)

[Question 5 (30 points total) 74](#_Toc507070142)

[Part A (5 points) 74](#_Toc507070143)

[Part B (5 points) 75](#_Toc507070144)

[Part C (5 points) 76](#_Toc507070145)

[Part D (5 points) 77](#_Toc507070146)

[Part E (5 points) 78](#_Toc507070147)

[Part F (5 points) 79](#_Toc507070148)

[Bonus (+5 points total) 79](#_Toc507070149)

[Part A (+2.5 points) 79](#_Toc507070150)

[Part B (+2.5 points) 80](#_Toc507070151)

[Homework #5 Solutions 81](#_Toc507070152)

[Question 1 (35 points total) 81](#_Toc507070153)

[Question 2 (30 points total) 87](#_Toc507070154)

[Question 3 (30 points total) 90](#_Toc507070155)

[Unit 6 HW Solutions 98](#_Toc507070156)

[Question 1 (25 points total) 98](#_Toc507070157)

[Question 2 (25 points total) 98](#_Toc507070158)

[Question 3 (50 points total) 103](#_Toc507070159)

[Part A (15 points) 109](#_Toc507070160)

[Part B (15 points) 114](#_Toc507070161)

[Bonus (+5 points total) 121](#_Toc507070162)

[Part A (+2 points) 121](#_Toc507070163)

[Part B (+2 points) 122](#_Toc507070164)

[Part C (+1 point) 124](#_Toc507070165)

# Homework #1 Solutions

***Note: You earned 8 points just for turning in the assignment!***

## Causality (20 points total)

What is the difference between a randomized experiment and a random sample? Under what type of study/sample can a causal inference be made?

**(10 points) In a randomized experiment, the groups in the study are allocated via a chance mechanism. In a random sample, the units/subjects in a study are included from some population via a chance mechanism.**

**(10 points) Causal inferences can be made only with a randomized experiment.**

*Note: your solution does not need to be word-for-word with the above. For example, for the first question you could have phrased the first part as “in a randomized experiment, subjects in the study are randomly assigned to groups.”*

## Inference (10 points total)

In 1936, the Literary Digest polled 1 out of every 4 Americans and concluded that Alfred Landon would win the presidential election in a landon-slide. Of course, history turned out dramatically different (see [link](http://historymatters.gmu.edu/d/5168/) for further details). The magazine combined three sampling sources: subscribers to its magazine, phone number records, and automobile registration records. Comment on the desired population of interest of the survey and what population the magazine actually drew from.

**(10 points) The population of interest for this study was intended to be all the people who were going to vote in the 1936 election (not even necessarily all people in the country as many people don’t vote). The population that the study actually referenced were more affluent people (based on appealing to those with magazine subscriptions and cars during one of the lowest points in the great depression). As FDR’s “New Deal” and other social programs were much more attractive to lower income people, they were not included in anything resembling representative numbers.**

*Note: To get full credit, you must correctly identify the desired population AND what they actually ended up with. If you miss one, around 5 points will be subtracted. There is no one correct answer for the discussion - the one given is an example.*

## Causality and Inference (32 points total)

Suppose we have developed a new fertilizer that is supposed to help corn yields. This fertilizer is so potent that a small vial of it sprayed over an entire field is a sufficient dose. We find that the new fertilizer results in an average yield of 60 more bushels over the old fertilizer with a p-value of 0.0001. Write up a scope of inference under the following study designs that generated this data.

*Note: To get full credit, you must correctly discuss whether the results can be extended to any broader population (4 points) and whether causal inferences can be made (4 points). This makes each sub-question worth a total of 8 points.*

### Part A

We offer the new fertilizer at a discount to customers who have purchased the old fertilizer along with a survey for them to fill out. Some farmers send in a survey after the growing season, reporting their crop yield. From our records, we know which of these farmers used the new fertilizer and which used the old one.

**As the farmers self-select into the study, extending the results to any other group or population (such as all the customers of the company) is speculative. Also, we do not get to randomize the assignment to “old” and “new” fertilizer, so no causal inference can be made.**

### Part B

When a customer makes an order, we randomly send them either the old or new fertilizer. At the end of the season, some of the farmers send us a report of their yield. Again, from our records, we know which of these farmers used the new fertilizer and which used the old.

**As the farmers self-select into the study, extending the results to any other group or population (such as all the customers of the company) is speculative. This is a randomized experiment as we randomly assign the farmers to the two fertilizers. Hence, causal inferences can be made - but again, only to the particular farmers that sent us their yields. However, because only some of the farmers send their yields (after treatment has been applied), this could introduce a confounding variable into the study via nonresponse bias (a form of selection bias). This could be a cause for concern for the validity of the entire study.**

### Part C

When a customer makes an order, we randomly send them either the old or new fertilizer. At the end of the season, we sub-select from the fertilizer orders and send a team out to count those farmers’ crop yields.

**Due to us randomly sampling from all of the fertilizer orders, we can extend our results to all farmers that ordered fertilizers this year. This is a randomized experiment as we randomly assigned the farmers to the two fertilizers. Hence, causal inferences can be made. However, because only some of the farmers send their yields (after treatment has been applied), this could introduce a confounding variable into the study via nonresponse bias (a form of selection bias). This could be a cause for concern for the validity of the entire study.**

### Part D

We offer the new fertilizer at a discount to customers who have purchased the old fertilizer. At the end of the season, we sub-select from the fertilizer orders and send a team out to count those farmer’s crop yields. From our records, we know which of these farmers used the new fertilizer and which used the old one.

**Due to us randomly sampling from all of the fertilizer orders, we can extend our results to all farmers that ordered fertilizers this year. We did not get to randomize the assignment to “old” and “new” fertilizer, though, so no causal inference can be made.**

## 2 Sample Univariate (30 points total)

We polled a Business Stats class here at SMU and asked them how much money (cash) they had in their pocket at that very moment. The idea was that we wanted to see if there was evidence that those in charge of the vending machines should include the expensive bill/coin acceptor or if it should just have the credit card reader. We asked a professor from Seattle University last year to poll her class with the same question. Below are the results of our polls.

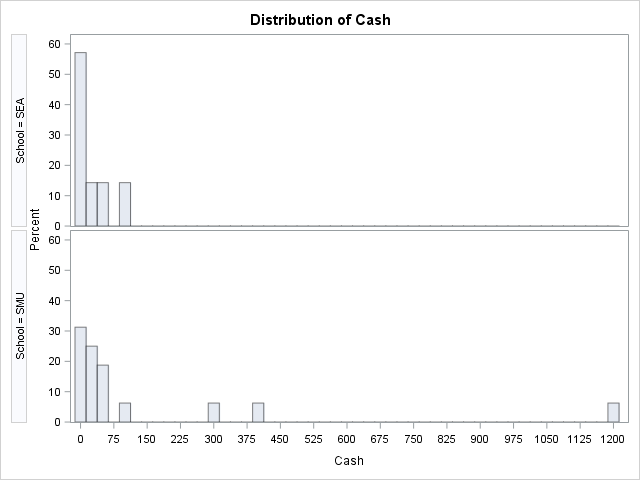
**SMU** > 34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0 **Seattle U** > 20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

### Part A (10 points)

Use SAS to make a histogram of the amount of money in a student’s pocket from each school. Does it appear there is any difference in population means? What evidence do you have? Discuss your thoughts.

*There are multiple ways to get these histograms in SAS. If you used a different method, that is fine… provided that the shapes are the same. As long as the plot is correct, you should receive nearly full credit for the discussion.*

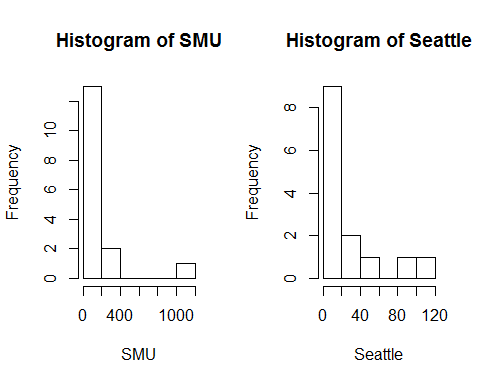
My SAS code:  
proc univariate data=problem4;  
class school;  
histogram class;  
run;



**It is tough to tell given the scale of the histograms. There may be some evidence of a difference in population means. We conduct a formal permutation test next to test for a difference in means.**

### Part B (5 points)

SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)  
Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)  
par(mfrow=c(1,2))  
hist(SMU)  
hist(Seattle)



###Part C (15 points)

Run a permutation test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a statistical conclusion and scope of inference (similar to the one from the powerpoint; this should include identifying the Ho and Ha as well as the pvalue).

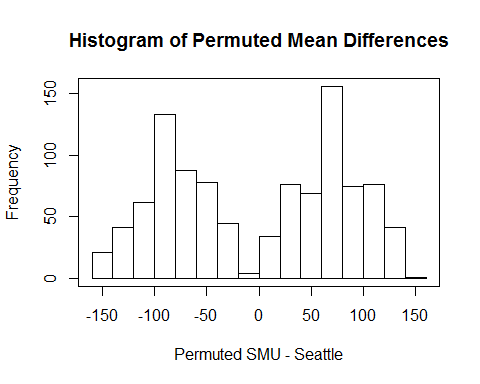
*Note: Your answers will vary. As long as you’ve used code similar to that below, your answer should be within a few decimal points.*

*Note: The point breakdown is as follows: 5 points for running the test, 5 points for getting the correct p-value (which, in all likelihood, should be between 0.1 and 0.2), and 5 points for writing up a conclusion that includes stating the null and alternative hypothesis, as well as drawing the correct conclusion (i.e., there is not sufficient evidence that a difference exists).*

school1 <- rep('SMU', 16)  
school2 <- rep('Seattle', 14)  
school <- as.factor(c(school1, school2))  
all.money <- data.frame(name=school, money=c(SMU, Seattle))  
  
t.test(money ~ name, data=all.money)

##   
## Welch Two Sample t-test  
##   
## data: money by name  
## t = -1.4945, df = 15.499, p-value = 0.1551  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -277.64481 48.39481  
## sample estimates:  
## mean in group Seattle mean in group SMU   
## 27.000 141.625

number\_of\_permutations <- 1000  
xbarholder <- numeric(0)  
counter <- 0  
observed\_diff <- mean(subset(all.money, name == "SMU")$money)-mean(subset(all.money, name == "Seattle")$money)  
  
set.seed(123)  
for(i in 1:number\_of\_permutations)  
{  
scramble <- sample(all.money$money, 30)  
smu <- scramble[1:16]  
seattle <- scramble[17:30]  
diff <- mean(smu)-mean(seattle)  
xbarholder[i] <- diff  
if(abs(diff) > observed\_diff)  
counter <- counter + 1  
}  
hist(xbarholder, xlab='Permuted SMU - Seattle', main='Histogram of Permuted Mean Differences')  
box()



pvalue <- counter / number\_of\_permutations  
pvalue

## [1] 0.135

**To test for a difference of population means between the SMU and Seattle groups, a permutation test was conducted on 1,000 random permutations of the data. A histogram of the 1,000 differences of sample means from the 1,000 permutation can be seen above. The observed difference was $114, in which 135 of the 1000 permutations yielded a difference in sample means that was as extreme or more extreme than this observed difference (p-value = 135/1000 = 0.135). This does not provide sufficient evidence against the null hypothesis that the mean pocket cash of SMU students is equal to that of Seattle University students. These 30 students were not from a random sample; therefore, inference cannot be drawn beyond the 30 subjects in the sample. (Since we failed to reject Ho, there is no need to write up whether casual inference can be drawn).**

# Homework #2 Solutions

***Note: You get 5 points just for turning in the assignment!***

## Question 1 (15 points total)

The world’s smallest mammal is the bumblebee bat, also known as the Kitti’s hog nosed bat. Such bats are roughly the size of a large bumblebee! Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8g. *Beware: This data is not the same as in the lecture slides!*

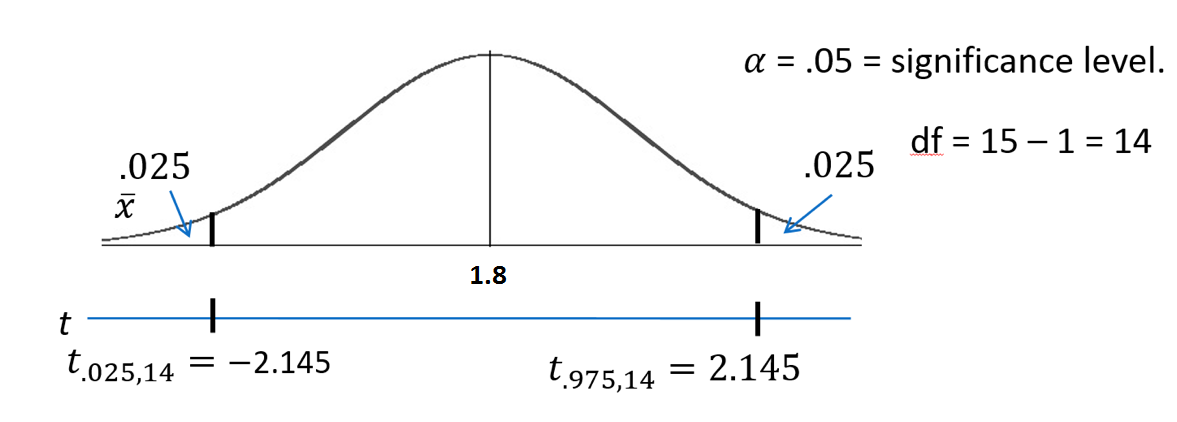
Sample: 1.7, 1.6, 1.5, 2.0, 2.3, 1.6, 1.6, 1.8, 1.5, 1.7, 1.2, 1.4, 1.6, 1.6, 1.6

### Part A (14 points total)

Perform a complete analysis using SAS: the six step hypothesis test with a conclusion that includes a statistical conclusion, a confidence interval and a scope of inference (as best as can be done with the information above. there are many correct answers with the vagueness of the description of the sampling mechanism).

**Step 1 - Hypotheses (2 points):**

**Step 2 - Identification of Critical Value (1 point for drawing, 1 point for value):**

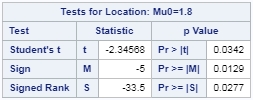


**Step 3 - Value of Test Statistic (2 points):**

**Step 4 - Give p-value (2 points):**

**Step 5 - Decision (2 points): Reject**

**Step 6 - Conclusion (3 points for the statistical conclusion, 1 point for the confidence interval, 1 point for discussing the scope): On the basis of this test, there is enough evidence to reject the claim that the mean weight of bumblebee bats is equal to 1.8g ( from a two sided t-test). A 95% confidence interval is [1.5065, 1.7869] grams. The problem was ambiguous on the randomness of the sample, thus we will assume that it was not a random sample which makes inference to all bats strictly speculative.**

**SAS output**  


### Part B (1 point)

Inspect and run this R Code and compare the results (t statistic, p-value and confidence interval) to those you found in SAS. To run the code simply copy and paste the below code into R.

sample = c(1.7, 1.6, 1.5, 2.0, 2.3, 1.6, 1.6, 1.8, 1.5, 1.7, 1.2, 1.4, 1.6, 1.6, 1.6)  
t.test(x=sample, mu = 1.8, conf.int = "TRUE", alternative = "two.sided" )

**You should observe exactly the same results.**

## Question 2 (40 points total)

In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, let’s suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below:

**Fired**  
> 34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56  
**Not Fired**  
> 27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54

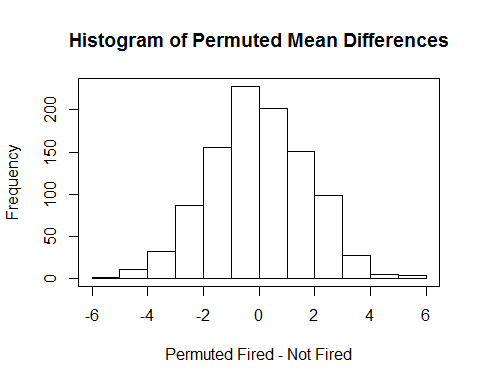
### Part A (3 points for running the test, 1 point for the p-value, and 1 point for the correct conclusion)

Perform a permutation test to test the claim that there is age discrimination. Provide the and , the p-value, and full statistical conclusion including the scope. Note: this was an example in Live Session 1. You may start from scratch or use the sample code and PowerPoint from Live Session 1.

fired <- c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)  
not.fired <- c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)  
label1 <- rep('fired', 21)  
label2 <- rep('not.fired', 30)  
label.all <- as.factor(c(label1, label2))  
samoa <- data.frame(status=label.all, age=c(fired, not.fired))  
  
t.test(age ~ status, data=samoa)

##   
## Welch Two Sample t-test  
##   
## data: age by status  
## t = 1.079, df = 40.268, p-value = 0.287  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.678993 5.526612  
## sample estimates:  
## mean in group fired mean in group not.fired   
## 45.85714 43.93333

number\_of\_permutations <- 1000  
xbarholder <- numeric(0)  
counter <- 0  
observed\_diff <- mean(subset(samoa, status == "fired")$age)-mean(subset(samoa, status == "not.fired")$age)  
  
set.seed(123)  
for(i in 1:number\_of\_permutations)  
{  
scramble <- sample(samoa$age, 51)  
fired.new <- scramble[1:21]  
not.fired.new <- scramble[22:51]  
diff <- mean(fired.new)-mean(not.fired.new)  
xbarholder[i] <- diff  
if(abs(diff) > observed\_diff)  
counter <- counter + 1  
}  
hist(xbarholder, xlab='Permuted Fired - Not Fired', main='Histogram of Permuted Mean Differences')  
box()



pvalue <- counter / number\_of\_permutations  
pvalue

## [1] 0.269

**There is not sufficient evidence to suggest that the mean score of those who were fired is greater than the mean age of those who were not fired (). Since this was a random sample of employees in Samoa, we can generalize the inference to all employed people in Samoa. Since we failed to reject the null hypothesis, we do not need to discuss whether causal conclusions can be drawn.**

*Note: Remember, your p-value will not be exactly the same. However, it should be very close (likely within 0.05).*

### Part B (15 points total, scored exactly as Problem 1, part A)

Now run a two sample t-test appropriate for this scientific problem (use SAS) (Note: we may not have talked much about a two-sided versus a one-sided test. If you would like to read the discussion on p.44 (Statistical Sleuth), you can run a one-sided test if it seems appropriate. Otherwise, just run a two-sided test as in class. There are also example in the Statistics Bridge Course). Be sure to include all six steps, a statistical conclusion and scope of inference.

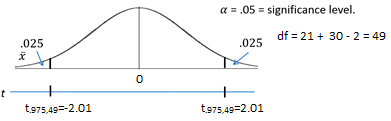
*Note: the solutions will look slightly different if you do a one-sided test. Either option can still receive full credit, though.*

*Note: If your data is sorted in the opposite order, you will get all the same numbers but with opposite signs (for example, -1.10 instead of 1.10). This is totally fine and your answers should match up, including p-values. If your signs are opposite and your p-value doesn’t match up, it means you are reporting the wrong shaded area. In this case, if you subtract your p-value from 1 then it should match with the p-value given in the solution.*

**Step 1 - Hypotheses (2 points):**

**or**

**Step 2 - Identification of Critical Value (1 point for drawing, 1 point for value): (2-sided) or (1-sided)**



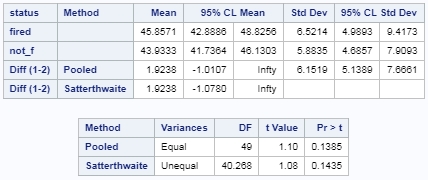
**Step 3 - Value of Test Statistic (2 points):**

**Step 4 - Give p-value (2 points): (2-sided) or (1-sided)**

**Step 5 - Decision (2 points): Fail to Reject**

**Step 6 - Conclusion (3 points for the statistical conclusion, 1 point for the confidence interval, 1 point for discussing the scope): On the basis of this test, there is not enough evidence to suggest that the mean ages of the fired and not fired groups are different. In other words, there is not enough evidence to suggest that there is discrimination based on age ( from a two-sample t-test). A 95% confidence interval for this difference is years. Since the subjects in this sample were randomly sampled, inference can be generalized to the population of all employees in the American Samoa Government.**

*Note: If you did a 1-sided test, the 95% CI is .*

**SAS output**  


### Part C (5 points total)

Compare this p-value to the randomized p-value found in the previous problem.

**P-values will vary from test to test (and if you did a 1-sided test your p-value will be about half of the one from the previous problem). The p-value from the solution to the last problem was 0.269 which is very close to the one that is provided by t-test. These p-values are close since the distribution of sample means from the permutation test is approximately normal.**

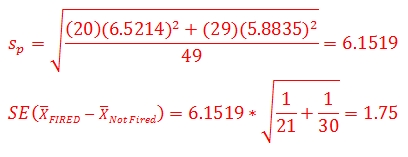
### Part D (5 points total)

The jury wants to see a range of plausible values for the difference in means between the fired and not fired groups. Provide them with a confidence interval for the difference of means and an interpretation.

**A 95% confidence interval for this difference is years (or the 1-sided interval reported above). With 95% confidence, we expect the true mean difference in ages to fall in this interval. Because 0 is included in the interval and is thus a probable value, we don’t have enough evidence to conclude a difference exists.**

### Part E (5 points total)

Given the sample standard deviations from SAS, calculate by hand:  
i. (2.5 points) Pooled Standard Deviation   
ii. (2.5 points) Standard Error of



### Part F (5 points total)

Inspect and run this R Code and compare the results (t statistic, p-value and confidence interval) to those you found in SAS. To run the code simply copy and paste the below code into R.

Fired = c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)  
Not\_fired = c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)  
t.test(x = Fired, y = Not\_fired, conf.int = .95, var.equal = TRUE, alternative = "two.sided")

**You should observe exactly the same results. No real discussion is needed aside from making this observation!**

## Question 3 (25 points total)

In the last homework I mentioned I polled my Business Stats class here at SMU and asked them how much money (cash) they had in their pocket at that very moment. The idea was that we wanted to see if there was evidence that those in charge of the vending machines should include the expensive bill/coin acceptor or if it should just have the credit card reader. However, I met a professor from Seattle University last year and asked her to poll her class with the same question. Below are the results of our polls.

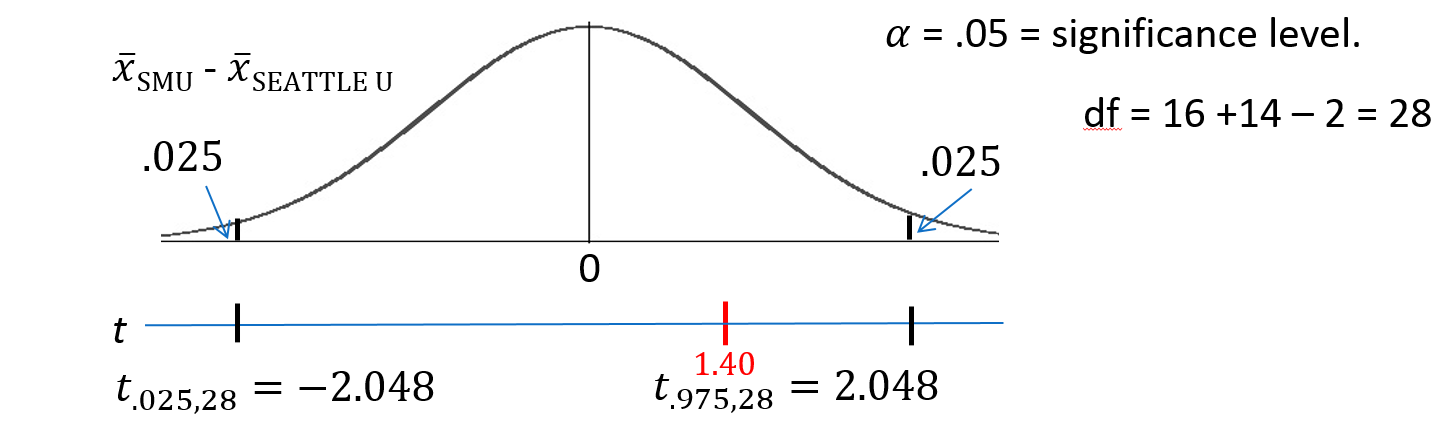
**SMU**  
> 34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0  
**Seattle U**  
> 20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

### Part A (15 points, score like the other hypothesis tests)

Run a two sample t-test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a complete analysis: all 6 steps, including a statistical conclusion and scope of inference (similar to the one from the powerpoint). This should include identifying the and as well as the p-value. Also, include the appropriate confidence interval. FUTURE DATA SCIENTIST’S CHOICE!: YOU MAY USE SAS OR R TO DO THIS PROBLEM!

**Step 1 - Hypotheses (2 points):**

**Step 2 - Identification of Critical Value (1 point for drawing, 1 point for value):**



**Step 3 - Value of Test Statistic (2 points):**

**Step 4 - Give p-value (2 points):**

**Step 5 - Decision (2 points): Fail to Reject**

**Step 6 - Conclusion (3 points for the statistical conclusion, 1 point for the confidence interval, 1 point for discussing the scope): On the basis of this test, there is not enough evidence to suggest that the mean amount of pocket cash of the SMU students is different than that of the students from Seattle U ( from a two-sided t-test). A 95% confidence interval for this difference is [-$53, $282]. Since the subjects in this sample were not randomly sampled, the results only generalize to the subjects in the study (no need to discuss causal conclusions for a non-significant result).**

SMU = c(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)  
Seattle = c(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)  
school1 <- rep('SMU', 16)  
school2 <- rep('Seattle', 14)  
school <- as.factor(c(school1, school2))  
all.money <- data.frame(name=school, money=c(SMU, Seattle))  
  
t.test(money ~ name, data=all.money, var.equal=T, alternative='two.sided')

##   
## Two Sample t-test  
##   
## data: money by name  
## t = -1.3976, df = 28, p-value = 0.1732  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -282.62112 53.37112  
## sample estimates:  
## mean in group Seattle mean in group SMU   
## 27.000 141.625

*Note: I have reported the solution using R, but SAS is totally fine as well!*

### Part B (10 points)

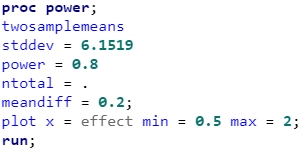
**The p-value from the two sample t-test was 0.1732, while the pvalue from the permutation test last week was somewhere near 0.15 (answers will vary due to randomness). Recall from last week that the distribution of the permuted differences of sample means was very non-normal (if you need to refresh your memory, refer to the histogram in the Unit 1 solutions). This shows that the normal distribution (or one that is approximated by a t distribution) is not always a good approximation of the distribution of a statistic. Although the normal distribution is obviously a poor fit for this distribution, it just so happens that it does a good job of approximating the tails of the bimodal permuted distribution. The main point here is that the outlier in this data had a drastic effect on the shape of the distribution of sample means, thus disqualifying the normal distribution from being a good approximation of the distribuiotn in order to approximate p-values. The p-values for the t-test and permutation test are close here, but this cannot be depended on to be consistent.**

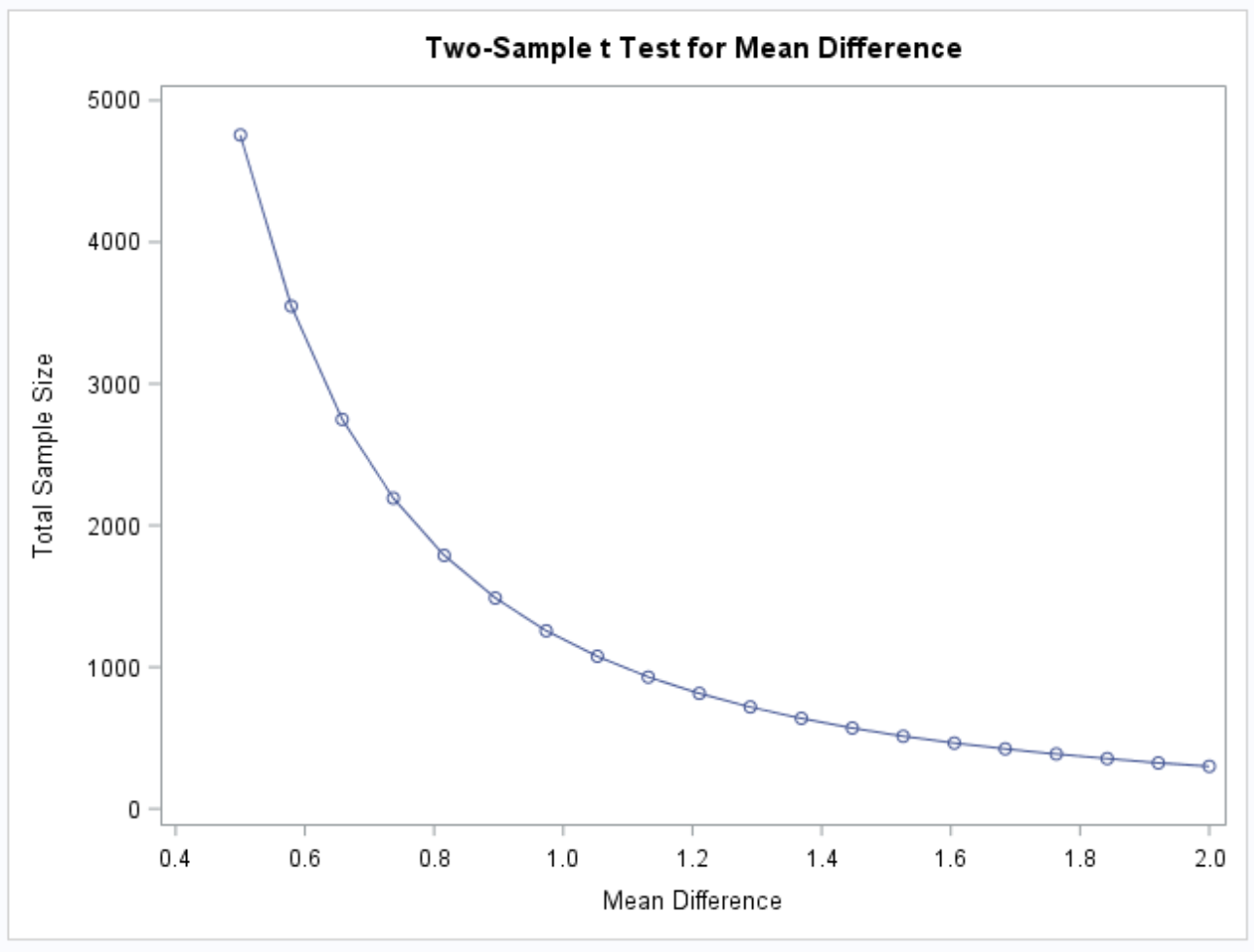
*Note: It should be noted that the p-values are close, but pay careful attention to the discussion above. The critical piece to mention is that the outlier has a drastic effect on the outcomes.* ***If you do not discuss the outlier in any capacity, 2 points are taken off.***

## Question 4 (15 points total)

### Part A (5 points)

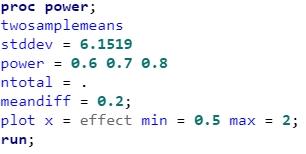
Calculate the estimate of the pooled standard deviation from the Samoan discrimination problem. Use this estimate to build a power curve. Assume we would like to be able to detect effect sizes between 0.5 and 2 and we would like to calculate the sample size required to have a test that has a power of .8. Simply cut and paste your power curve and SAS code. HINT: USE THE CODE FROM DR. McGEE’s lecture. Instead of using groupstddevs, use stddev since we are using the pooled estimate.

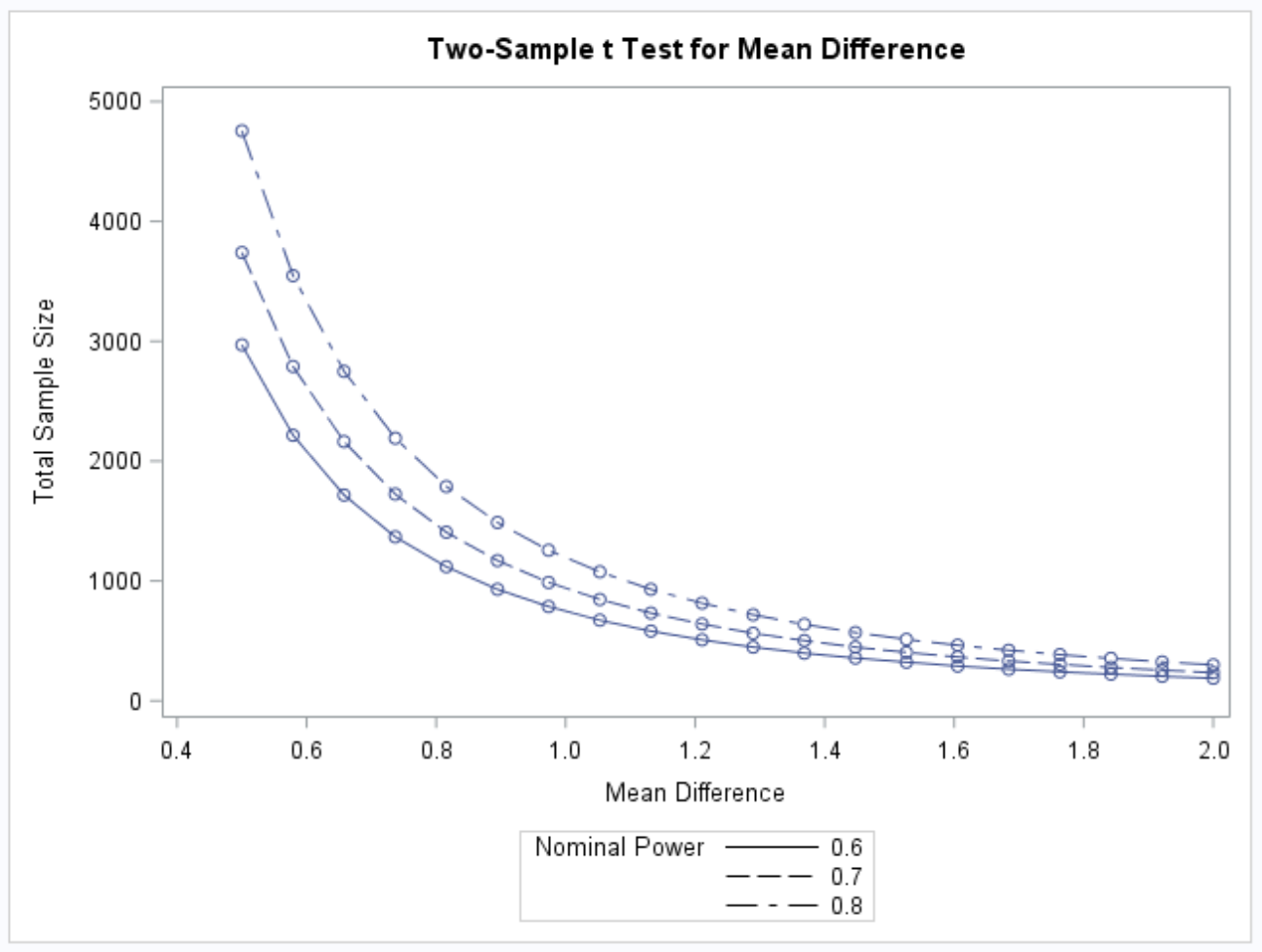
**SAS Code**  


**SAS Output**  


### Part B (5 points)

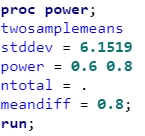
Now let’s say that we decided that we may be able to live with slightly less power if it means savings in sample size. Provide the same plot as above but this time calculate curves of sample size (yaxis) vs effect size (.5 to 2) (x axis) for power = 0.8, 0.7 and 0.6. There should be three plots on your final plot. Simply cut and paste your power curve and SAS code. HINT: USE THE CODE FROM DR. McGEE’s lecture… instead of groupstddevs, use stddev since we are using the pooled estimate.

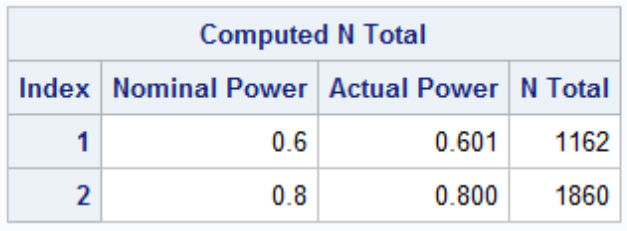
**SAS Code**  


**SAS Output**  


### Part C (5 points)

Using your last plot, estimate the savings in sample size from a test aimed at detecting an effect size of 0.8 with a power of 80% versus a power of 60%.

**SAS Code**  


**SAS Output**  


**Savings: 1860 - 1162 = 698**

*Note: If you only reported the output and not the savings, you lose one point.*

# Unit 3 HW Solutions

**Question 1 (30 points total)**

In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, let’s suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below:

**Fired**

* 34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56

**Not Fired**

* 27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54

**Part A (10 points total)**

Check the assumptions (with SAS) of the two-sample t-test with respect to this data. Address each assumption individually as we did in the videos and live session and make sure and copy and paste the histograms, q-q plots or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?

SAS code for histograms, q-q plots, and box plots as well as the t-test below

\*/Assumes a data file named samoa with variables ForNF and age has been uploaded,

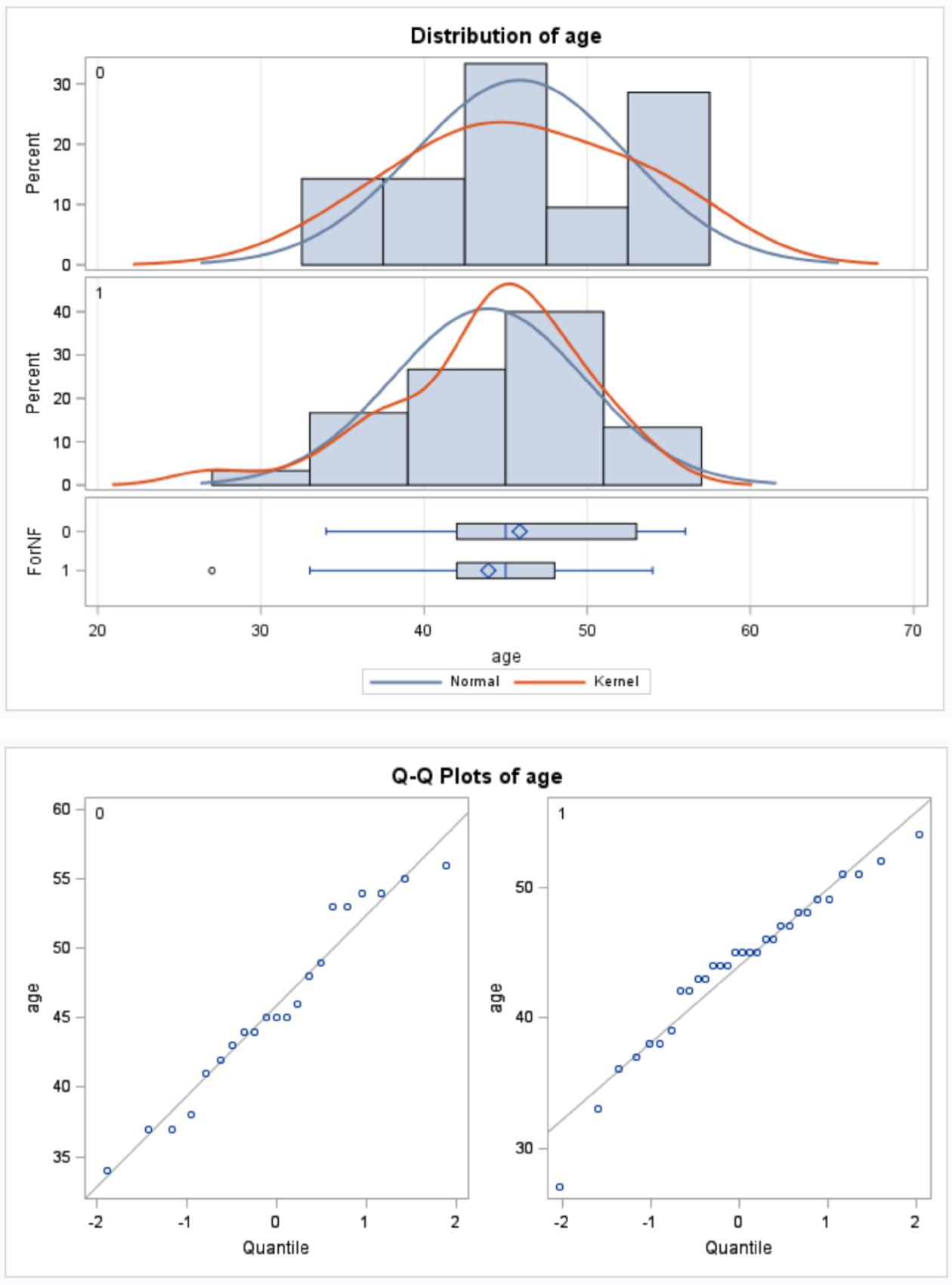
where 1 is Fired and 0 is Not Fired;

proc ttest data = samoa;

class ForNF;

var age;

run;



**(3 points) Normality: There is little if any evidence from the histograms and QQ plots of any departures from normality from these data. We will assume they do come from normal distributions.**

**(3 points) Equal standard deviations: There is little evidence that the samples are pulled from distributions that have different standard deviations, thus we will assume that the standard deviations are equal.**

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

**(1 point) Decision: the two-sample t-test and confidence intervals are appropriate to use for these data.**

*Note: Remember, your graphs do not have to look exactly the same as the answer key. There are multiple ways to create histograms in SAS - just ensure the same information is being displayed.*

**Part B (10 points)**

Check the assumptions with R and compare them with the plots from SAS.

fired <- **c**(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)

not.fired <- **c**(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47 , 48, 48, 49, 49, 51, 51, 52, 54) label1 <- **rep**('fired', 21)

label2 <- **rep**('not.fired', 30)

label.all <- **as.factor**(**c**(label1, label2))

samoa <- **data.frame**(status=label.all, age=**c**(fired, not.fired))

**par**(mfrow=**c**(2,2))

**hist**(fired,xlab='Age of Fired Employees',main='Fired Group')

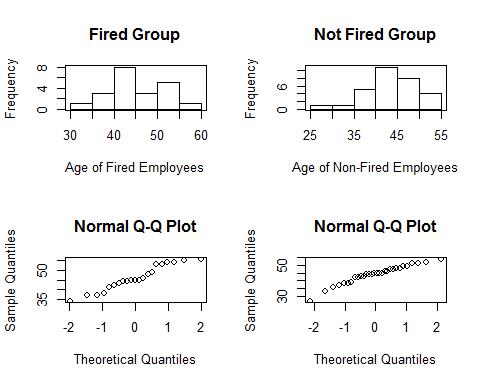
**box**()

**hist**(not.fired,xlab='Age of Non-Fired Employees',main='Not Fired Group')

**box**()

**qqnorm**(fired)

**qqnorm**(not.fired)



**You should observe exactly the same results, aside from slight formatting differences between SAS and R. No real discussion is needed aside from reaching the same conclusion as before!**

**Part C (10 points)**

Now perform a complete analysis of the data. You may use either the permutation test from HW 1 or the t-test from HW 2 (copy and paste) depending on your answer to part a. In your analysis, be sure and cover all the steps of a complete analysis:

1. State the problem.
2. Address the assumptions of t-test (from part A)
3. Perform the t-test if it is appropriate and a permutation test if it is not. (Judging from your analysis of the assumptions.
4. Provide a conclusion including the p-value and a confidence interval.
5. Provide the scope of inference.

(Steps 3-5 are from your previous HW; you are just putting everything together.)

NOTE: THIS QUESTION SHOULD BE EASY AS YOU ARE SIMPLY FORMATTING YOUR RESULTS FROM EARLIER IN THE ABOVE FORM. (Steps 3 -5 are from your previous HW; you are just putting everything together.) IT REALLY JUST EQUATES TO ADDING A STATEMENT OF THE PROBLEM AND ADDRESSING THE ASSUMPTIONS (1 and 2 above). You can basically copy and paste the rest. We are simply putting everything together to make a complete report.

**Problem (1 point): We wish to test the claim that the mean age of the fired group is different than the mean age of the not fired group for this population of American Samoa working citizens. In other words, we are testing the claim that age discrimination exists for this population. (Note that a one-sided test addresses the problem more directly, as age discrimination typically refers to preferential treatment of younger workers. However, both a one-sided test and two-sided test are acceptable for this assignment.)**

**The assumptions (1 point) are as stated in parts A and B. Because the assumptions for the t-tools are met, we will proceed with a t-test.**

**What follows is a repeat from HW #2, but you can see how everything flows together in a complete analysis. If you chose the permutation test (with an approximate p-value or exact p-value), then see the homework solutions from the permutation test.**

**Step 1 - Hypotheses (1 point):**

: =

: ≠

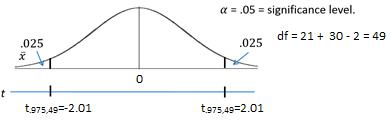
**or**

: ≤

: >

Note that the latter set of hypotheses more closely aligns with our question of interest, as age discrimination typically focuses on older workers being treated less favorably.

**Step 2 - Identification of Critical Value (1 point):** ±2.01 **(2-sided) or** 1.677 **(1-sided)**

****

This graph is for a two-sided test.

**Step 3 - Value of Test Statistic (1 point):** = 1.10

**Step 4 - Give p-value (1 point):** = 0.2771 **(2-sided) or** = 0.1385 **(1-sided)**

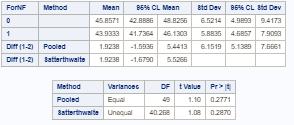
**Step 5 - Decision (1 point): Fail to Reject** at significance level alpha = 0.05.

**Step 6 - Conclusion (1 point for the statistical conclusion, 1 point for the confidence interval, 1 point for discussing the scope): On the basis of this test, there is not enough evidence to suggest that the mean ages of the fired and not fired groups are different. In other words, there is not enough evidence to suggest that there is discrimination based on age (** = 0.2771 **from a two-sample t-test). A 95% confidence interval for this difference is** [−1.60,5.44] **years. Since the subjects in this sample were randomly sampled, inference can be generalized to the population of all employees in the American Samoa Government.**

*Note: if you did a 1-sided test and provided a 2-sided 90% CI, this interval should be [-1.01, 4.86]. The 1-sided 95% CI is [1.01, inf], although when we consider that a confidence interval should be the set of all plausible values for the difference in means, the interpretive value of this CI is not as strong as a 2-sided 90% CI.*

SAS Code:

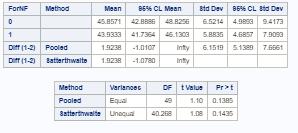
\*/For a two-sided test at alpha = 0.05; proc ttest data = samoa; class ForNF;



var age;

run;

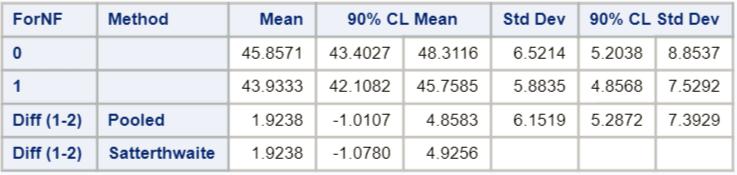
\*/For a one-sided test at alpha = 0.05; proc ttest data = samoa sides = u; class ForNF;



var age;

run;

\*/For a 90% two-sided confidence interval to match the 0.05 one-sided hypothesis test; proc ttest data = samoa sides = u alpha = 0.1;



class ForNF;

var age;

run;

\*/Critical value for a two-sided test at alpha



* 0.05; data critval;

cv = quantile("T", .975, 49);

proc print data = critval;

run;

\*/Critical value for a one-sided test at alpha



* 0.05; data critval;

cv = quantile("T", .95, 49);

proc print data = critval;

run;

Code for the permutation test can be found

in prior assignments.

*Note: Perhaps you might be wondering at this point in the HW, “Why are we always testing the assumptions of the t-test? Is it the best test? Should we always run the t-test when we can?” These are very good questions and open questions that are up for debate! The one thing that is mathematically proven and not up for debate is that if the assumptions are met, the two-sample t-test is the most powerful (in terms of Power = 1 - beta) test in the universe at testing the claim of the difference of means. Two questions may arise here. 1. Do we every really have the assumptions fully met in the real world and just how much power do we give up at varying degrees of violation of the assumptions? 2. Do we always want inference on the equality/difference of means? We will continue to answer these questions in Chapter 4. (Also note that we started to answer number two with a t-test of log transformed data. The inference there is on the equality (ratio) of medians, which may be a better measure of center when dealing with right or left skewed data!*

**Question 2 (30 points total)**

In the last homework, it was mentioned that a Business Stats class here at SMU was polled and students were asked how much money (cash) they had in their pockets at that very moment. The idea was to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin acceptor or if they should just have the credit card reader. However, a professor from Seattle University polled her class with the same question. Below are the results of the polls.

**SMU**

* 34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0

**Seattle U**

* 20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

**Part A (9 points total)**

Check the assumptions (with SAS or R) of the two-sample t-test with respect to this data. Address each assumption individually as we did in the videos and live session and make

sure to copy and paste the histograms, q-q plots, or any other graphic you use (boxplots, etc.) to defend your written explanation. Do you feel that the t-test is appropriate?

SMU = **c**(34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20 , 10, 0)

Seattle = **c**(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)

school1 <- **rep**('SMU', 16)

school2 <- **rep**('Seattle', 14)

school <- **as.factor**(**c**(school1, school2))

all.money <- **data.frame**(name=school, money=**c**(SMU, Seattle))

**par**(mfrow=**c**(2,2))

**hist**(SMU,xlab='SMU Pocket Cash',main='SMU')

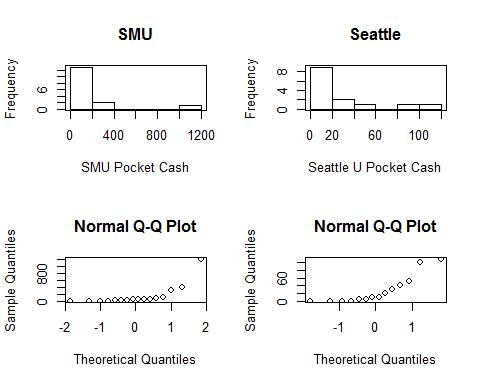
**box**()

**hist**(Seattle,xlab='Seattle U Pocket Cash',main='Seattle')

**box**()

**qqnorm**(SMU)

**qqnorm**(Seattle)



**(3 points) Normality: There is significant evidence from the histograms and q-q plots of severe departures from normality for these data. We will assume they do not come from normal distributions. In addition, the sample size does not look adequate to make the t-test robust to this assumption.**

**(3 points) Equal standard deviations: There is significant evidence to suggest that the standard deviations of these distributions are different.**

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

The assumptions do not appear to be met; we will NOT proceed with the t-tools.

Assumptions check in SAS:

\*To address assumptions of the t-test with histograms and q-q plots;

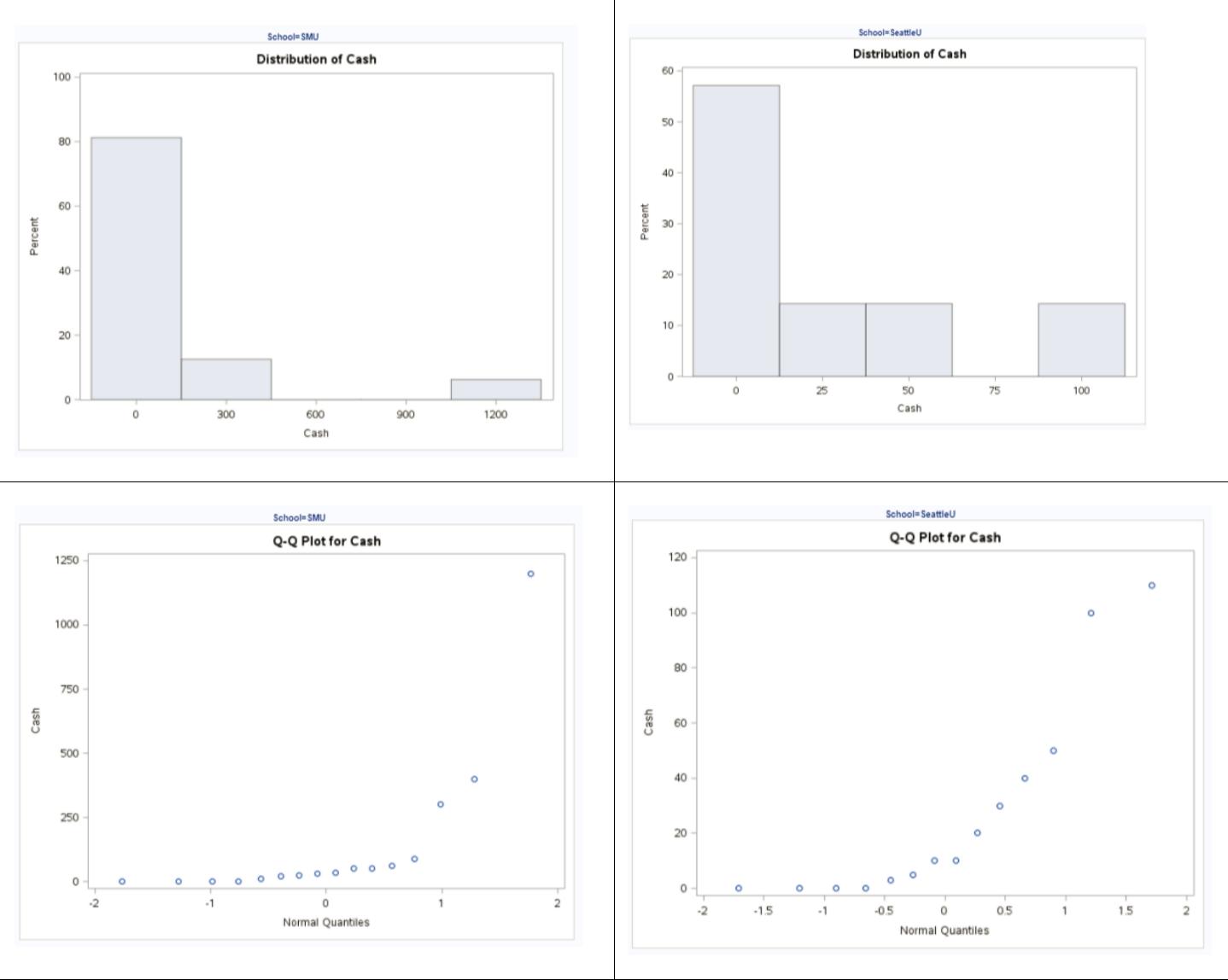
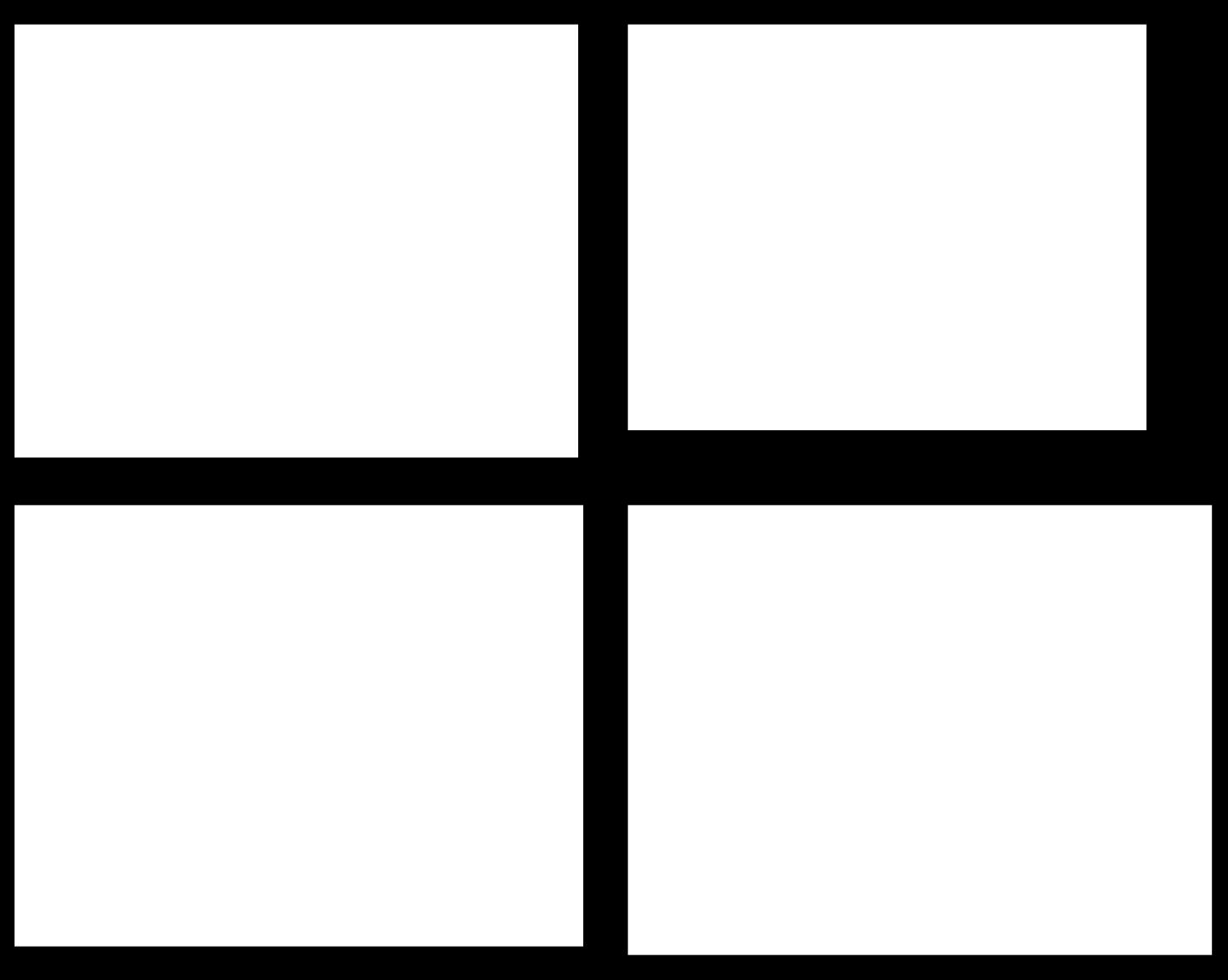
proc univariate data = schoolcash;

by school;

histogram cash;

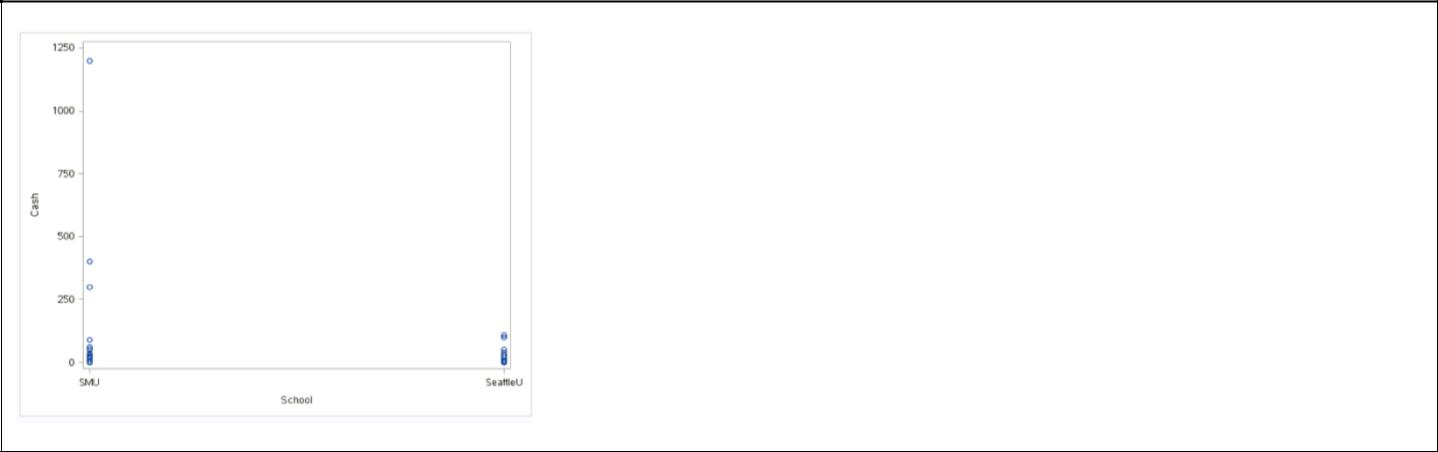
qqplot cash;

run;



\*To address t-test assumptions with scatter plot; proc sgplot data = schoolcash; scatter x= school y = cash;

run;



**Part B (15 points total)**

Now perform a complete analysis of the data. You may use either the permutation test from HW 1 or the t-test from HW 2 (copy and paste) depending on your answer to part a. In your analysis, be sure to cover all the steps of a complete analysis.

1. State the problem.
2. Address the assumptions of the t-test (from part a)
3. Perform the t-test if it is appropriate and a permutation test if it is not (judging from your analysis of the assumptions).
4. Provide a conclusion, including the p-value and a confidence interval.
5. Provide the scope of inference.

NOTE: AGAIN, THIS QUESTION SHOULD BE EASY, AS YOU ARE SIMPLY FORMATTING YOUR RESULTS FROM EARLIER IN THE ABOVE FORM. IT REALLY JUST EQUATES TO ADDING A STATEMENT OF THE PROBLEM AND ADDRESSING THE ASSUMPTIONS (1 or 2 above.) Steps 3-5 are from your previous HW; you are just putting everything together. You can basically copy and paste the rest. We are simply putting everything together to make a complete report.

**Problem (2 points): Test the claim that the mean amount of pocket cash in SMU’s and Seattle U’s students’ pockets is different.**

**Assumptions (3 points): The assumptions are as stated in parts A and B (you do not necessarily need to re-state them). Since the assumptions of the t-test are not met, we will instead conduct a permutation test for the difference in sample means.**

**What follows is an abridged version of the problem from HW #1, but you can see how everything flows together in a complete analysis. The code, output, and steps have been omitted for parsimony, but they are available in the HW #1 solutions.**

*Note: Remember, your p-values may be slightly different but will in all likelihood be within 0.05 of this answer key.*

**(10 points for the test and conclusion) To test for a difference of population means between the SMU and Seattle groups, a permutation test was conducted on 1,000 random permutations of the data. Note that a full permutation test is ideal, but SAS will likely crash due to the computing work required for this data. Hence, we perform our analysis on 1,000 random permutations of the data. A histogram of the 1,000 differences of sample**

**means from the 1,000 permutations can be viewed (in prior HW). The observed difference was $114, where 135 of the 1000 permutations yielded a difference in sample means that was as extreme or more extreme than this observed difference (p-value = 135/1000 = 0.135). This does not provide sufficient evidence against the null hypothesis that the mean pocket cash of SMU students is equal to that of Seattle University students. These 30 students were not from a random sample; therefore, inference cannot be extended beyond the 30 subjects in the sample. Since we failed to reject Ho, there is no need to write up whether causal inference can be drawn.**

**Part C (6 points total)**

Note the potential outlier in the SMU data set. Re-check the assumptions in SAS or R without the outlier. Does this change your decision about the appropriateness of the t-tools? Compare the p-value from t-test with and without the outlier. Based on your analysis so far, what should we do with this outlier? Consult the outlier flowchart in Section 3.4.

*Note: 3 points for running the t-test and 3 points for reporting and discussing the results.*

##Remove the $1,200 from SMU

SMU = **c**(34, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0)

Seattle = **c**(20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0)

school1 <- **rep**('SMU', 15)

school2 <- **rep**('Seattle', 14)

school <- **as.factor**(**c**(school1, school2))

all.money <- **data.frame**(name=school, money=**c**(SMU, Seattle))

**par**(mfrow=**c**(2,2))

**hist**(SMU,xlab='SMU Pocket Cash',main='SMU')

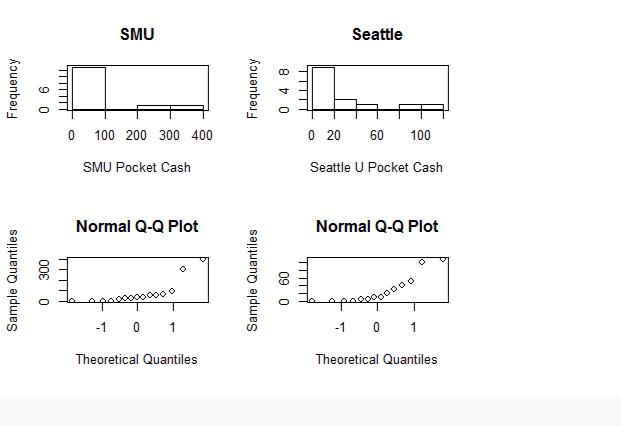
**box**()

**hist**(Seattle,xlab='Seattle U Pocket Cash',main='Seattle')

**box**()

**qqnorm**(SMU)

**qqnorm**(Seattle)



**t.test**(money **~** school,data=all.money,var.equal=T)

##

* Two Sample t-test
* data: money by school
* t = -1.3402, df = 27, p-value = 0.1913
* alternative hypothesis: true difference in means is not equal to 0
* 95 percent confidence interval:
* -111.53155 23.39821
* sample estimates:

|  |  |  |
| --- | --- | --- |
| ## | mean in group Seattle | mean in group SMU |
| ## | 27.00000 | 71.06667 |

**Even without the outlier, there is still considerable evidence against normality and equal standard deviations; while the sample size may be big enough to make the test robust to the normality assumption and the similar sample size may imply that the test is also robust to the equal standard deviation assumption, it is more conservative to go with a permutation test here. The p-value from the t-test with the outlier was 0.1732 and without the outlier it was 0.1913. We note that it does not make a large difference (and no difference in the decision not to reject Ho at alpha = 0.05), although we also did not think the t-test was appropriate for these data. A more appropriate analysis would be to run the permutation test with and without the outlier. This was done and the p-value with the outlier was 0.135 (above) and the p-value without the outlier was 0.261. Again, the decision is the same, thus we will keep the analysis above and report the result with the outlier.**

\*SAS Code without outlier;

\*To remove outlier;

data schoolcashNoOutlier;

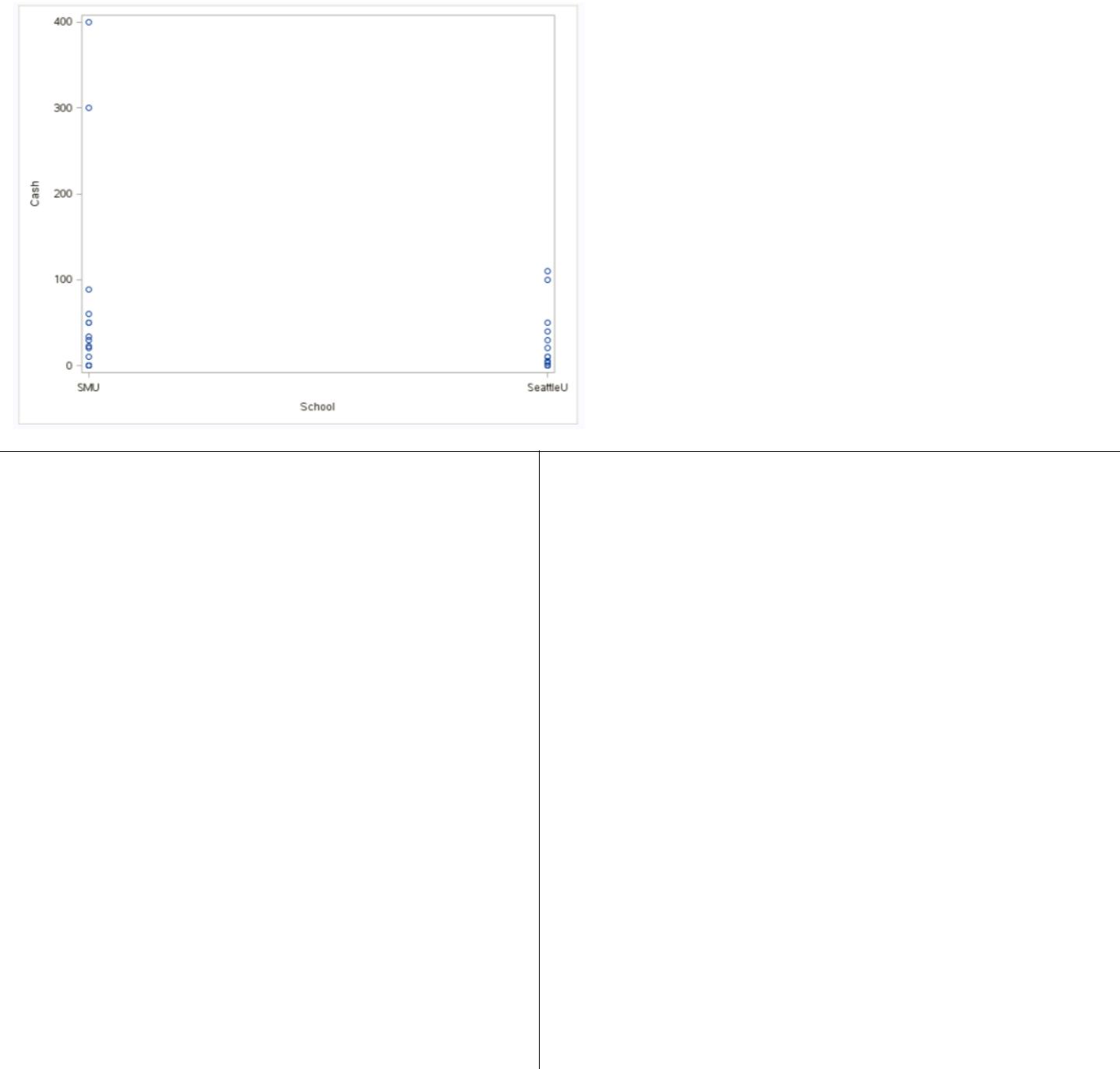
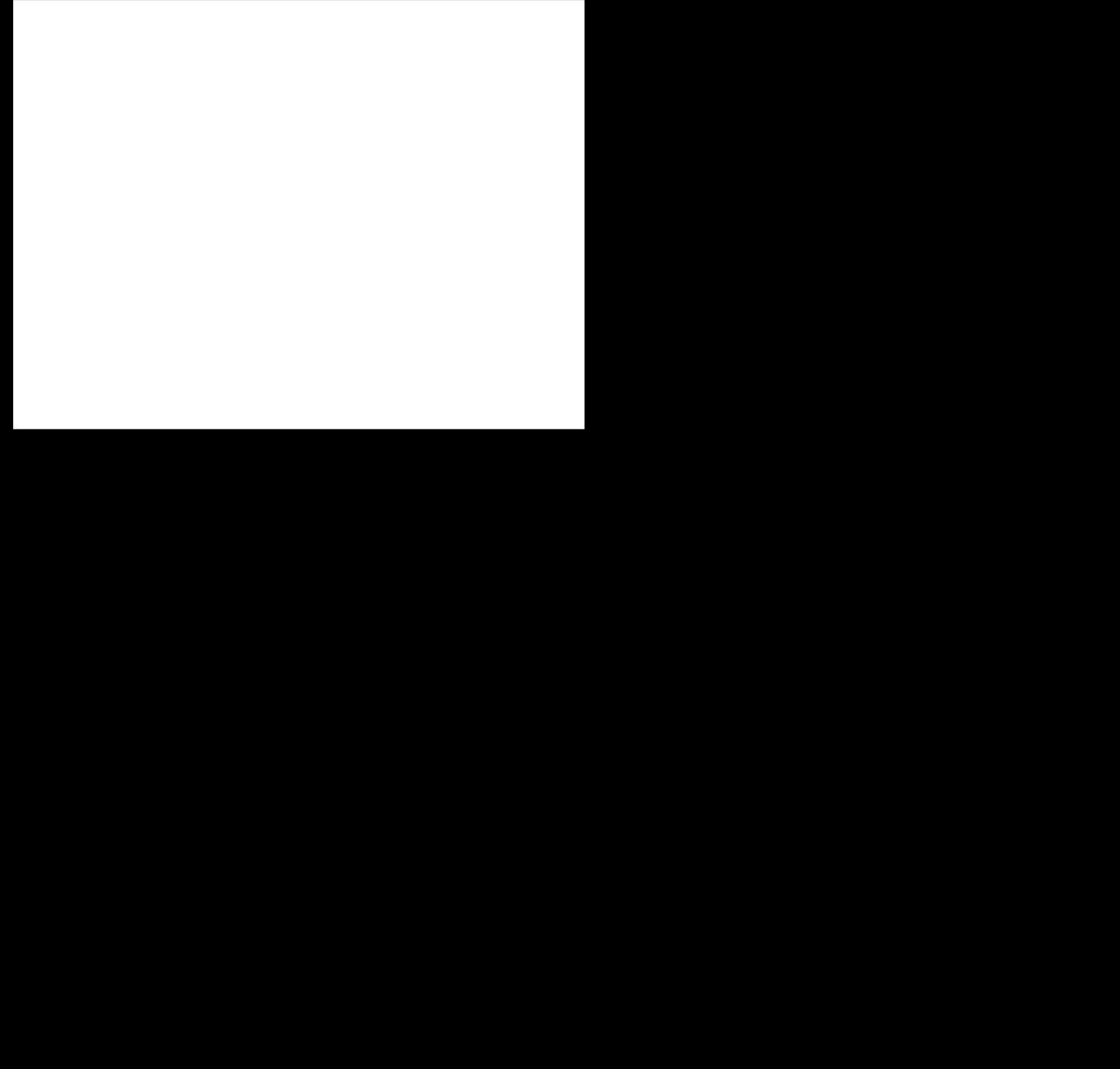
set schoolcash;

where cash < 1200;

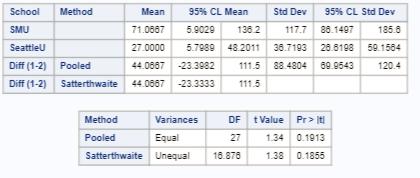
run;

\*To address t-test assumptions with scatter plot without outlier; proc sgplot data = schoolcashNoOutlier; scatter x= school y = cash;

run;



\*To run a t-test and check assumptions of t-test with histograms and q-q plots without outlier;

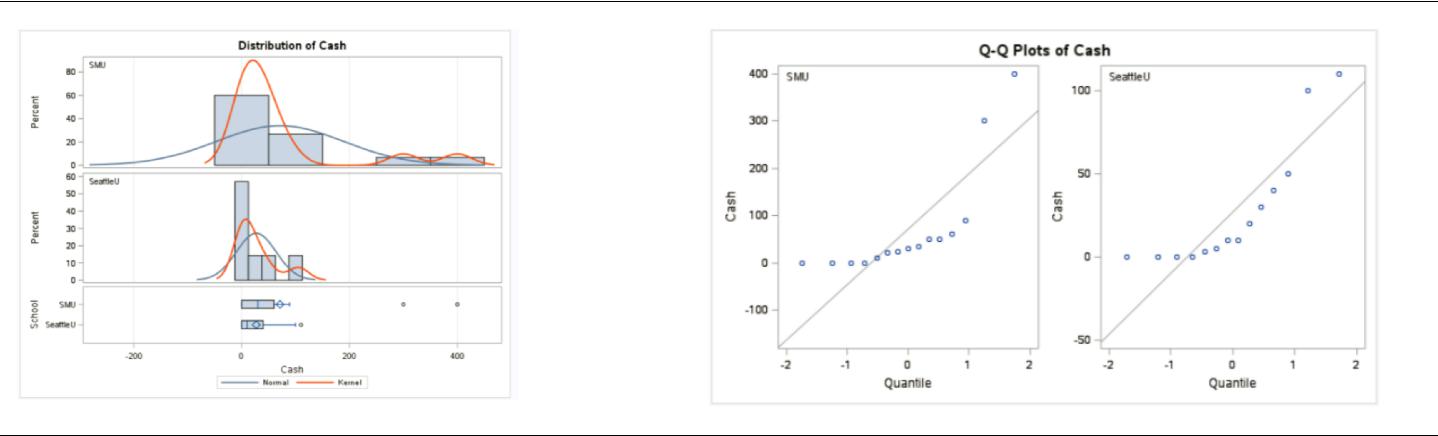
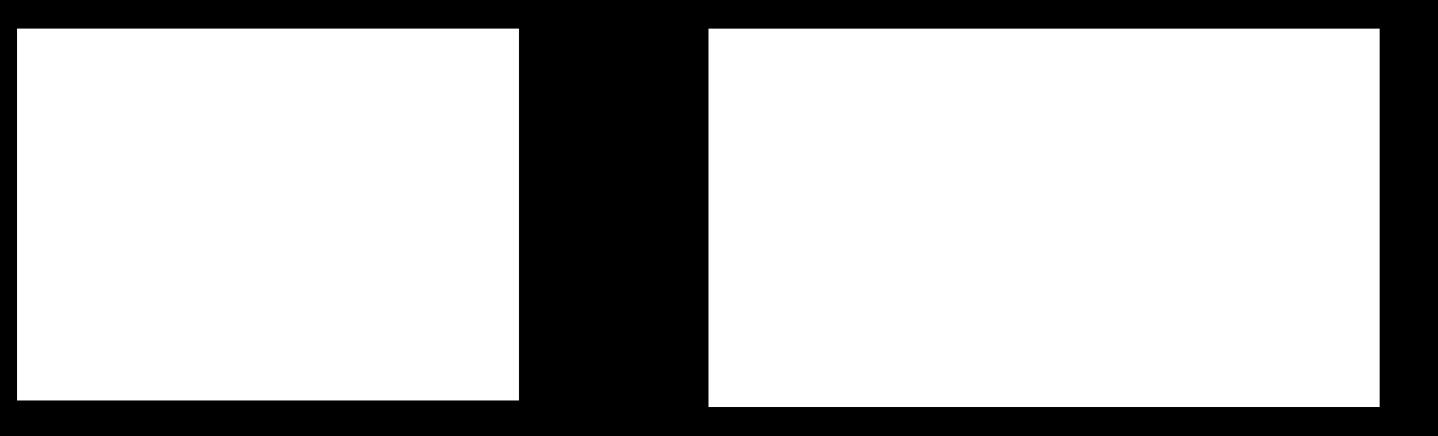


proc ttest data = schoolcashNoOutlier;

class school;

var cash;

run;



**Question 3 (40 points total)**

Find the “Education Data” data in the course materials. This data set includes annual incomes in 2005 of the subset of National Longitudinal Survey of youth (NLSY79) subjects who had paying jobs in 2005 and who had completed either 12 or 16 years of education by the time of their interview in 2006. All the subjects in this sample were between 41 and 49 years of age in 2006. Test the claim that the distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education. (Hint: pay careful attention to the ratio between the largest and smallest incomes in each group… also… is the bigger mean associated with the bigger standard deviation? Transformation?)

*Note: There is some SAS code in the course materials to help you download the data into SAS. It is a very large dataset… “datalines” is not a good idea here! You could also use the File/Import option.*

Finally, make sure you present your findings as you would to a client:

1. State the Problem.
2. Address the Assumptions (graphically and using words).
3. Perform the Most Appropriate (Powerful) Test (in reality, this may be a pooled t-test on the original data, a t-test on the log transformed data, or a permutation test on the original data, since these are the ones we have studied so far. For now, assume you must choose between the pooled t-test on the original data or on the log transformed data.)
4. Provide a conclusion including a p-value and a confidence interval.
5. Provide a scope of inference.

**(5 points) Problem: Test the claim that the distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education.**

**(10 points) Assumptions: In order to test the claim that the “distribution” of incomes of those with 16 years of education exceed the “distribution” of incomes with 12 years of education we will focus on the location parameters: the median. We will test if there is sufficient evidence to suggest that the median income of those with 16 years of education exceeds the median income of those with only 12 years of education. The reason to look at medians is that the t-test is not an appropriate test for the raw data. The histogram and box plot below indicate strong evidence of inequality of variance between the two populations. A quick look at sample size indicates that the smaller sample size is associated with the larger standard deviation, which is when the t-test is least robust.**

##Read in the data, note your directory will be different ##You could've used SAS as well!

edu <- **read.csv**('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 3/HW/EducationData.csv')

**par**(mfrow=**c**(2,2))

**hist**(**subset**(edu, Educ**==**12)**$**Income2005,xlab='Income, 12 Years Education',main='12 Years')

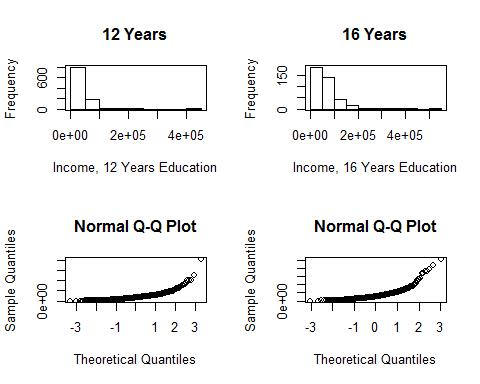
**box**()

**hist**(**subset**(edu, Educ**==**16)**$**Income2005,xlab='Income, 16 Years Education',main='16 Years')

**box**()

**qqnorm**(**subset**(edu, Educ**==**12)**$**Income2005)

**qqnorm**(**subset**(edu, Educ**==**16)**$**Income2005)



**However, the larger mean is associated with the larger standard deviation, which indicates that a log transformation of the data may be appropriate. The histogram and box plot below are for the log transformed data and indicate that there is evidence that the standard deviations may be equivalent for the log transformed data. While the q-q plots looks more normal, the normality of the original data is of little concern, given that the large sample size should ensure the sampling distribution of the means is normal (from the Central Limit Theorem CLT). Since the visual check provides strong evidence for the equality of standard deviations, we will proceed with a t-test to test the difference of means of the log transformed data. Note: this is actually a test of the ratio of medians of the original data.**

edu**$**log.income <- **log**(edu**$**Income2005)

**par**(mfrow=**c**(2,2))

**hist**(**subset**(edu, Educ**==**12)**$**log.income,xlab='Log Income, 12 Years Education',main='12 Years')

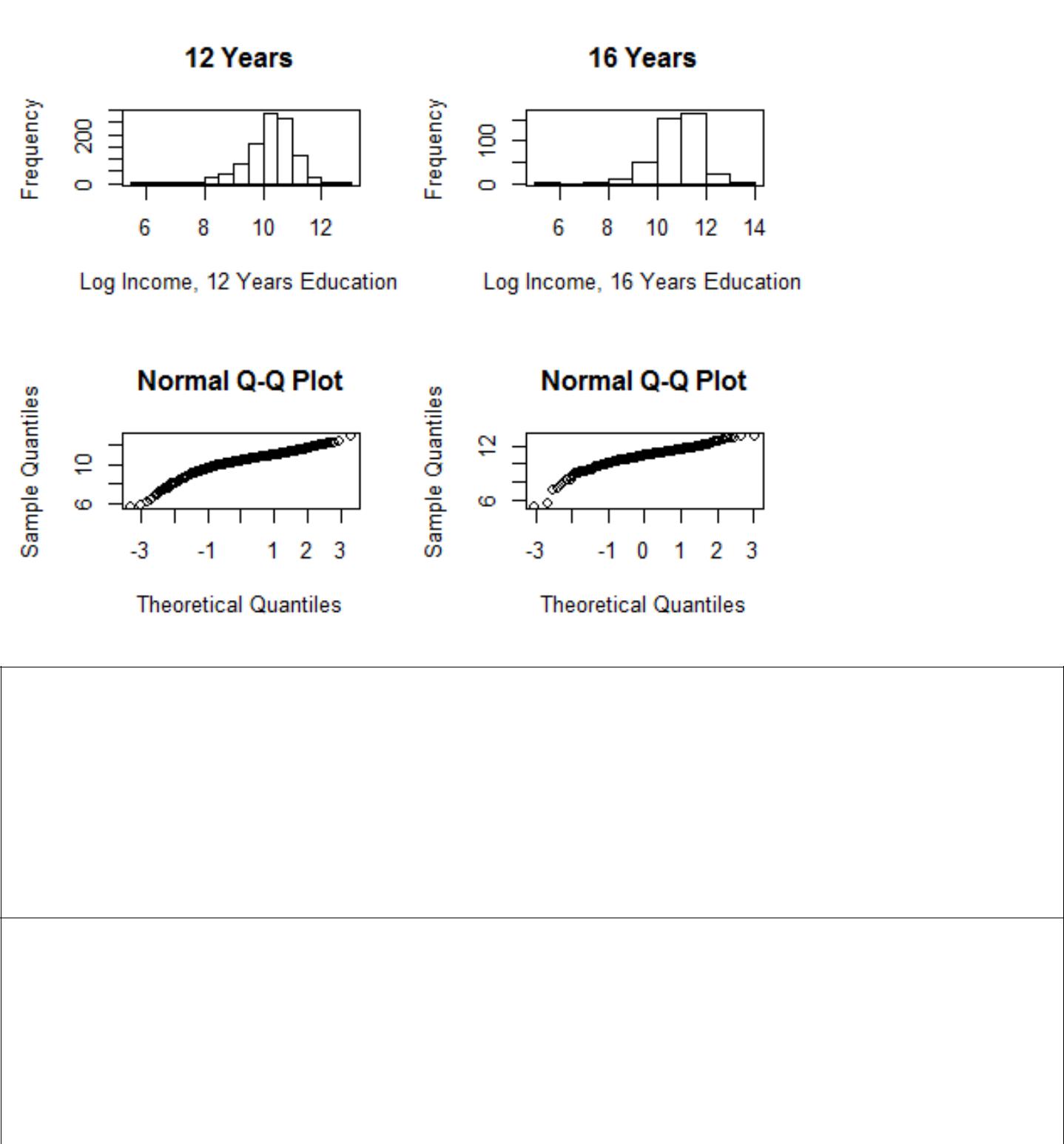
**box**()

**hist**(**subset**(edu, Educ**==**16)**$**log.income,xlab='Log Income, 16 Years Education',main='16 Years')

**box**()

**qqnorm**(**subset**(edu, Educ**==**12)**$**log.income)

**qqnorm**(**subset**(edu, Educ**==**16)**$**log.income)



\*To import education data;

FILENAME REFFILE '/home/sadiet0/my\_courses/bsadler0/MSDS 6371/UNIT 3/EducationData.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=education;

GETNAMES=YES;

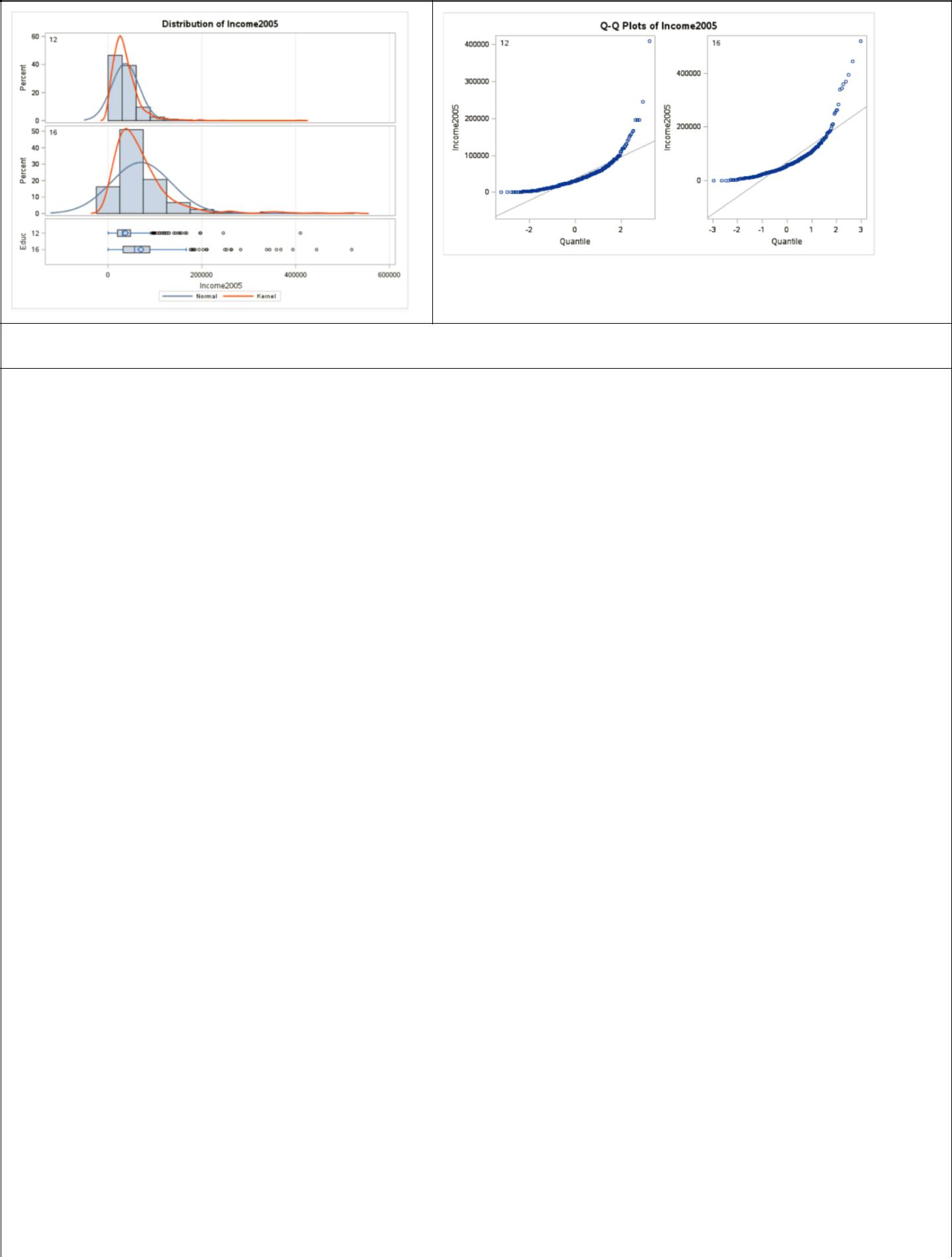
RUN;

**Original Data:**

\*To check t-test assumptions on original data; proc ttest data = education sides = l; class educ;

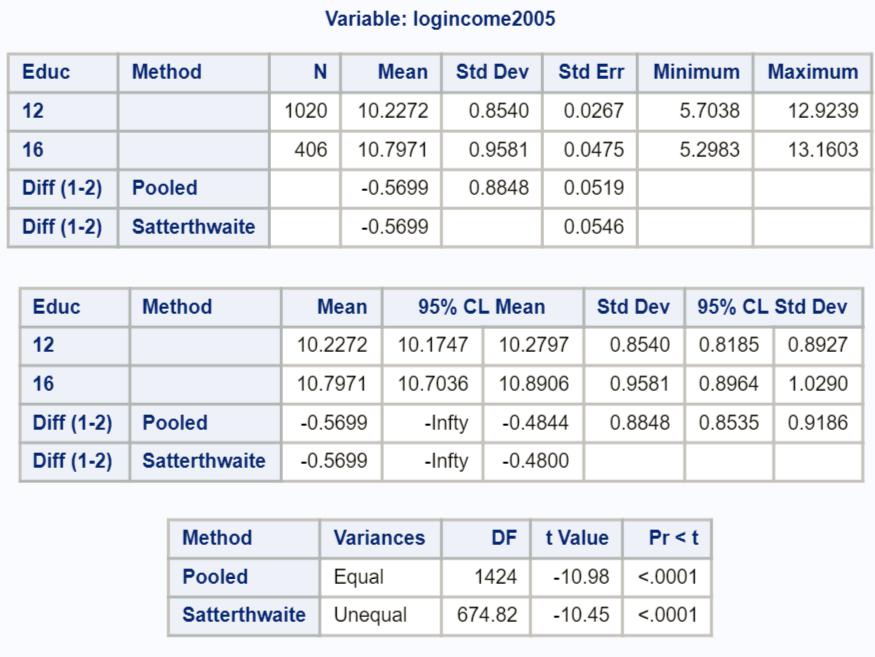
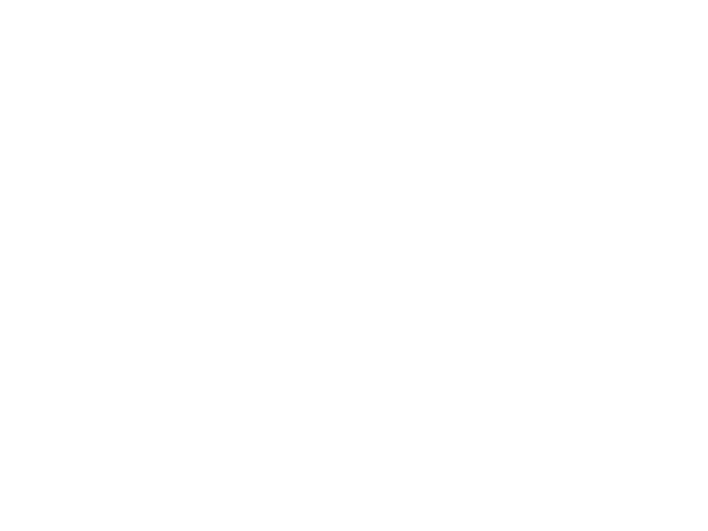
var income2005;

run;



Logged data:

**\*To create a logged variable; data education; set education;**

****

**logincome2005 = log(income2005); run;**

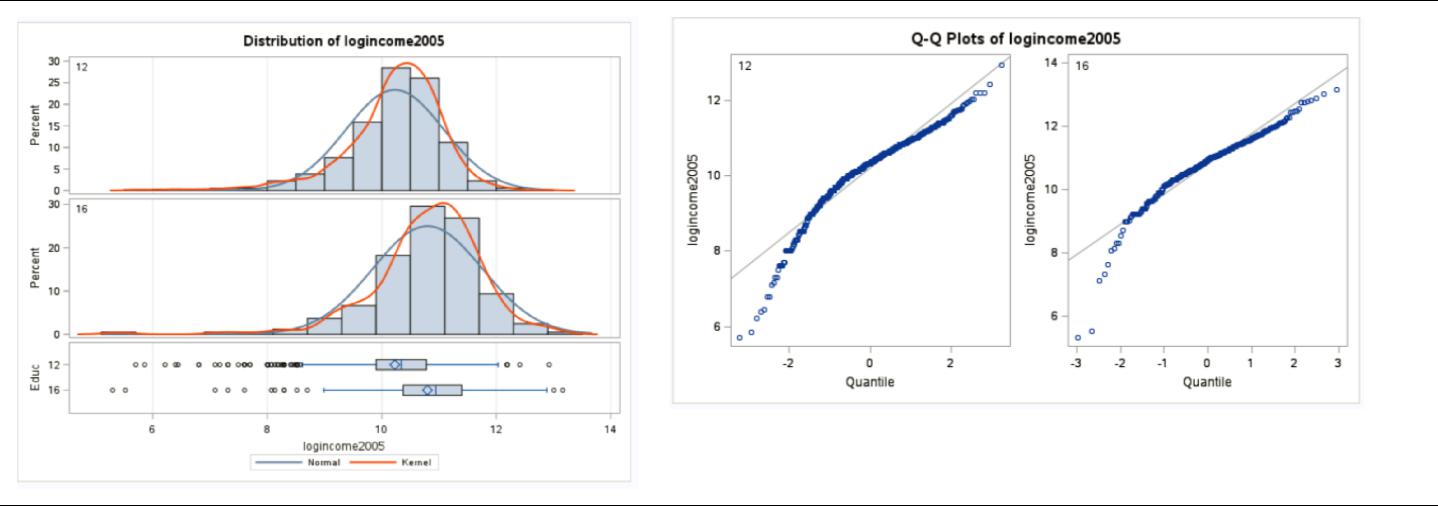
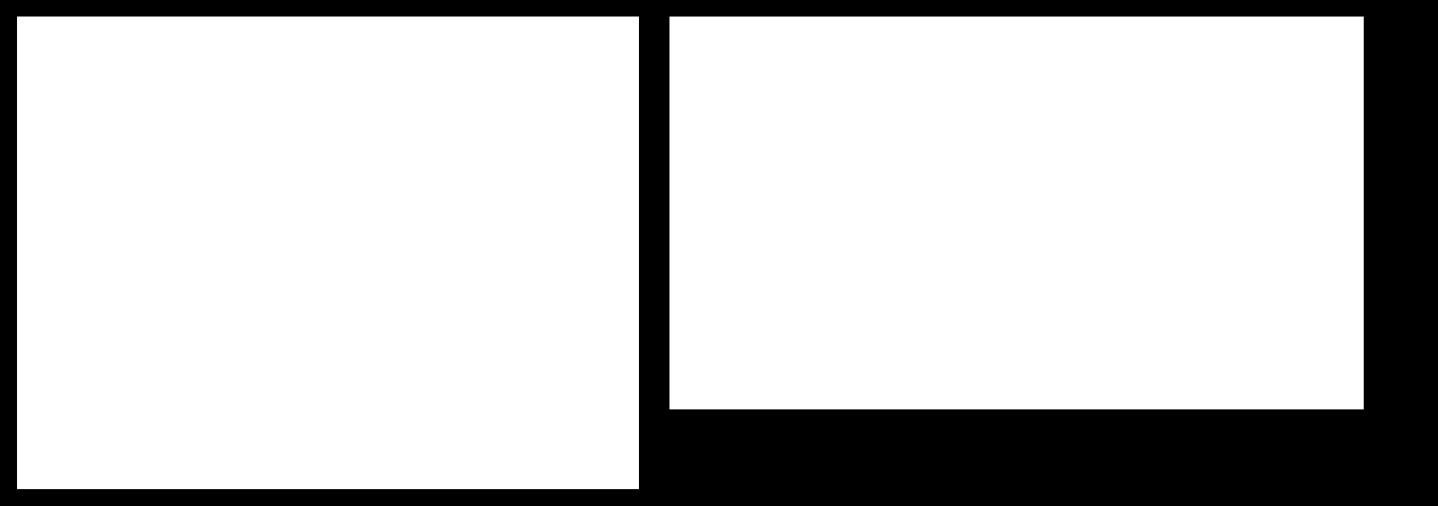
**\*To perform the t-test and check assumptions on log transformed data;**

**proc ttest data = education sides = l; class educ;**

**var logincome2005;**

**run;**

\*Note that the test statistic is negative because of the way the data is sorted.



**\*To get critical value for one-sided t-test at level**

****

**alpha = 0.05;**

**data critval;**

**cv = quantile("t", .05, 1424); run;**

**proc print critval;**

**run;**



**\*To find a 90% confidence interval to align with**

**an alpha = 0.05 one-sided hypothesis test;**

**proc ttest data = education sides = 2 alpha = 0.1;**

**class educ;**

**var logincome2005;**

**run;**

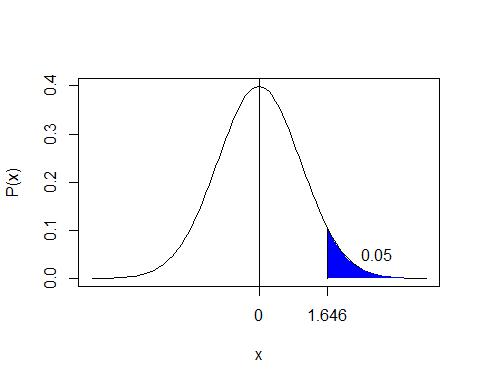
**Step 1 - Hypotheses (2 points):**

: (16 ) = (12 )

: (16 ) > (12 )

*Note: the null hypothesis could also be less than or equal to. If you did a 2-sided test, 2 points should be deducted here and the remaining work should be evaluated as if a 2-sided test were acceptable. (In other words, we don’t want you to miss the entire problem because you tested the wrong hypothesis.)*

**Step 2 - Identification of Critical Value (1 point for drawing, 1 point for value):** 1.646



**Step 3 - Value of Test Statistic (2 points):** = 10.98

**Step 4 - Give p-value (2 points):** < 0.0001

**Step 5 - Decision (2 points): Reject**

**Step 6 - Conclusion (5 points for the statistical conclusion, 5 points for the confidence interval, 5 points for discussing the scope): There is overwhelming evidence at the alpha = 0.05 level of significance (** < 0.0001**) that the median income in 2005 for people with 16 years of education is 1.77 times as large as the median income for those in the study that had only 12 years of education. A 90% confidence interval for this factor is [e-.6553, e-**

**.4844]= [1.62, 1.93]. This was an observational study, and thus we cannot confirm that the years of education caused the increase in income, only that they are associated with each other. There is little detail about the randomness of the sample, although it is doubtful that it was a random sample. We must limit the inference gained from this study to only the subjects of this sample.**

*Note: an alternative way to state the conclusion would be to say the median income for those with 16 years of education was 77% larger than the median income for those that had only 12 years of education, with the confidence interval being 62% to 93%.*

*Note: if the order of your data was different, you would have gotten that the median income for those with 12 years of education was 0.57 times the median income of those with 16 years of education. The confidence interval would have been [e-.6553, e-.4844] =[0.519, 0.616]. This is completely fine as long as you interpret the results properly.*

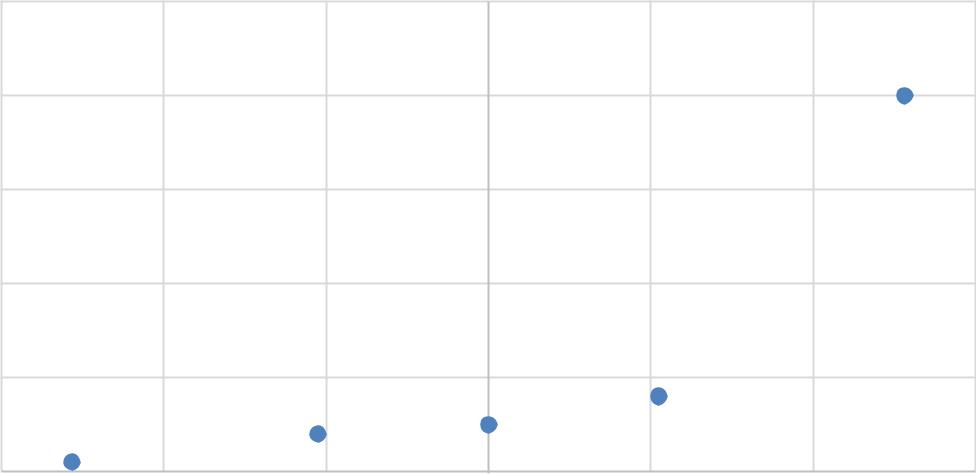
**Bonus (+5 points total)**

Create 2 q-q plots (by hand) for the original data in Chapter 3, question 20 of the text book. A q-q plot for the In-State and a q-q plot for the Out-Of-State data. Show all work by filling in a table like the one below (one for In-State and one for Out-of-State):

|  |  |  |  |
| --- | --- | --- | --- |
| Original Data | Percentage for | Z-Score of | Z-Scores for |
| (In-State) | Percentiles | Original Data | Percentiles |
|  |  |  |  |
| 1000 | 0.1 | -0.65954 | -1.28155 |
| 4000 | 0.3 | -0.47288 | -0.5244 |
| 5000 | 0.5 | -0.41066 | 0 |
| 8000 | 0.7 | -0.22399 | 0.524401 |
| 40000 | 0.9 | 1.76708 | 1.281552 |

Original Data (Out-of-State)

50000



40000

30000

20000

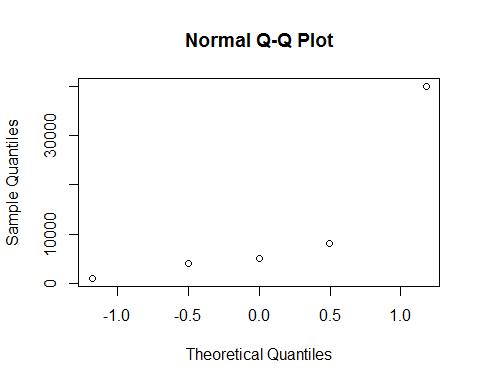
10000

0

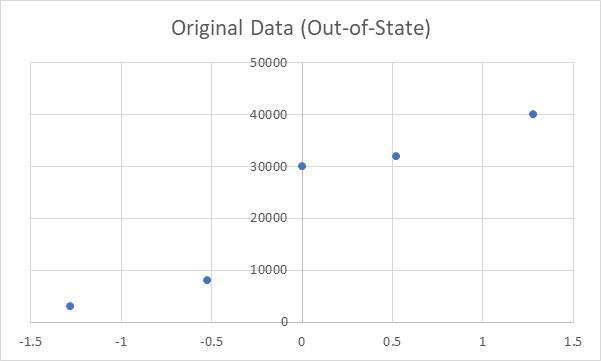
-1.5 -1 -0.5 0 0.5 1 1.5

in.state <- **c**(1000,4000,5000,8000,40000)

**qqnorm**(in.state)

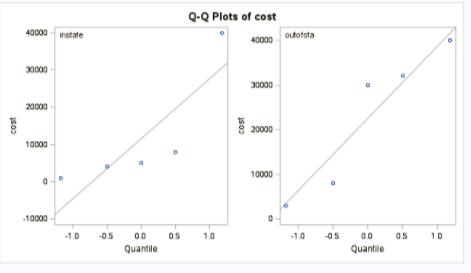


|  |  |  |  |
| --- | --- | --- | --- |
| Original Data | Percentage for | Z-Score of | Z-Scores for |
| (Out-of-State) | Percentiles | Original Data | Percentiles |
|  |  |  |  |
| 3000 | 0.1 | -1.21367 | -1.28155 |
| 8000 | 0.3 | -0.90406 | -0.5244 |
| 30000 | 0.5 | 0.458224 | 0 |
| 32000 | 0.7 | 0.582069 | 0.524401 |
| 40000 | 0.9 | 1.077446 | 1.281552 |



Check your q-q plots by comparing them with the ones from proc ttest. (Run proc ttest but just for the q-q plots. You do not need to run a full hypothesis test.) What would you conclude about the normality of the distributions these data came from?

\*To get q-q plots;



proc ttest data = tuition; class location; var cost;

run;

**The out-of-state q-q plot provides more evidence of normality than the in-state. However, with a sample size of only 5, neither q-q plot provides evidence of extreme departures from normality.**

# Homework #4 Solutions

*This class allows you to practice preparing professional looking reports. Make sure all reports are typed and all graphs (unless otherwise noted) are computer generated and copied and pasted into your report. If you would like help with Word or Excel please don’t hesitate to ask.*

## Question 1 (0 points total)

Read Chapter 4 from Statistical Sleuth and answer the conceptual problems at the end of the chapter. Note: You do not need to type these up and turn them in. The answers are at the very end of the chapter.

## Question 2 (20 points total)

When wildfires ravage forests, the timber industry argues that logging the burned trees enhances forest recovery; the EPA argues the opposite. The 2002 Biscuit Fire in southwest Oregon provided a test case. Researchers selected 16 fire-affected plots in 2004-before any logging was done and counted tree seedlings along a randomly located transect pattern in each plot. They returned in 2005, after nine of the plots had been logged, and counted the tree seedlings along the same transects. The percent of seedlings lost from 2004 to 2005 is recorded in the table below for logged (L) and unlogged (U) plots:

Test the EPA’s assertion (and thus the opposite of the logging industry’s assertion) that logging actually increases the percentage of seedlings lost from 2004 to 2005.

### Part A (15 points total)

Perform a complete analysis using a rank sum test in SAS.

**Problem (1 point): Test the EPA’s assertion (and thus the opposite of the logging industry’s assertion) that logging actually increases the percentage of seedlings lost from 2004 to 2005.**

**Assumptions (2 points): The data are ordinal as they are percentages of seedlings. We will assume the observations are independent, although significant spatial correlation may be present. You may also have assumed that the distributions were the same, except for location.**

**Step 1 - Hypotheses (2 points):**

**The distribution of the percent of lost seedlings in the logged plots is equal to that of the unlogged plots.**  
 **The distribution of the percent of lost seedlings in the logged plots is more than that of the unlogged plots.**

*Note: you may also state the hypotheses in terms of medians.*

**Step 2 - Identification of Critical Value (2 points): critical value for Normal approximation (you may leave this out if you used “Exact” p-values) (1-sided, )**

**Step 3 - Value of Test Statistic (2 points):**

**Step 4 - Give p-value (2 points): (1-sided)**

**Step 5 - Decision (1 point): Reject (assuming )**

**Step 6 - Conclusion (2 points): There is sufficient evidence to suggest that the distribution of the percent seedlings lost in the logged plots is more than that of the unlogged plots ( from a one-sided rank sum test). A 95% confidence interval for the increase in median percent seedlings lost in logged plots is . Note that a 90% confidence interval may be more appropriate for a one-sided test with significance level , so that the results of thy hypothesis test and confidence intervals match up.**

**Scope of inference (1 point): Since the plots were not randomized to receive either the logging or not logging treatment, no causation can be implied here. Since the transect patterns were randomly selected, this inference can be generalized to the 16 larger plots.**

*Note: the exact p-value 0.0058 could have been used and steps 2 and 3 left out.*

SAS:

\*To compute rank sum test;

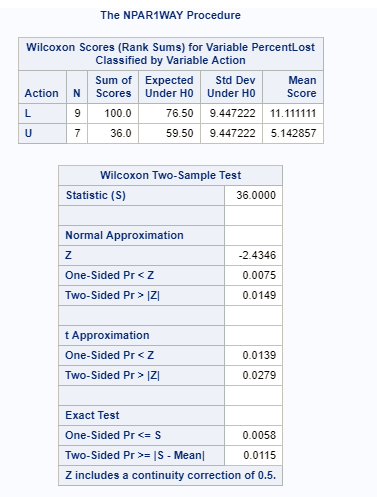
proc npar1way data = loggingdata wilcoxon;

class Action;

var PercentLost;

exact HL Wilcoxon;

run;



To get a corresponding 90% HL Confidence interval (to match with the alpha = 0.05 one-sided hypothesis test).

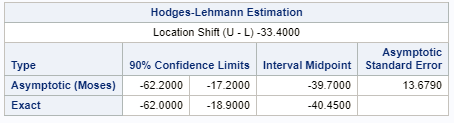
proc npar1way data = loggingdata wilcoxon alpha = 0.1;

class Action;

var PercentLost;

exact HL Wilcoxon;

run;



\*To get critical value for normal approximation;

data mycritval;

cv = quantile("normal", 0.95);

run;

proc print data = mycritval;

run;



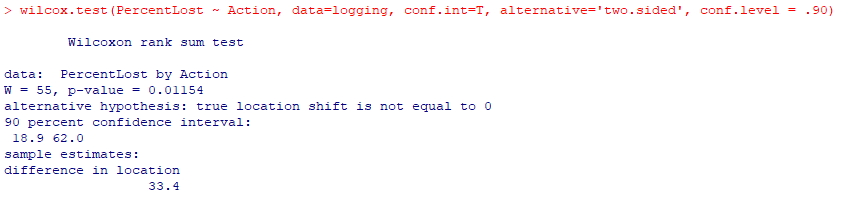
### Part B (5 points total)

Verify the p-value and confidence interval by running the rank sum test in R (using R function Wilcox.test). You do not need to repeat the complete analysis. Simply cut and paste a screen shot of your code and the output. You may use: <https://www.r-bloggers.com/wilcoxon-mann-whitney-rank-sum-test-or-test-u/> for reference.

logging <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 4/Live Session/Logging.csv')  
  
wilcox.test(PercentLost ~ Action, data=logging, conf.int=T, alternative='greater')

##   
## Wilcoxon rank sum test  
##   
## data: PercentLost by Action  
## W = 55, p-value = 0.005769  
## alternative hypothesis: true location shift is greater than 0  
## 95 percent confidence interval:  
## 18.9 Inf  
## sample estimates:  
## difference in location   
## 33.4

R code for matching 90% two-sided confidence interval:



*Note: by default, R provides the exact p-value rather than the normal approximation for the p-value for smaller sample sizes. Also, the ‘conf.int’ option provides the HL confidence limits which should match SAS exactly*

## Question 3 (25 points total)

Conduct a Welch’s two-sample t-test on the Education Data from HW 3 (untransformed).  
Perform a complete analysis using SAS to test the claim that the mean income of college educated people (16 years of education) is greater than the mean of those with a high school education only (12 years of education).

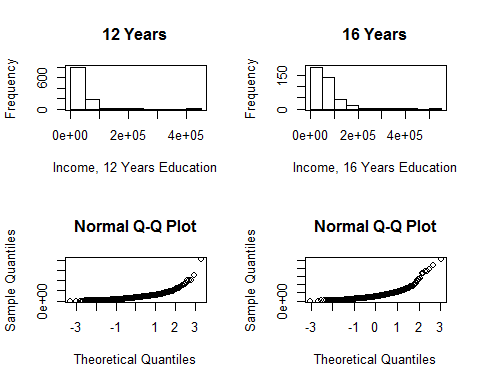
### Part A (5 points)

State the problem, address the assumptions. Be sure to support with your knowledge of theory (CLT) as well as with histograms, box plots, q-q plots, etc.

**Problem (1 point): Test the claim that the mean income in 2005 of those with a college education (16 years of education) is greater than those with only a high-school education (12 years of education).**

**Assumptions:**

##Read in the data, note your directory will be different  
##You could've used SAS as well!  
  
edu <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 3/HW/EducationData.csv')  
  
par(mfrow=c(2,2))  
hist(subset(edu, Educ==12)$Income2005, xlab='Income, 12 Years Education', main='12 Years')  
box()  
hist(subset(edu, Educ==16)$Income2005, xlab='Income, 16 Years Education', main='16 Years')  
box()  
qqnorm(subset(edu, Educ==12)$Income2005)  
qqnorm(subset(edu, Educ==16)$Income2005)



**Normality (1 point): There is strong evidence that the incomes come from right skewed distributions. We have large enough sample size (, ) to ensure a t-test will be robust to this assumption in this case.**

**Equal Standard Deviations (1 point): There is visual evidence of different standard deviations. Because the sample sizes are significantly different, the test is not robust to this assumption.**

**Independence (1 point): We will assume independence, although, since the subjects were often from the same family, this is very risky (and probably incorrect on some level). We will proceed with caution.**

**(1 point) Since the regular t-test is robust to the normality assumption for large sample sizes but not robust to the equal standard deviation assumption, we will again assume independence and run the Welch’s t-test (Satterthwaite).**

SAS Code:

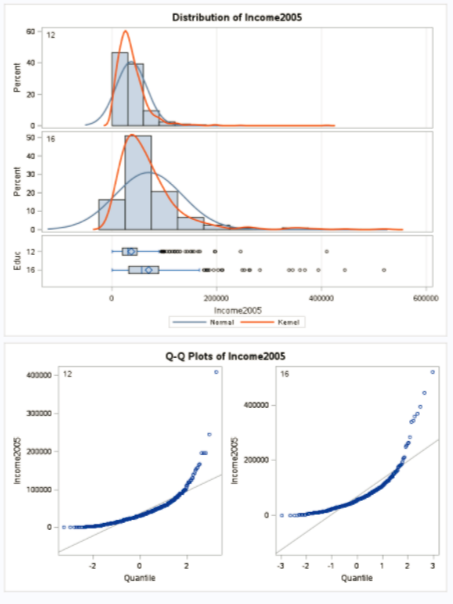
\*To check t-test assumptions on original data;

proc ttest data = education sides = l;

class educ;

var income2005;

run;



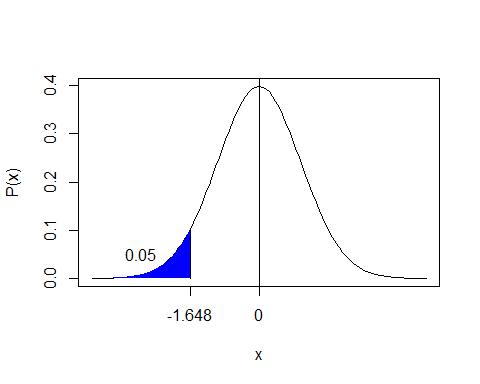
### Part B (5 points)

Show all 6 steps, including a thoughtful, thorough, yet non-technical conclusion. Include a confidence interval.

**Step 1 - Hypotheses (1 point):**

*Note: the null hypothesis could also be less than or equal to.*

**Step 2 - Identification of Critical Value (1 point):**



**Step 3 - Value of Test Statistic (1 points):**

**Step 4 - Give p-value (1 points):**

**Step 5 - Decision (0 points): Reject**

**Step 6 - Conclusion (1 point): There is overwhelming evidence at the level of significance () that the mean income in 2005 for people with 16 years of education is larger than the incomes for those in the study that had only 12 years of education. A 90% confidence interval for this increase is .**

**SAS Output:**

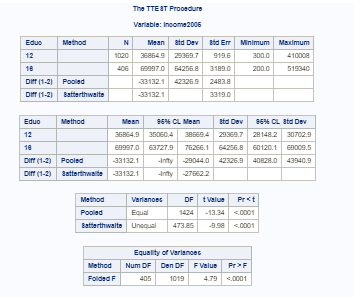
\*To perform t-test;

proc ttest data = education sides = l;

class educ;

var income2005;

run;



### Part C (5 points)

Include a scope of inference at the end. (You may copy and paste this from a previous HW if you’d like.)

**This was an observational study; therefore, we cannot conclude that the extra education caused the change (increase) in mean incomes. Households were selected from a random sample of a previously selected “area of the United States” and the subjects in this study are the members of those households. Therefore, since every member of the “area” had the same chance of being selected, it is a random sample of the “areas.” However, no indication is given on how the “areas” were selected. In conclusion, the association between education and income above can be generalized to all the members of the “areas” that were selected for this study, but not generalized to the U.S. as a whole.**

### Part D (5 points)

Verify the Welch’s t statistic and p-value with R (using R function t.test). Simply cut and paste your R code and output. You may use: <http://rcompanion.org/rcompanion/d_02.html> for reference.

##First, run the 1-sided test  
t.test(Income2005 ~ Educ, var.equal=F, data=edu, alternative='less')

##   
## Welch Two Sample t-test  
##   
## data: Income2005 by Educ  
## t = -9.9827, df = 473.85, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
## -Inf -27662.19  
## sample estimates:  
## mean in group 12 mean in group 16   
## 36864.90 69996.97

##Then, run a 2-sided test ONLY for the confidence interval  
t.test(Income2005 ~ Educ, var.equal=F, data=edu, conf.level=0.9)

##   
## Welch Two Sample t-test  
##   
## data: Income2005 by Educ  
## t = -9.9827, df = 473.85, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 90 percent confidence interval:  
## -38601.97 -27662.19  
## sample estimates:  
## mean in group 12 mean in group 16   
## 36864.90 69996.97

*Note: the alternative was ‘less’ because R is testing whether the mean of 12 years minus the mean of 16 years is less than 0, which is equivalent to testing whether the mean of 16 years minus the mean of 12 years is greater than 0.*

### Part E (5 points)

Would you prefer to run the log transformed analysis you ran in HW3, or do you feel this analysis is more appropriate? Why or why not? (Make mention of the assumptions as well as the parameters that each test provides inference on. As you know, they are different.)

**Since the original distributions of income are right skewed, the median may be a more valuable measure of center. For this reason, the log transformation with the test analysis may be preferred. This is largely a subjective/opinion question as long as it is defended appropriately. Alternatively, a rank sum test may be conducted as well here. This will give inference with respect to the median with no distributional assumptions.**

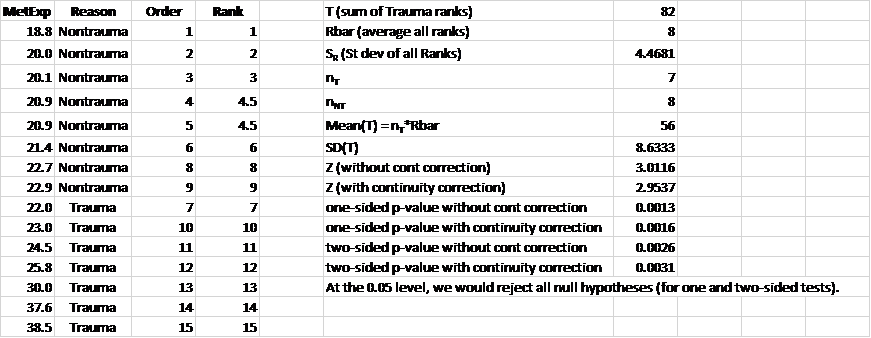
*Note: any reasonable argument can receive full credit.*

## Question 4 (25 points total)

### Part A (10 points)

Chapter 4, Problem 20 from the text. Show all work. “By hand” here means actually by hand. Simply take a picture of your work and include it in your pdf/doc file. Include your sorted, labeled, and ranked data; your calculations of the mean and standard deviation of the assumed distribution of the rank sum statistic under Ho; your calculation of the Z statistic with a continuity correction; your p-value; and conclusion (no confidence interval is necessary here).

*Note: it is not entirely clear if the question of interest calls for a 1-sided or 2-sided analysis, so both are permissible. Give 1 point for T, 1 point for , 1 point for , 1 point for Mean(T), 1 point for SD(T), 1 point for Z, 1 point for the p-value, and 3 points for the conclusion. Solutions for both 1-sided and 2-sided tests are given below.*



### Part B (5 points)

Problem 21 from the text. Take a screen capture of the SAS output in addition to your response.

**The one from SAS is 0.0016 from the normal approximation and 0.0006 from the exact test. Yes, SAS uses the continuity correction.**

\*To perform rank sum test;

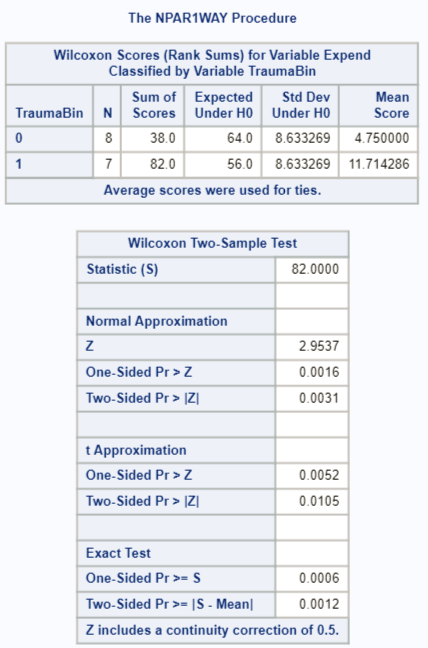
proc npar1way data=Metabolic WILCOXON;

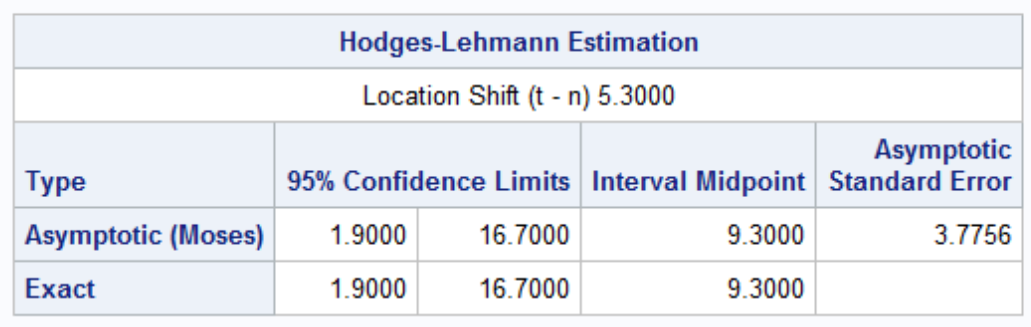
class Traumabin;

var Expend;

exact;

run;





### Part C (10 points)

Write up a complete analysis using the information you have gained from A and B to test the claim that the distributions are different.  
i. State the problem.  
ii. State the assumptions you are making and why you are making them. Justify your decisions. Print out any histograms, q-q plots, box plots, etc. that you use in your justification.  
iii. Show all 6 steps of the hypothesis test for the rank sum test of the trauma data. Use the critical values, test statistics, p-values, etc. obtained above. Add a confidence interval from the Hodges-Lehmann procedure (from SAS).  
iv. Also include a scope of inference statement.

**Problem (1 point): Test if the distribution of metabolic expenditure of non-trauma patients is different than that of trauma patients.**

*Note that a 1-sided test is permissible, given the limited information we have. Other values of alpha are also permissible.*

**Assumptions (1 point): Metabolic expenditure are ratio level data and thus ordinal. We will assume the subjects are independent. You may also have assumed that the distributions were the same, except for location.**

**Step 1 - Hypotheses (1 point):**

**The distribution of the metabolic expenditures for trauma and non-trauma patients are equal.**  
 **The distribution of the metabolic expenditures for trauma and non-trauma patients are different.**

*Note: you may also state the hypotheses in terms of medians.*

**Step 2 - Identification of Critical Value (1 point): critical value for Normal approximation (you may leave this out if you used “Exact” p-values) (2-sided, )**

**Step 3 - Value of Test Statistic (1 point):**

**Step 4 - Give p-value (1 point): (2-sided)**

**Step 5 - Decision (1 point): Reject (assuming )**

**Step 6 - Conclusion (2 points): There is sufficient evidence to suggest that the distribution of metabolic expenditure for the trauma group is different from that of the non-trauma group ( from a two-sided rank sum test). A 95% confidence interval for the difference in medians is kcal/kg/day.**

**Scope of inference (1 point): The data are strictly observational, and thus only an association between the presence of trauma and metabolic expenditure can be drawn. It is unclear how the sample was taken; therefore, we will assume the patients were not selected at random and the inference here is limited to the 15 subjects in the survey.**

SAS Code:

\*Critical value for two-sided test;

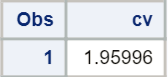
data mycritval;

cv = quantile ("normal", 1-(0.05/2));

run;

proc print data = mycritval;

run;



## Question 5 (30 points total)

A study was performed to test a new treatment for autism in children. In order to test the new method, parents of children with autism were asked to volunteer for the study in which 9 parents volunteered their children for the study. The children were each asked to complete a 20 piece puzzle. The time it took to complete the task was recorded in seconds. The children then received a treatment (20 minutes of yoga) and were asked to complete a similar but different puzzle. The data from the study is below:

|  |  |  |
| --- | --- | --- |
| Child | Before | After |
| 1 | 85 | 75 |
| 2 | 70 | 50 |
| 3 | 40 | 50 |
| 4 | 65 | 40 |
| 5 | 80 | 20 |
| 6 | 75 | 65 |
| 7 | 55 | 40 |
| 8 | 20 | 25 |
| 9 | 70 | 30 |

### Part A (5 points)

Calculate the statistic S for a signed rank test by hand showing the final table with the absolute differences, the signs, and the ranks. Also, show your calculation of the z-statistic (standardized S statistic).

**Table (1 point):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Child | Before | After |  | Sign | Rank |
| 1 | 85 | 75 | 10 | - | 1 |
| 2 | 70 | 50 | 20 | + | 3 |
| 3 | 40 | 50 | 10 | - | 3 |
| 4 | 65 | 40 | 25 | + | 3 |
| 5 | 80 | 20 | 60 | + | 5 |
| 6 | 75 | 65 | 10 | + | 6 |
| 7 | 55 | 40 | 15 | + | 7 |
| 8 | 20 | 25 | 5 | + | 8 |
| 9 | 70 | 30 | 40 | + | 9 |

**(1 point)**

**(1 point)**

**(1 point)**

**(1 point)**  (The research question lends itself to a one-sided test.)

\*To find a one-sided p-value associated with a z-statistic of 2.13;

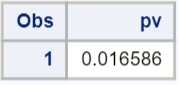
data pvalue;

pv = 1-probnorm(2.13);

run;

proc print data = pvalue;

run;



### Part B (5 points)

Verify your calculation in both SAS and R. Simply cut and paste your code and relevant output.

*Note: each software output is worth 2.5 points.*

\*To create the variable that is a difference between the two measurements;

data autism;

set autism;

diff = before - after;

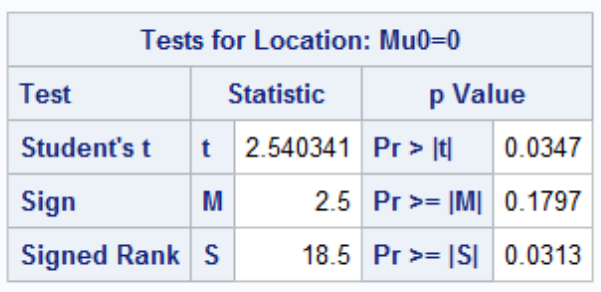
run;

\*To get a signed rank test;

proc univariate data = autism;

var diff;

run;



*Note: the p-value for the one-sided signed-rank test in SAS 0.0313/2 = 0.0157, which is close but not exactly the same as in R. SAS, R, and the textbook all use a slightly different test statistic.*

autism <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 4/HW/Autism.csv')  
  
wilcox.test(autism$Before, autism$After, pair=T)

## Warning in wilcox.test.default(autism$Before, autism$After, pair = T):  
## cannot compute exact p-value with ties

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: autism$Before and autism$After  
## V = 41, p-value = 0.03236  
## alternative hypothesis: true location shift is not equal to 0

wilcox.test(autism$Before, autism$After, pair=T, alternative='greater')

## Warning in wilcox.test.default(autism$Before, autism$After, pair = T,  
## alternative = "greater"): cannot compute exact p-value with ties

##   
## Wilcoxon signed rank test with continuity correction  
##   
## data: autism$Before and autism$After  
## V = 41, p-value = 0.01618  
## alternative hypothesis: true location shift is greater than 0

### Part C (5 points)

Conduct the six step hypothesis test using your calculations from above to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle.

**Step 1 - Hypotheses (1 point):**

**The median difference in time to finish the puzzle is equal to 0.**  
 **The median difference in time to finish the puzzle is greater than 0.**

*Note: this assumes you calculate the differences as Before - After. If you did it the other way around, the sign on your alternative hypothesis should be the other way around.*

**Step 2 - Identification of Critical Value (1 point): critical value for Normal approximation is (1-sided, )**

**Step 3 - Value of Test Statistic (1 point):**

**Step 4 - Give p-value (1 point): (1-sided)**

**Step 5 - Decision (0 points): Reject (assuming )**

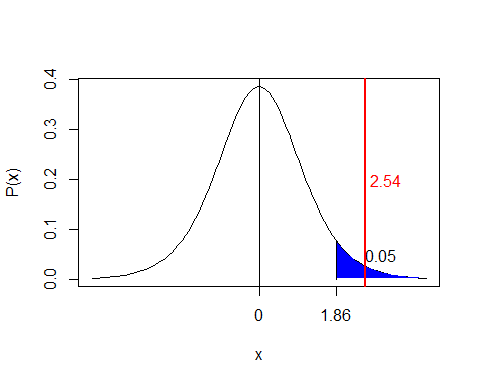
**Step 6 - Conclusion (1 point): There is evidence that the median difference in time to finish the puzzle was greater than 0 ( from a one-sided signed rank test). There is evidence that yoga is associated with shorter puzzle solution times for the individuals who participated. Because the participants were volunteers, we cannot infer that the analysis would hold for those outside the study. Nonetheless, the results could prompt further research.**

### Part D (5 points)

Use SAS to conduct a six step hypothesis test using a paired t-test to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle.

**Step 1 - Hypotheses (1 point):**

**Step 2 - Identification of Critical Value (1 point):**



**Step 3 - Value of Test Statistic (1 points):**

**Step 4 - Give p-value (1 points): (had to divide the p-value by 2 for a 1-sided test)**

**Step 5 - Decision (0 points): Reject**

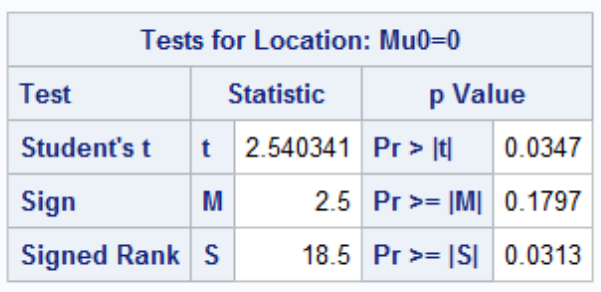
**Step 6 - Conclusion (1 point): There is evidence at the level of significance () that mean difference in time to finish the puzzle was greater than 0.**

**SAS Output:**

proc univariate data = autism;

var diff;

run;



### Part E (5 points)

Verify your calculations in R. Simply cut and paste your code and relevant output.

t.test(autism$Before, autism$After, paired=T, alternative='greater')

##   
## Paired t-test  
##   
## data: autism$Before and autism$After  
## t = 2.5403, df = 8, p-value = 0.01735  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 4.913201 Inf  
## sample estimates:  
## mean of the differences   
## 18.33333

### Part F (5 points)

Use your data from above to construct a “complete analysis” of the test that you feel is most appropriate to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle. This is simply formatting your results. You should be able to cut and paste most of the work from above.

*Note: the only thing you need to add here is the statement of the problem and the assumptions for the test that you think is most appropriate. You could make an argument for any of the tests, so you may give yourself full credit provided you get the assumptions correct. The statement of the problem and assumptions given below correspond to the signed-rank test.*

**Problem (0.5 points): Test if the median difference in time is greater than 0 (i.e. the median of all Before - After values is > 0). More generally, test whether yoga was effective in reducing time to finish the puzzle.**

**Assumptions (0.5 points): The subjects are independent, a random sample from a fixed population, and the differences are symmetric.**

*Note: everything else will simply be a copy/paste of the test from whichever test you choose. You may give yourself full credit as long as you’ve stated the problem and the assumptions.*

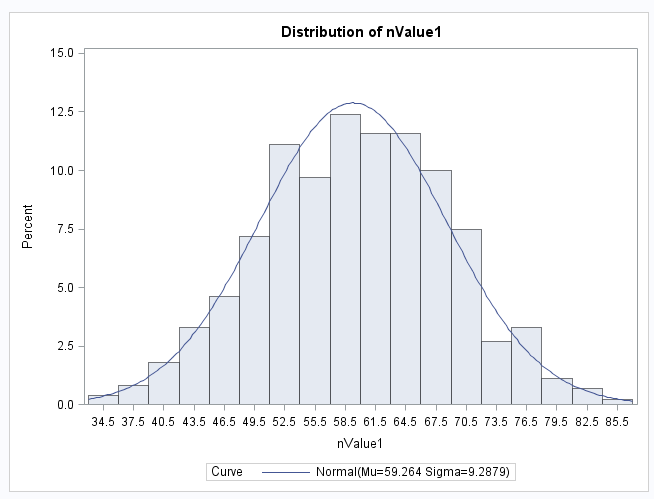
## Bonus (+5 points total)

Using our permutation test SAS code that we have used in prior HWs, do the following:

### Part A (+2.5 points)

Build the permutation distribution for the rank sum statistic for the Trauma data used above. Use 5000 permutations. Use SAS to fit/overlay a normal curve to the resulting histogram. Compare the mean and standard deviation of this normal curve that was fit to the permutation/randomization distribution to the mu and sigma you found earlier in the homework.

**The mean of the normal distribution here was 59 and the standard deviation was about 9.3. The mean of the normal approximation found through theory in the prior problem was 56 with a standard deviation of about 8.6. The first one was found by fitting the normal curve to the permutation test while the second one came from theory. We don’t expect them to be identical, although we do expect them to be close.**



### Part B (+2.5 points)

Compare the one-sided p-value found in this permutation distribution with the one found in prior questions.

|  |  |
| --- | --- |
| HINT: Don’t mind the highlight; the whole thing is the hint. You will need to work code similar to what is to the right into the permutation test SAS code we used before (In place of Proc ttest.) You will also have to do some research on how to get your hands on the sum of the ranks statistic (a good start is to print the outnpar data set!). |  |

**The one-sided p-value found in the prior question was .0016 while the one from this test was . Your answers for the p-value from the permutation test will vary. SAS may crash for the 5000 permutation version. It appears to be one of the limitations of SAS.**

# Homework #5 Solutions

## Question 1 (35 points total)

Simply Answer Question 25 on pg. 147 from the Statistical Sleuth (read it!):

Plot the raw data, and also plot the data after a log transform. After a log transform, do the data satisfy the assumptions better? The data is in ex0525.csv or ex0525.xlsx. Perform this analysis in SAS.

Regardless of whether the assumptions of the original data or log transformed data are met, please include a **complete analysis** on the **log transformed** data.

1. State the Problem
2. Address the assumptions. Comment on each assumption (Use the visual test, as the Brown-Forsythe test will be overpowered due to the large sample size. This simply means that it is able to detect very small effect sizes-here, differences in standard deviations-which may not be big enough to practically affect the test). Comment on your thoughts of the assumptions, but, in the end, assume there is not enough visual evidence to suggest the standard deviations of the log transformed data are different.
3. Conduct the Test (an example is in the Unit 5 PowerPoint).
4. Write a conclusion (an example is in the Unit 5 PowerPoint).
5. State the Scope. (Can we generalize to the entire population or just the sample that was taken? Is there a causal relationship present?)

*Looking to the future! This is not an additional problem. Just FYI: The next step will be to look at these pairwise if we reject the to discover WHICH pairs have evidence of different means/medians.*

ADDITIONAL THINGS TO INCLUDE (for the logged data):

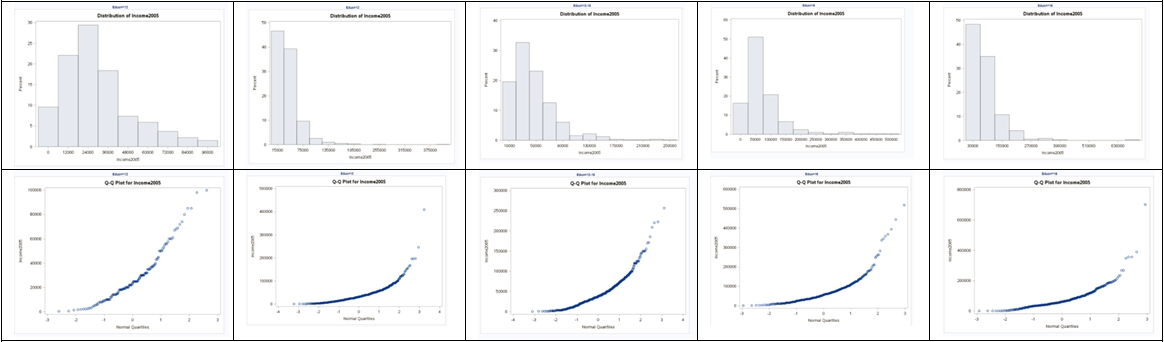
1. Please also identify .
2. Also specify the mean square error and how many degrees of freedom were used to estimate it.
3. Provide the code to perform the ANOVA in R and a screen shot of the output.

**Problem (1 point): How strong is the evidence that at least one of the five population distributions of education level has a different mean 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.**

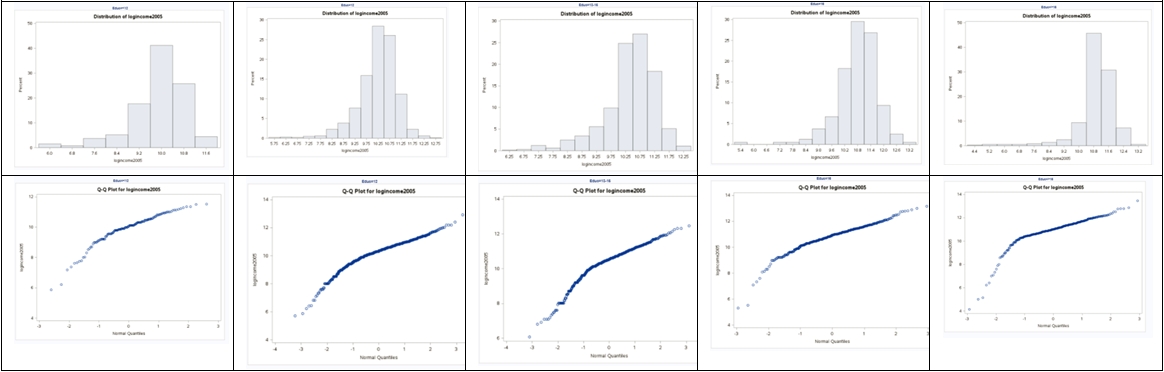
**Normality (3 points): The histograms and QQ plots below (of original data) appear to each show strong evidence of right skew, and thus provide evidence against coming from a normal distribution. This is not unexpected, as income data is often right skewed. However, each group has a sample size greater 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;



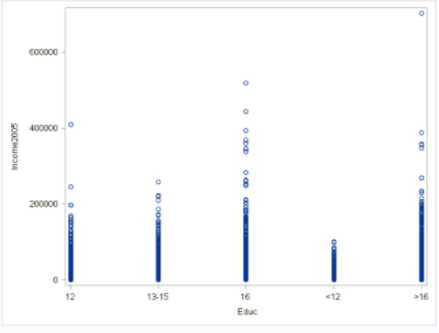
\*Perform a log transform;  
data incomedata;  
set incomedata;  
logincome2005 = log(income2005);  
run;

\*To address ANOVA assumptions on log transformed data with histograms and QQ plots;  
proc univariate data = incomedata;  
by educ;  
histogram logincome2005;  
qqplot logincome2005;  
run;

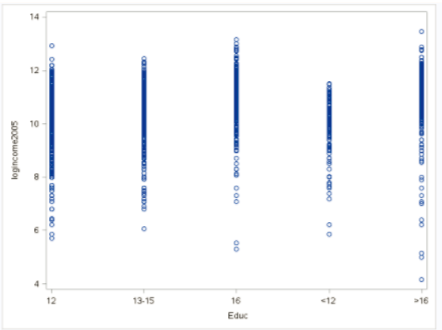


**Equal Standard Deviations (3 points): It appears that the original data shows evidence against equal standard deviations in the scatter plot. We were given in the problem that we are able to assume that the standard deviations between the groups are equal (homoscedasticity) for log transformed data. This is a safe assumption visually. Small deviations in sd will have less effect on the test than larger deviations. Remember, all models are wrong, but some are useful. [George Box]**

\*To address ANOVA assumptions on original data with scatter plots;  
proc sgplot data = incomedata;  
scatter x= educ y = income2005;  
run;



\*To address ANOVA assumptions on log transformed data with scatter plots;  
proc sgplot data = incomedata;  
scatter x= educ y = logincome2005;  
run;



**Independence (3 points): We will assume the data are independent, both between and within groups, and proceed with the ANOVA to test for differences in mean log income (median income) between the five levels of education. Note: this is risky assumption, as it turns out the sample is a random sample of households in which all members of the household were recruited into the survey. More pertinent information on the sample can be found in the first paragraph of the “Sampling Procedures” section that can be found by following this link:** [**https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/sample-design-screening-process**](https://www.nlsinfo.org/content/cohorts/nlsy79/intro-to-the-sample/sample-design-screening-process)**.**

**Step 1 - Hypotheses (2 points):**

**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).**

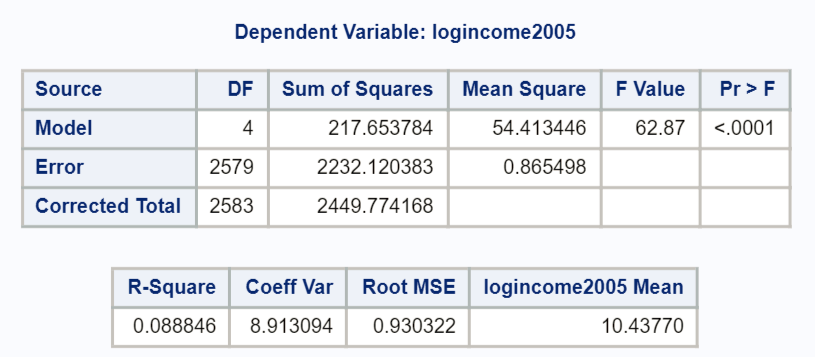
**Step 3 - Value of Test Statistic (2 points):**

**Step 4 - Give p-value (2 points):**

**Step 5 - Decision (2 points): Reject**

**Step 6 - Conclusion (5 points): 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;

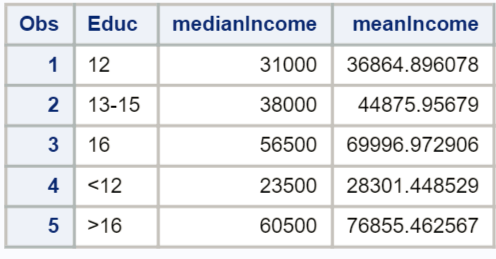


##Here is how to answer the problem using R  
##Read in the data, note your directory will be different  
  
edu <- read.csv('C:/Users/Charles/Documents/SMU/Online Teaching/MSDS 6371 - Statistical Foundations for Data Science/UNIT 5/HW/ex0525.csv')  
  
edu$log.income <- log(edu$Income2005)  
  
edu.anova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

**To answer the second part of the question in the textbook, we will look at the difference of means here given the table, although a difference in medians could also be provided. Several SAS procedures will produce means and medians, including proc univariate (as coded above).**

\*There are many possibilities to produce means and medians… you can get even fancier code that will compute the percentage difference between each consecutive jump in education;  
proc means data = incomedata nway;  
class educ;  
var income2005;  
output out = incomesummary median = medianIncome mean = meanIncome;  
run;  
proc print data = incomesummary;  
var Educ medianIncome meanIncome;  
run;



* **The differences are as follows:**  
  + **(1 point) The estimated difference in mean income between those with less than a high school education and those with a high school degree only is $8,563.45 ($36,865.90 - $28,301.45), which is a 30.3% ($8,563.45/$28,301.45) increase in means from less than high school to only high school levels of education. The estimated difference in median incomes between those with less than a high school education and those with a high school degree only is $7,500 ($31,000 - $23,500), which is a 31.9% ($7,500/$23,500) increase in medians from less than high school to only high school levels of education.**
  + **(1 point) The estimated difference in mean income between those with only a high school education and those with only some college is $8,011, with a 21.7% increase in means from only high school to some college only. The estimated difference in median income between those with only a high school education and those with only some college is $7,000, with a 22.6% increase in medians from only high school to some college only.**
  + **(1 point) The estimated difference in mean income between those with only some college and those with only a college degree $25,121, with a 56.0% increase in means from only some college to only a college degree. The estimated difference in median income between those with only some college and those with only a college degree $18,500, with a 48.7% increase in medians from only some college to only a college degree.**
  + **(1 point) The estimated difference in mean income between those with only a college degree and those with more than 16 years of education (more than a college degree) is $6,858, with a 9.8% increase in means from only college to more than college. The estimated difference in median income between those with only a college degree and those with more than 16 years of education (more than a college degree) is $4,000, with a 7.1% increase in medians from only college to more than college.**

**Scope of Inference (5 points): this is an observational study, and thus, we cannot assign causal inference to this relationship. (Education does not necessarily cause the difference in income.) The NLSY is a random sample of households and, thus, is a random sample but not a simple random sample of subjects in the desired population. Inference can be generalized to the population of areas sampled in the United States, although one should be wary of the standard deviations and standard errors estimated here. Cluster sampling of households was employed, which introduces dependency/correlation at the cluster (household) level. We will address this adjustment/calculation later in the Sampling Course.**

**(1 points)**   
**(2 points)**

## Question 2 (30 points total)

Use an extra sum of squares F-test (BYOA: Build Your Own ANOVA!) to use all the data (to increase the degrees of freedom and thus the power of the test) to compare only the bachelor’s degree group (16) income to the more than bachelor’s degree group (>16) income. Show your final ANOVA table and your 6-step complete analysis. You will need to assume that the standard deviations of the log-transformed data are again equal to proceed here. A two-sample t-test between these two groups (assuming equal standard deviations on logged data) yields a p-value of **0.1648** (try it!), but it only uses 778 degrees of freedom (from a pooled t-test). Make note again of how many degrees of freedom were used to estimate the pooled standard deviation in your extra sum of squares test. You may use SAS or R.

**Problem (1 point): Test whether there is a difference in income between those with a bachelor’s degree and those with more than a bachelor’s degree.**

*Note: alternatively, you could list the full and reduced model as is stated below:*

**Full Model:**   
**Reduced Model:**

**Step 1 - Hypotheses (2 points):**

**Step 2 - Identification of Critical Value: You may skip step 2 (critical value) in ANOVA settings, although the correct critical value is .**

**Step 3 - Value of Test Statistic (0 points, this is built in to the BYOA):**

**Step 4 - Give p-value (0 points, this is built in to the BYOA):**

**Step 5 - Decision (2 points): Fail to Reject**

**Calculation of the 95% CI:**

**Step 6 - Conclusion (5 points): There is not sufficient evidence to suggest that the median incomes of the bachelor’s degree group (16) and post bachelor’s degree group (>16) are different. A 95% confidence interval for this difference is . Now, a 95% CI for the multiplicative change in median incomes between these two groups is . Note that the multiplicative change of 1 is within our interval, consistent with the decision to fail to reject the null hypothesis. We will get to this confidence interval in Chapter 6, but see if you can reverse engineer this!**

(The confidence interval above is not necessary for full credit, as it is covered later in chapter 6.)

**Scope of Inference (5 points): As the results are not significant, we do not need to discuss assigning a causal inference in this relationship. The NLSY is a random sample of households and, thus, is a random sample but not a simple random sample of subjects in the desired population. Inference can be generalized to the population of areas sampled in the United States, although one should be wary of the standard deviations and standard errors estimated here. Cluster sampling of households was employed, which introduces dependency / correlation at the cluster (household) level. We will address this adjustment/calculation later in the Sampling Course.**

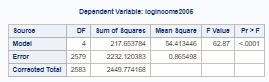
**DF Difference (5 points): note that in using the Extra Sums of squares, we were able to have 2579 degrees of freedom to estimate the pooled SD rather than only 778.**

**BYOA Table (10 points):**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 1 | 1.9786 | 1.9786 | 2.286 | 0.13067 |
| Error | 2579 | 2232.120383 | 0.865498 |  |  |
| Total | 2580 | 2234.099010 |  |  |  |

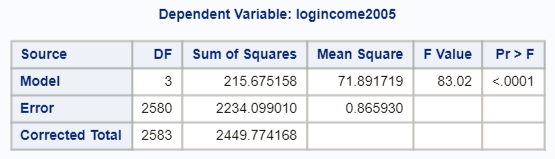
**SAS Code:**

\*To perform ANOVA on log transformed data (full model);  
proc glm data = incomedata;  
class educ;  
model logincome2005 = educ;  
run;



\*Recode the data so that college and more than college are in the same group;  
Data incomedata;  
set incomedata;  
recodededuc = educ;  
if (educ = ‘>16’ | educ = ‘16’) then recodededuc = ‘16 or >16’;  
run;

\*To perform ANOVA on log transformed data (reduced model);  
proc glm data = incomedata;  
class recodededuc;  
model logincome2005 = recodededuc;  
run;



 SAS Code:

\*To find P-value from an F statistic and degrees of freedom (numerator and denominator) for building our own ANOVA;

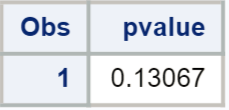
data pval;

pvalue = 1-probf(2.286, 1, 2579);

run;

proc print data = pval;

run;



\*To find critical value for an F test at alpha = 0.05;

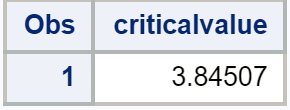
data critval;

criticalvalue = quantile("F", .95, 1, 2579);

run;

proc print data = critval;

run;



##Here is how to answer the problem using R  
##Combine the 16 and >16 groups  
edu$group.reduced <- ifelse(edu$Educ %in% c('>16', '16'), '16 plus',  
edu$Educ)  
  
##Full Model  
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

##Reduced Model  
edu.anova2 <- aov(log.income ~ group.reduced, data=edu)  
summary(edu.anova2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## group.reduced 3 215.7 71.89 83.02 <2e-16 \*\*\*  
## Residuals 2580 2234.1 0.87   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##Determine the critical value  
qf(0.95, 1, 2579)

## [1] 3.845067

##Determine the p-value  
pf(2.286, 1, 2579, lower.tail=F)

## [1] 0.1306685

## Question 3 (30 points total)

Now, suppose that you cannot assume the standard deviations are the same (for both the original or log transformed data). Conduct another complete analysis of the question in Chapter 5, problem 25 in Statistical Sleuth. Answer the question, “How strong is the evidence that at least one of the five population distributions (corresponding to the different years of education) is different from the others?” This question should be answered in at least 1 or 2 sentences after providing a **complete analysis** without the assumption of equal standard deviations for the logged data (or for the original data). Perform the test in SAS or R.

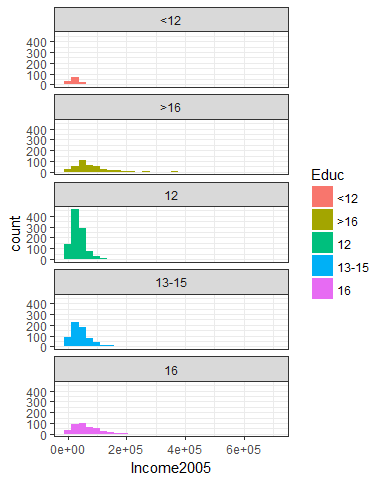
**Problem (3 points): How strong is the evidence that at least one of the five population distributions of education level has a different mean (or median, depending on your choice of test) income than any of the others?**

**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 independent between each group.**

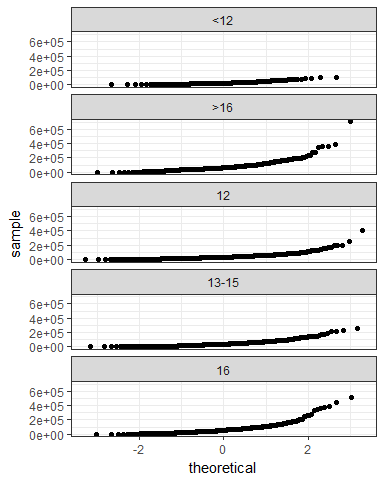
**Normality (3 points): this is exactly the same as in question 1 (plots below are generated using R, for reference).**

##There are multiple ways to do this, but the easiest is to use  
##the very popular ggplot package. First, install if necessary:  
##install.pacakges(ggplot2)  
  
library(ggplot2)   
  
##Histograms, Raw Data  
ggplot(data = edu, aes(x=Income2005)) +  
geom\_histogram(aes(fill=Educ)) +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

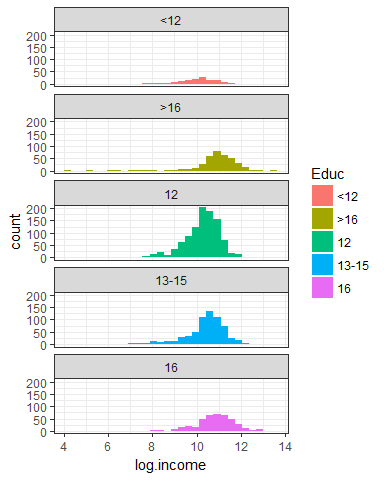


##QQ Plots, Raw Data  
ggplot(data = edu, aes(sample=Income2005)) +  
stat\_qq() +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

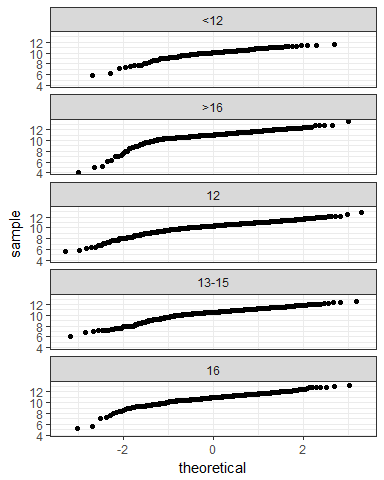


##Histograms, Logged Data  
ggplot(data = edu, aes(x=log.income)) +  
geom\_histogram(aes(fill=Educ)) +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

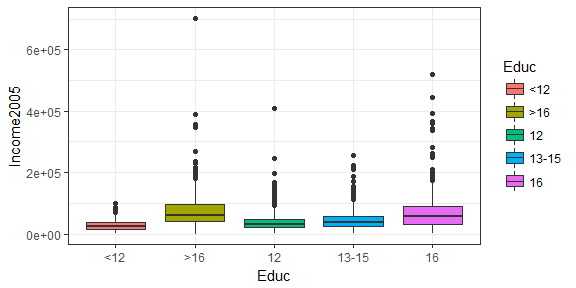


##QQ Plots, Logged Data  
ggplot(data = edu, aes(sample=log.income)) +  
stat\_qq() +  
facet\_wrap( ~ Educ, ncol=1) +  
theme\_bw()

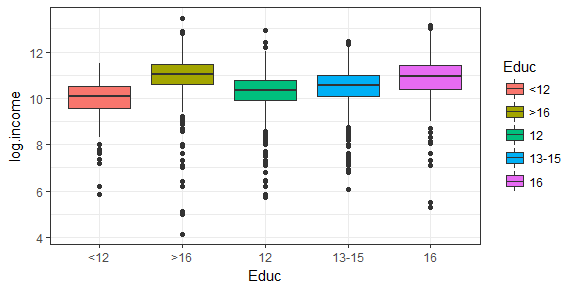


**Equal Standard Deviations (3 points): again, the plots are the same as in question 1 (plots below are generated using R, for reference). It appears that the original data show evidence against equal standard deviations in the scatter plot. We were given in the problem that we are NOT able to assume that the standard deviations between the groups are equal (homoscedasticity) for log transformed data.**

##Boxplots, Raw Data  
ggplot(data = edu, aes(x=Educ, y=Income2005)) +  
geom\_boxplot(aes(fill=Educ)) +  
theme\_bw()



##Boxplots, Logged Data  
ggplot(data = edu, aes(x=Educ, y=log.income)) +  
geom\_boxplot(aes(fill=Educ)) +  
theme\_bw()



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

**Because the CLT applies but the standard deviation assumption is violated, we have a few options. First, a Welch’s ANOVA can be used on the original data, or the nonparametric Kruskal-Wallis test can be used on the original data. Technically, a Welch’s ANOVA could be used on log transformed data, but that makes little sense if we still can’t overcome the equal standard deviation assumption violation by logging the data. Logging would only “fix” the normality issue, and with such large sample sizes, the sample means are likely to be normally distributed via the Central Limit Theorem. Still, it is a judgement call.**

**Step 1 - Hypotheses (2 points):**

**Option 1: Welch’s ANOVA**  
 **All mean incomes are the same across education levels.**  
 **At least one pair of income means are different between education levels.**

**Option 2: Kruskal-Wallis**  
 **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: May be omitted here.**

**Step 3 - Value of Test Statistic (2 points): For Welch’s ANOVA: ; for Kruskal-Wallis:**

**Step 4 - Give p-value (2 points): for both Welch’s ANOVA and Kruskal-Wallis**

**Step 5 - Decision (2 point): Reject**

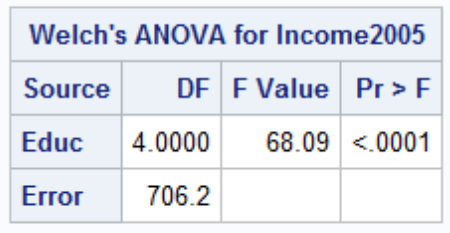
**Step 6 - Conclusion (5 points):**

**Welch’s ANOVA: At the level of significance, there is strong evidence () that the mean incomes between at least 1 pair of education levels are different.**

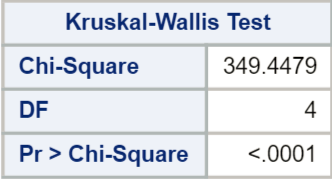
**Kruskal-Wallis: At the level of significance, there is strong evidence () that the median incomes between at least 1 pair of education levels are different.**

**Scope of Inference (5 points): This is an observational study, and thus, we cannot assign causal inference to this relationship (education does not necessarily *cause* the difference in income). The NLSY is a random sample of households, and thus is a random sample but not a simple random sample of subjects in the desired population. Inference can be generalized to the population of areas sampled in the United States, although one should be wary of the standard deviations and standard errors estimated here. Cluster sampling of households was employed, which introduces dependence/correlation at the cluster (household) level. We will address this adjustment/calculation later in the Sampling Course.**

\*For Welch’s ANOVA on original data (code for assumptions was run in Q1);  
proc glm data = incomedata;  
class educ;  
model income2005 = educ;  
means educ / hovtest = bf welch;  
run;



\*For Kruskal-Wallis test on original data;  
proc npar1way data = incomedata;  
class educ;  
var income2005;  
run;



##Using R  
##Welch's ANOVA  
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

##Kruskal-Wallis  
kruskal.test(Income2005 ~ Educ, data=edu)

##   
## Kruskal-Wallis rank sum test  
##   
## data: Income2005 by Educ  
## Kruskal-Wallis chi-squared = 349.45, df = 4, p-value < 2.2e-16

# 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**

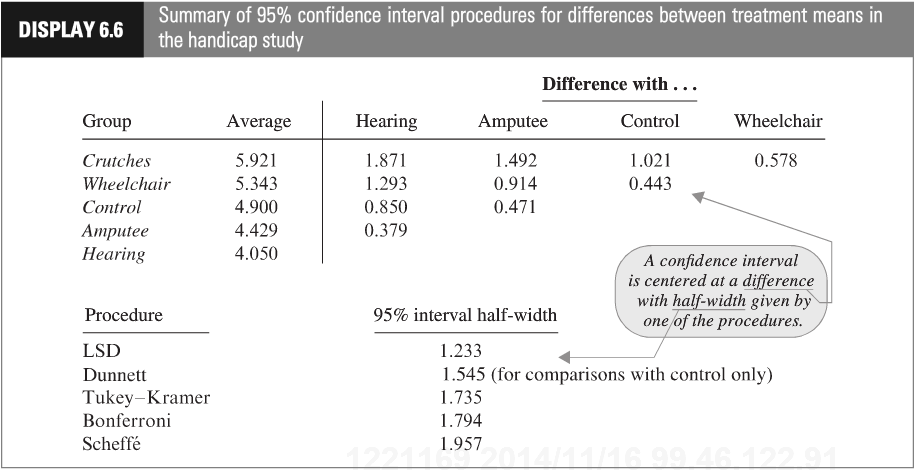
**95% CI for : .**

**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.

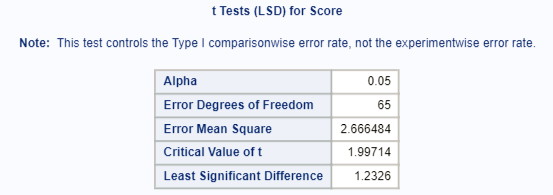
**SAS:**

\*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;

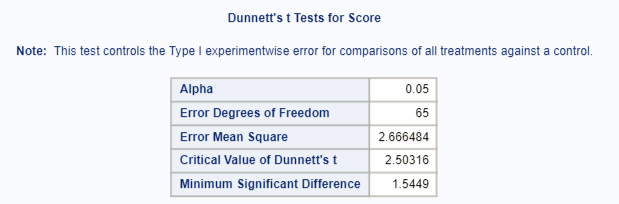
**We are fortunate that SAS provides the 95% CI half-width in the “least significant difference” box. For each confidence interval, we could also find the half-width using the critical value and standard error. The standard error for the difference in means of two of any two of these groups:**

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

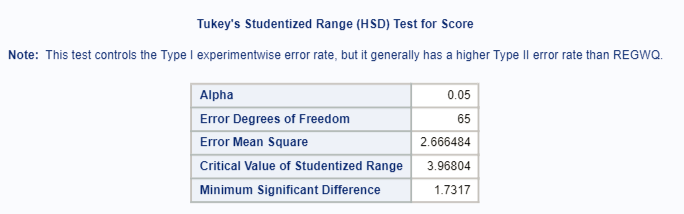
**LSD (t): Multiplier = 1.99714, 95% CI half-width =**



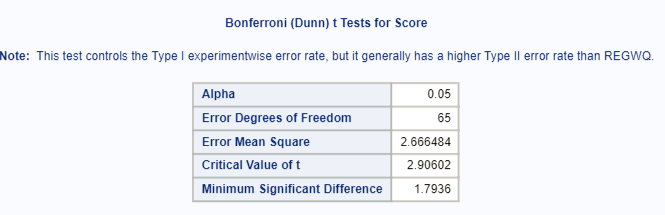
**Dunnett: Multiplier = 2.50316, 95% CI half-width =**



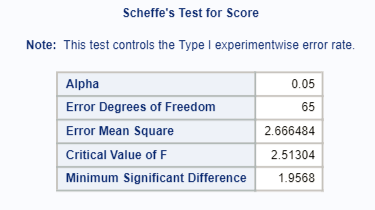
**Tukey: Multiplier = , 95% CI half-width =**



**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;

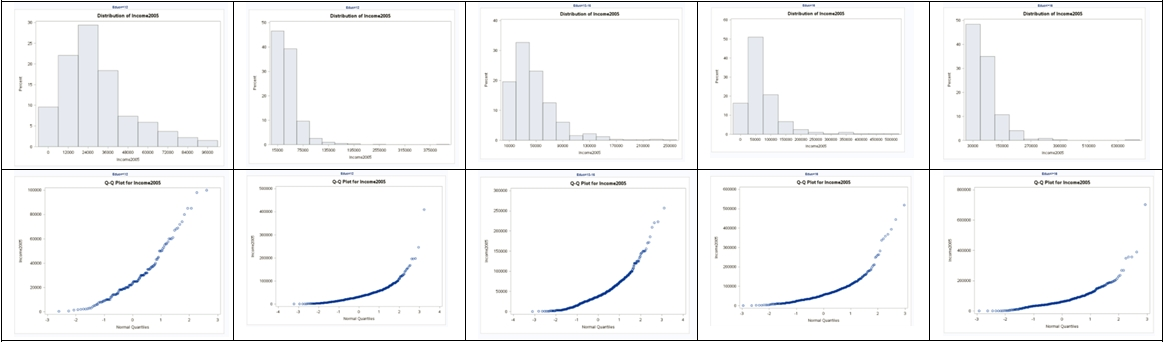
proc univariate data = incomedata;

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;

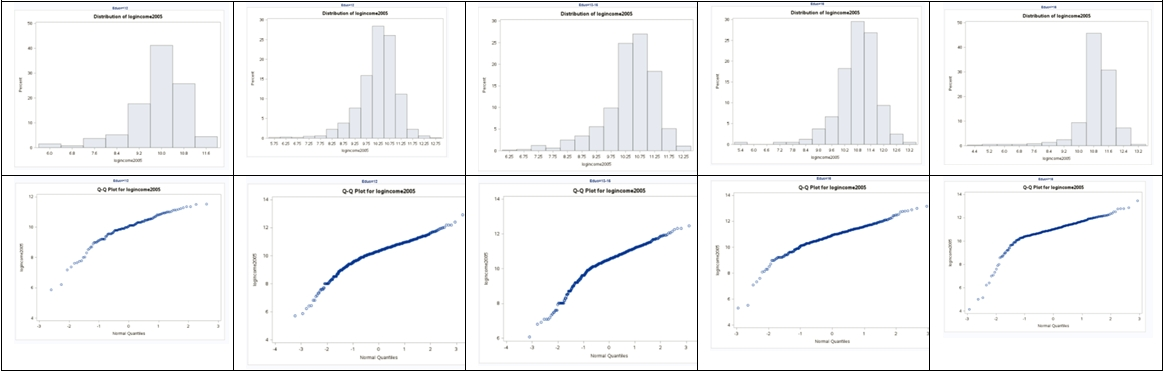
proc univariate data = incomedata;

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.**

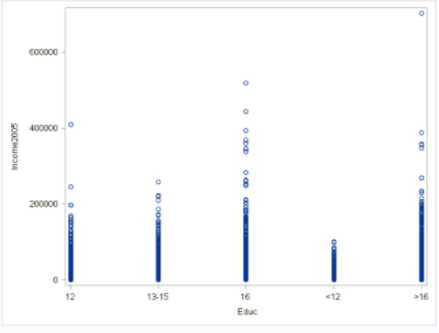
\*To address ANOVA assumptions on original data with scatter plots;

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);

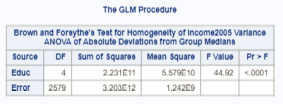
proc glm data = incomedata;

class educ;

model income2005 = educ;

means educ / hovtest = bf;

run;



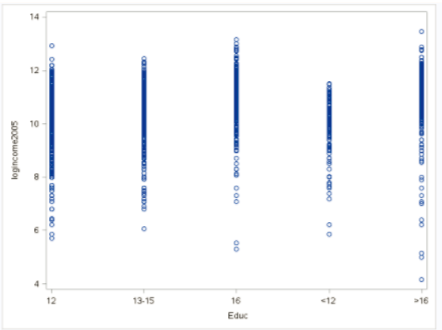
\*To address ANOVA assumptions on log transformed data with scatter plots;

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);

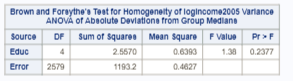
proc glm data = incomedata;

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).**

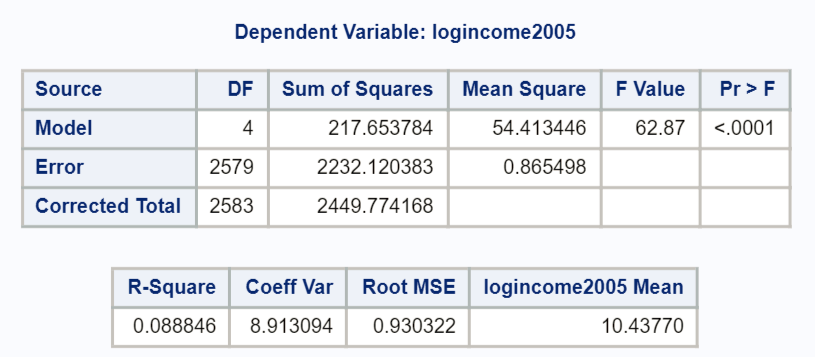
\*To perform ANOVA on log transformed data;

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;

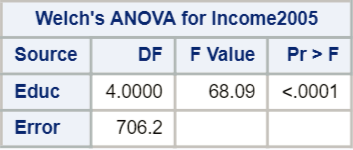
proc glm data = incomedata;

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

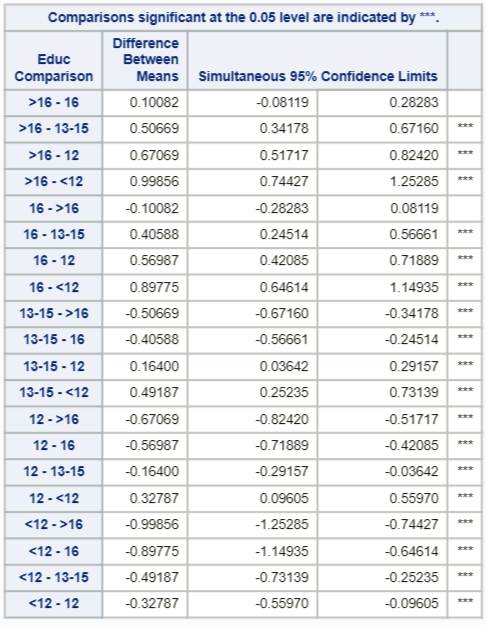
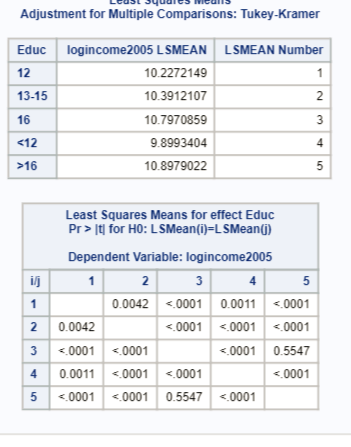
**Fortunately, Welch’s ANOVA on original data and pure ANOVA on log-transformed data are in agreement that at least one group is different from the others.**

**Now, onto multiple comparisons.**

### Part A (15 points)

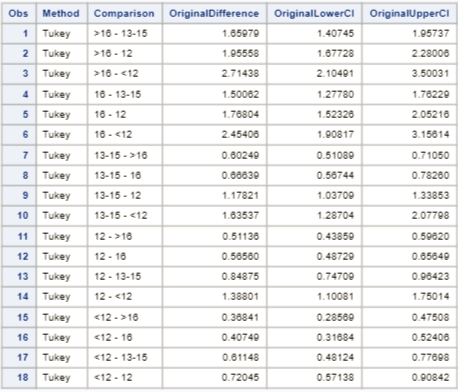
**Note that Tukey (Tukey-Kramer) requires normality and equal standard deviations, so we must perform the analysis on the logged data and make inferences accordingly. Confidence intervals and p-values for differences in means of logged data.**

*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)

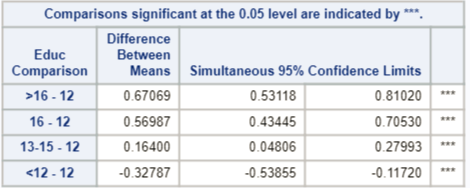
## diff lwr upr p adj  
## >16-<12 2.7143750 2.1049084 3.5003099 1.000000  
## 12-<12 1.3880148 1.1008146 1.7501448 1.001096  
## 13-15-<12 1.6353720 1.2870409 2.0779771 1.000000  
## 16-<12 2.4540642 1.9081654 3.1561368 1.000000  
## 12->16 0.5113570 0.4385857 0.5962028 1.000000  
## 13-15->16 0.6024856 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)

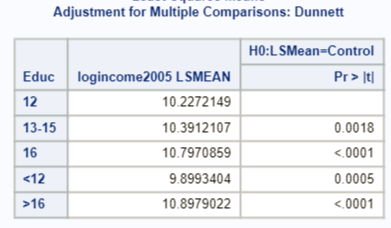
**Dunnett’s requires normality, equal standard deviations, and independence. So, we will proceed with Dunnett’s comparisons on the log transformed data.**

*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).*

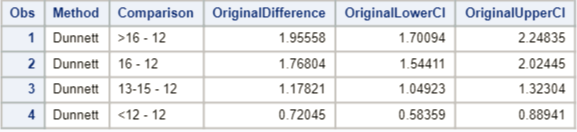
**95% confidence intervals on differences in means of logged data**



**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;

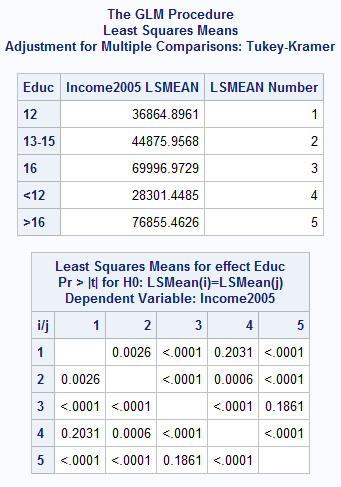
class educ;

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;

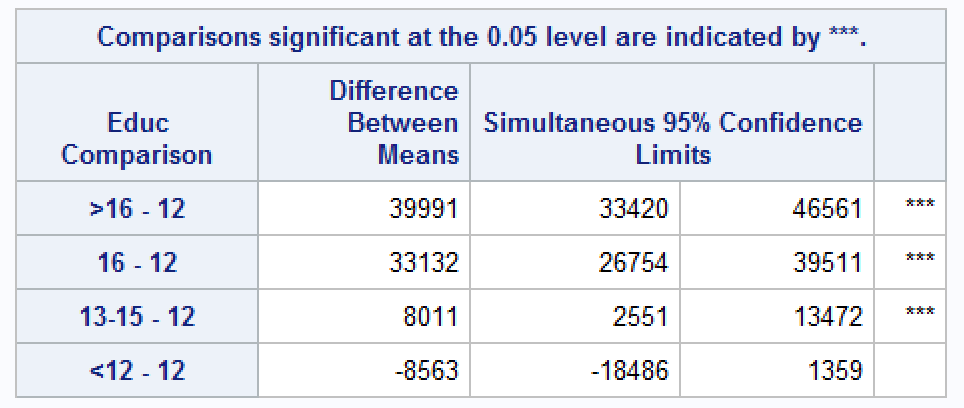
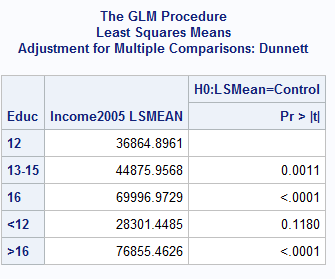
class educ;

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.**

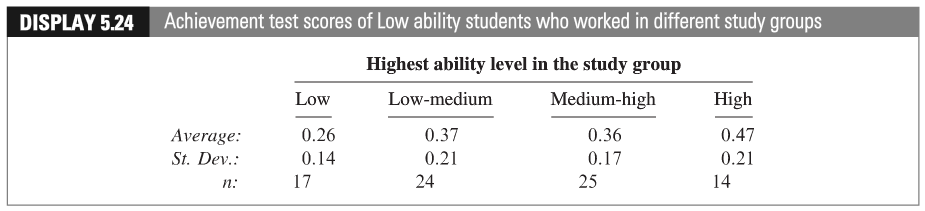
**We would certainly hope that a high school education is significant! But we shouldn’t let this idea drive our interpretation of the data.**

##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**

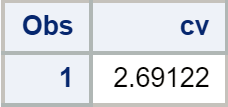
**The CI doesn’t include 0, which means low-ability students’ performances increase with the ability of the best students in the group (contrast is significantly different from 0).**

### 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**

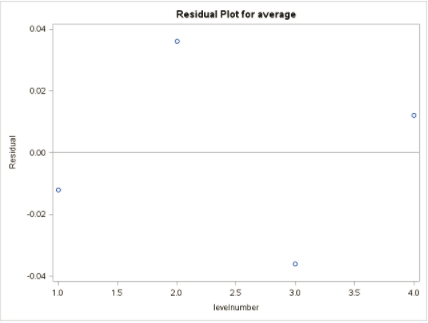
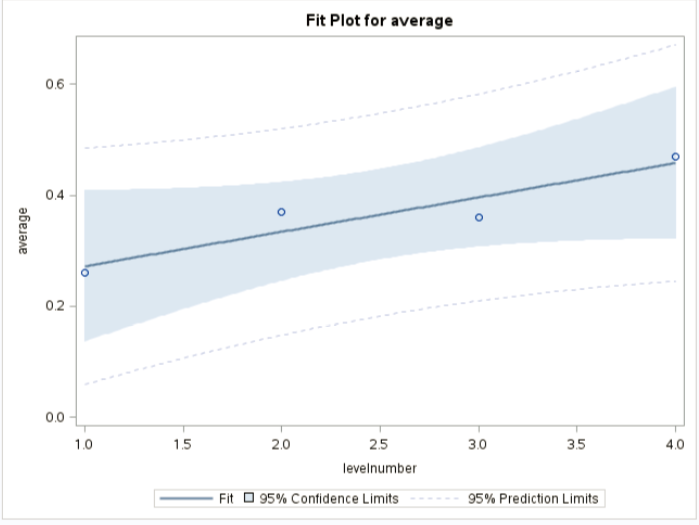
|  |  |  |  |
| --- | --- | --- | --- |
| Group | Low | Low-Medium | Medium-High |
| High | [-0.004, 0.204] | [-0.008, 0.168] | [0.055, 0.205] |
| Medium-High | [-0.128, 0.068] | [-0.131, 0.031] |  |
| Low-Medium | [-0.088, 0.128] |  |  |
| Low |  |  |  |

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

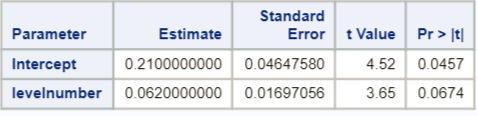
### Part C (+1 point)

\*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.**