UNIT 4 HW

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.

1. 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.
2. 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 industries assertion) that logging actually increases the percentage of seedlings lost from 2004 to 2005.

a. Perform a complete analysis using a rank sum test in SAS. (Logging data).

***Is there sufficient evidence to suggest that logging increases the percentage of seedlings lost in forest recovery from a fire?***

***Independence:***

***The problem states the samples are along a random located transect pattern and they returned a year later for the second observation.***

***There may be confounding variables that occurred over the course of a year however for this exercise we assume they are independent.***

***Normality:***

***Because this is a rank sum test no specific distributional assumptions required.***

***The sorted data below shows that the sample sizes are sufficient and there are no ties.***

Procsort data=import;

By descending PercentLost;

proc print data=import;



***Step 1:***

***: The percentage of seedlings lost in logged plots is equal to the percentage of seedlings lost in the unlogged plots***

***: The percentage of seedlings lost in logged plots is the greater than the percentage of seedlings lost in the unlogged plots***

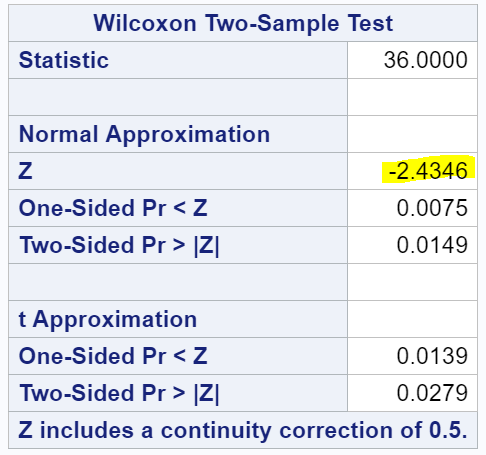
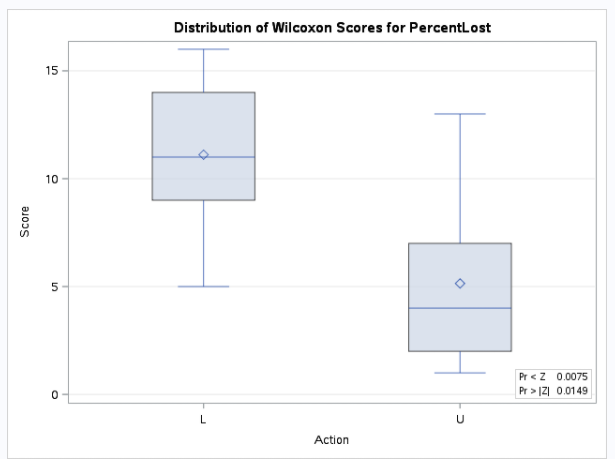
***Step 2: Normal Approximation Critical Value Z= -2.43***

**proc** **npar1way** data=WORK.IMPORT wilcoxon alpha=**.05**;

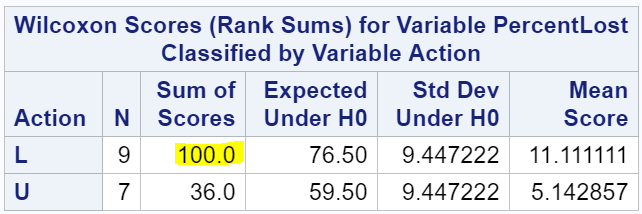
class Action;

var PercentLost;

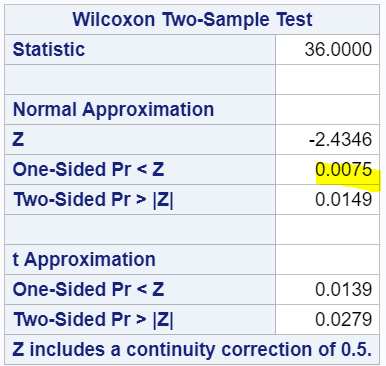
exact HL;

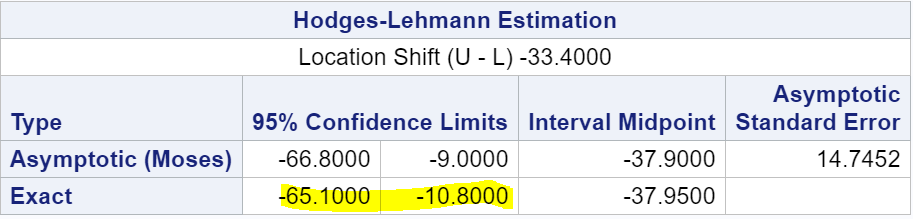
 

***Step 3:***



***Step 4:***





***Step 5: Reject***

***Step 6: The evidence suggests that the percentage of seedlings lost in logged plots is the greater than the percentage of seedlings lost in the unlogged plots (one-sided, normal approximation w/ Continuity Correction p-value = 0.0075, 95% confidence interval [-65.1000, -10.8000] from the rank-sum test).***

* 1. 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.

##### **Wilcoxon rank sum test**

Logged = c(85.6 ,85.4 ,75.5 ,53.1 ,46.7 ,45 ,43.2 ,40.8 ,18.2)  
Notlogged = c(56.1 ,34.2 ,23.6 ,18.1 ,13.3 ,-8.1 ,-20.1)  
wilcox.test(Logged,Notlogged, correct=TRUE)

## data: Logged and Notlogged  
## W = 55, p-value = 0.01154  
## alternative hypothesis: true location shift is not equal to 0

##### **Calculate Confidence Interval**

sum(rank(c(Logged,Notlogged))[1:9])

## [1] 100

sum(rank(c(Logged,Notlogged))[10:16]) #Sum ranks not Logged

## [1] 36

***I’m thinking SAS may be inaccurate if we assume the instructions in the URL provided are accurate. The book states in regard to the Rank Sum Test*** *“Its drawbacks are that associated confidence intervals are not computed by most of the statistical computer packages and that it does not easily extend to more complicated situations.”*

1. 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).
   1. 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.

***Test the claim that the distribution of incomes for those with 12 years of education is less than is less than the distribution for those with 16 years of education.***

***Independence:***

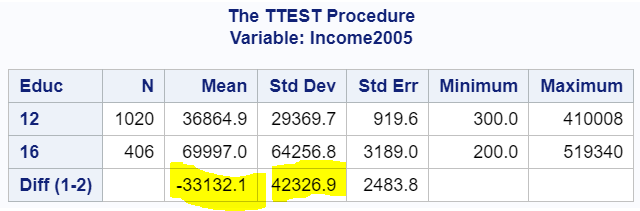
***The problem states the samples are from interviews. It is likely that the subjects volunteered therefore we have confounding variables.***

***Though it does say this is a subset; there is not enough information about how the samples were gathered to consider them random.***

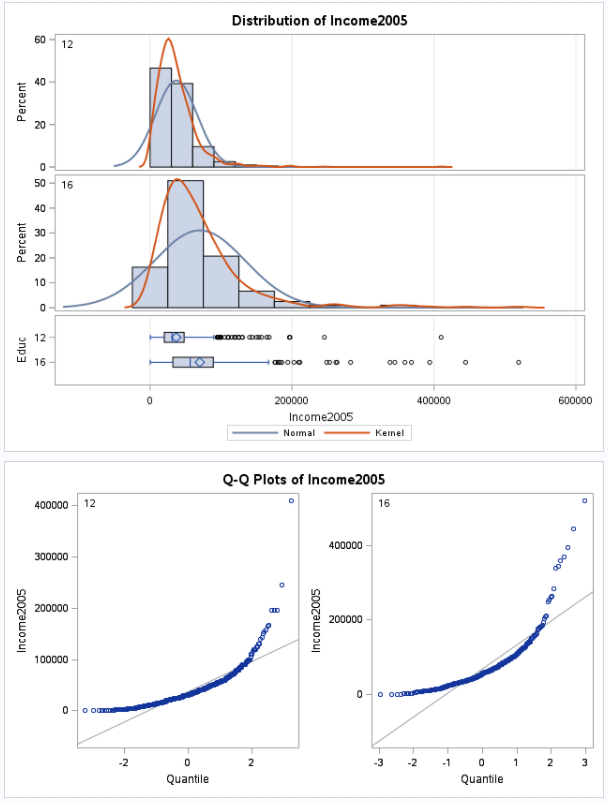
***For purposes of this homework we will treat this as an observational study.***

***Normality:***

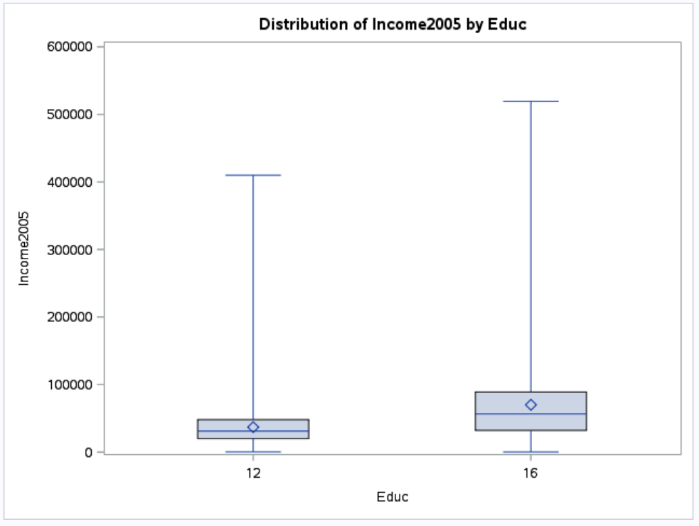
***Our sample sizes are 1020 and 406, there are significant differences in the sample means and standard deviations.***



***Our histograms have roughly the same shape but QQ plots are showing significant curves***

**

***We have extreme outliers***

**

***Welch’s 2 sample t-test is fairly robust with unequal sample sd assuming the populations are normal. Our populations likely have different population means therefore may be an inadequate summary.***

***We have large sample sizes therefore the Central Limit Theorem should help us.***

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

***1:***

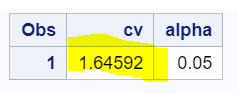
***2: Critical Value = 1.64, df = 1424***

**data** critval;

cv=quantile("T", **.95**, **1424**);

alpha=**.05**;

**proc** **print** data=critval;



***3: Test Statistic t = -9.98 (Satterthwaite)***

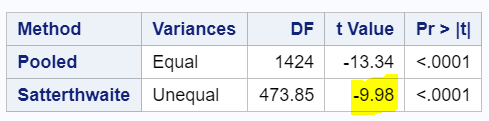
**proc** **Sort** data=IMPORT;

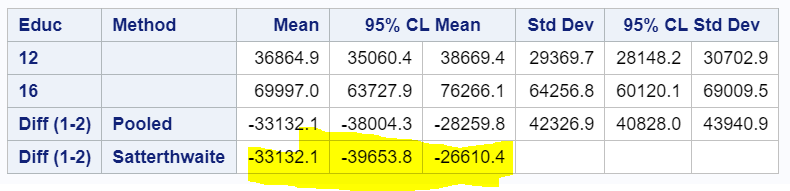
by Educ;

**proc** **ttest** data=WORK.IMPORT sides=**2**;

class Educ;

var Income2005;





***4: p = .0001 (Satterthwaite)***

***5: Reject***

***6: It is estimated that the mean income of subjects in this study with 12 years of education is less than the mean income of subjects with 16 years of education. (p-value = .0001) We are 95% confident that the 12-year educated earn between 3,9654 and 26,610 less than 16 the year educated subjects.***

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

***We cannot make causal inference and must limit our conclusions to differences of the groups in this study.***

* 1. 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.

library(readr)  
EducationData <- read\_csv("EducationData.csv",   
 col\_types = cols(Subject = col\_skip()))  
t.test(EducationData$Income2005 ~ EducationData$Educ, data=EducationData,   
 var.equal=FALSE,  
## Welch Two Sample t-test  
## t = -9.9827, df = 473.85, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -39653.77 -26610.39  
## sample estimates:  
## mean in group 12 mean in group 16   
## 36864.90 69996.97

* 1. 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.

***Though the conclusions from the 2 tests are the same; the difference in sample sizes, sample means, skew and the spread of the 16-year education sample make me favor logarithm transformation.***

***The Welch’s t-test was certainly easier to execute however it may have not have been ideal because we have to assume the populations are normal for it to remain resistant to the difference in the sample standard deviations.***

***The median is a more accurate measure of center when we have extreme outliers; however, reporting a difference in means is probably be easier for the audience to comprehend than the logarithm of the median of the data therefore I would generally favor the Welch’s when possible.***

* 1. 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 necessary here.)

***Problem 20 From Text***

*20. Trauma and Metabolic Expenditure. For the data in Exercise 18 in Chapter 3:*

*(a) Determine the rank transformations for the data.*

*(b) Calculate the rank-sum statistic by hand (taking the trauma patients to be group 1.)*

*(c) Mimic the procedures used in Display 4.5 and Display 4.7 to compute the Z-statistic.*

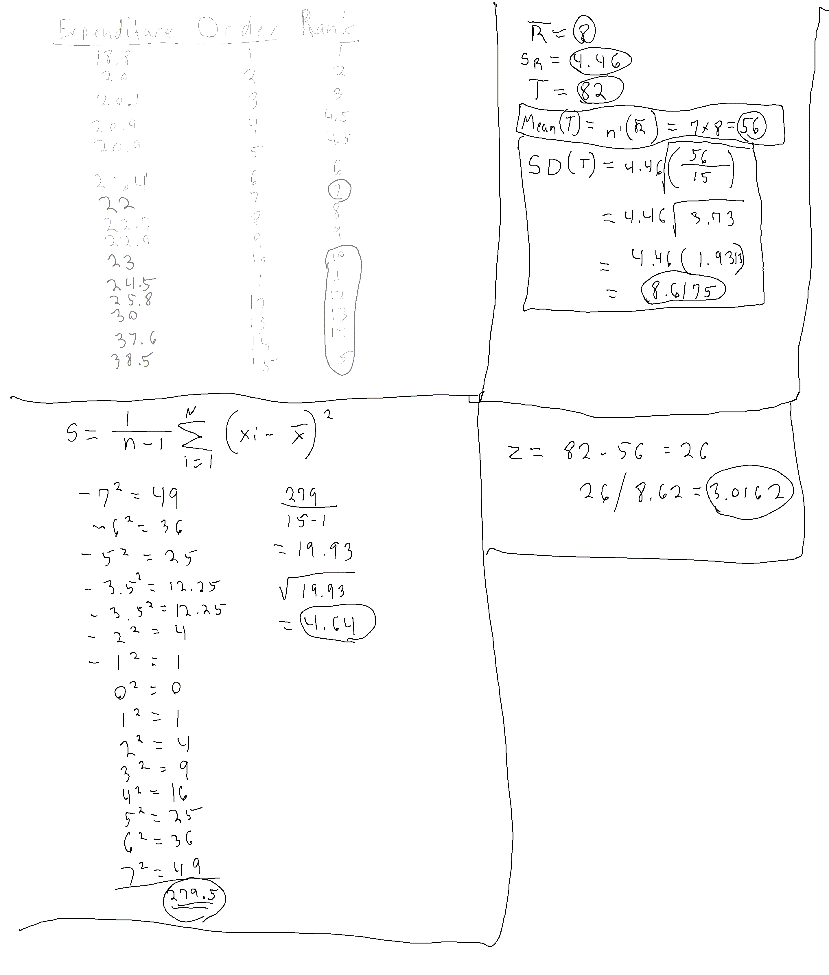
*(d) Find the one-sided p-value as the proportion of a standard normal distribution larger than the observed Z-statistic.*

***Exercise 18 From Text***

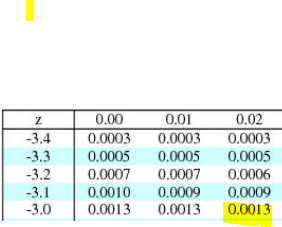
*18. Trauma and Metabolic Expenditure. The following data are metabolic expenditures for eight patients admitted to a hospital for reasons other than trauma and for seven patients admit- ted for multiple fractures (trauma).*

*Metabolic Expenditures (kcal/kg/day)*

*Nontraumapatients: 20.1 22.9 18.8 20.9 20.9 22.7 21.4 20.0 Traumapatients: 38.5 25.8 22.0 23.0 37.6 30.0 24.5*

**

***p-value = .0013***



***Conclusion: The evidence suggests that the metabolic expenditure of trauma patients is the greater than metabolic expenditure of non-trauma patients. (one-sided, normal approximation w/ Continuity Correction p-value = 0.0013 from the rank-sum test)***

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

*Problem 20 From Text*

*21. Trauma and Metabolic Expenditure. Use a statistical computer package to verify the rank-sum and the Z-statistic obtained in Exercise 20. Is the p-value the same? (Does the statistical package use a continuity correction?)*

***The p-value from SAS =.0016 which is slightly different than my manually calculated p-value of .0013.***

***This is likely due to rounding on my part or differences in the way SAS determines the Z-Score (Probably more accurately than the chart I used)***

***SAS does use Continuity Correction***

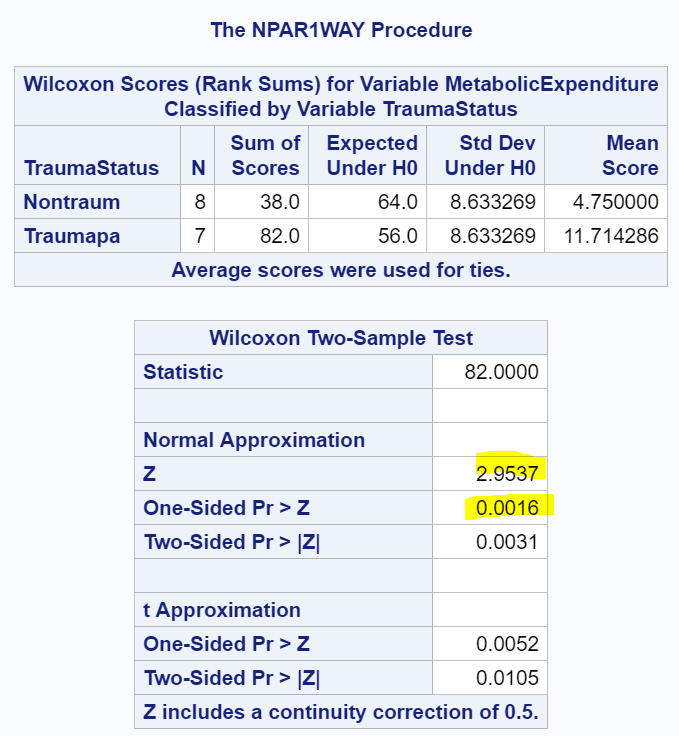
**proc** **npar1way** data=Patients wilcoxon alpha=**.05**;

class TraumaStatus;

var MetabolicExpenditure;

exact HL;

**run**;



* 1. Write up a complete analysis using the information you have gained from A and B to test the claim that the distributions are different.
     1. State the problem.

***Test the claim that the metabolic expenditure in trauma patients greater than that of non-trauma patients***

* + 1. 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.

***1. Independence: The problem does not give much information about the sampling method. We treat this is an observational study and draw inference only on the samples.***

***2. Normality:***

**proc** **univariate** data = Patients;

by TraumaStatus;

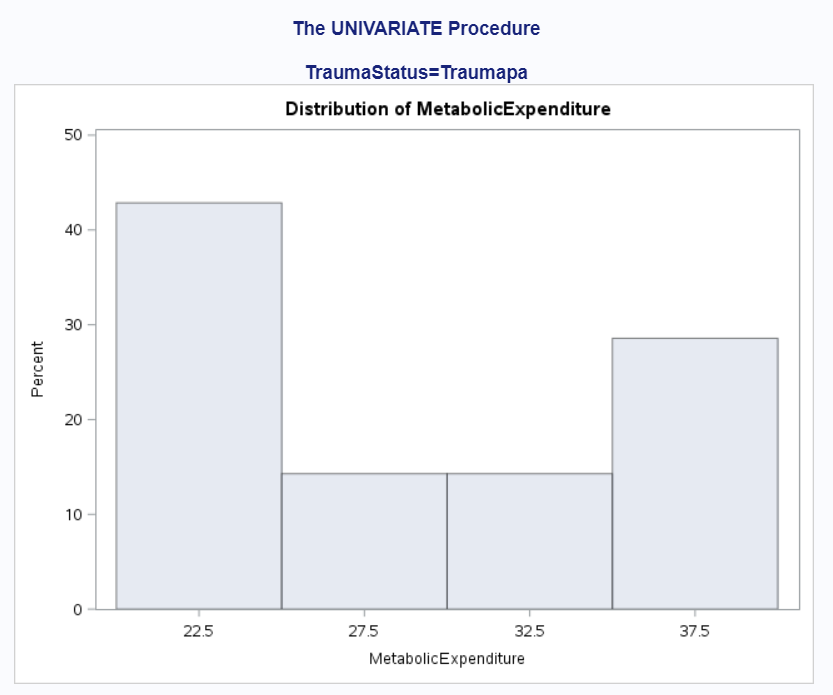
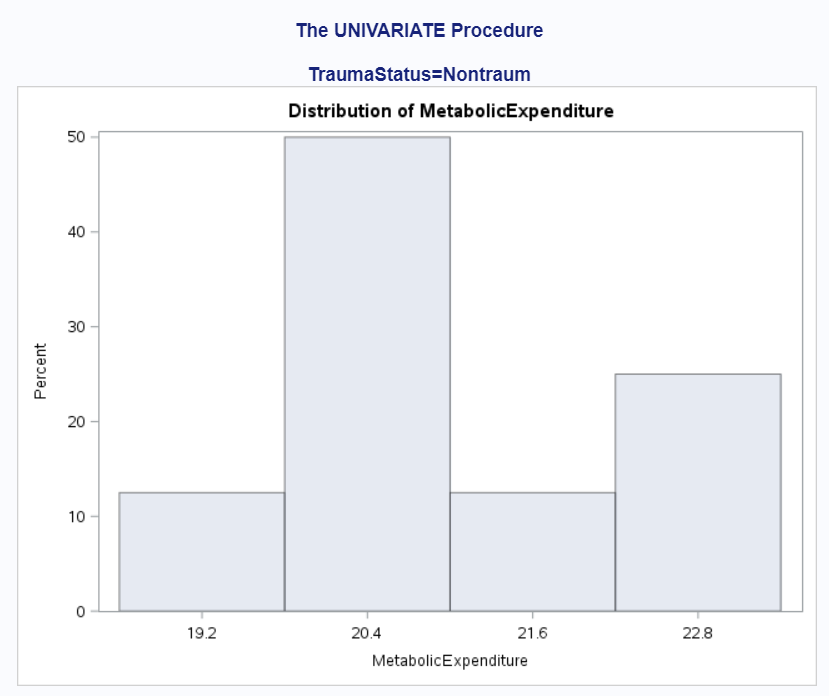
histogram;

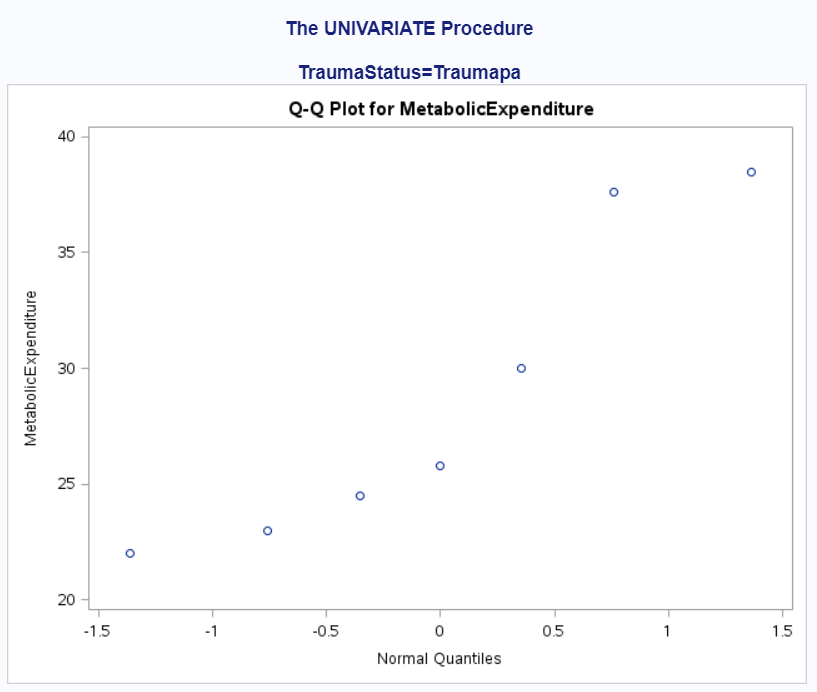
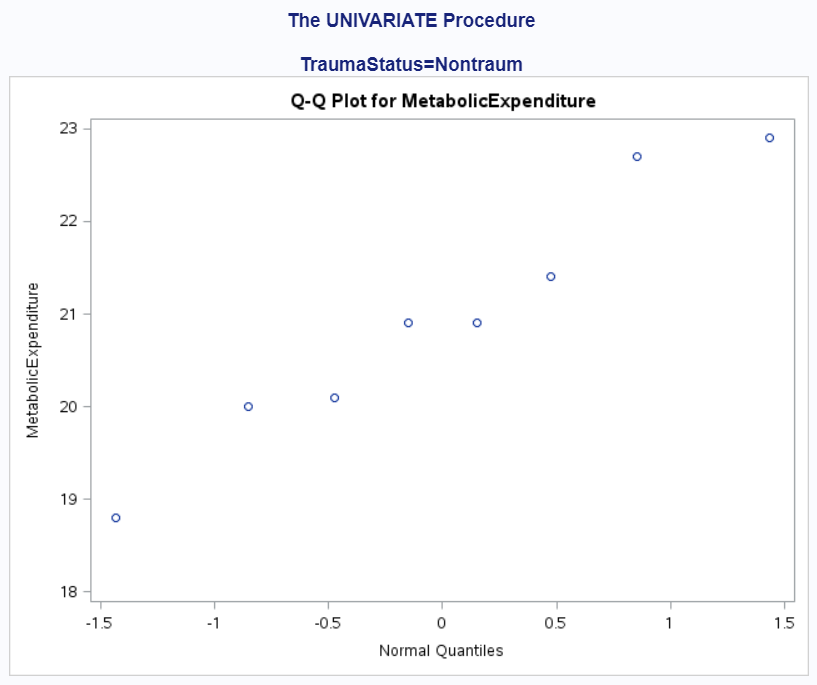
qqplot MetabolicExpenditure;

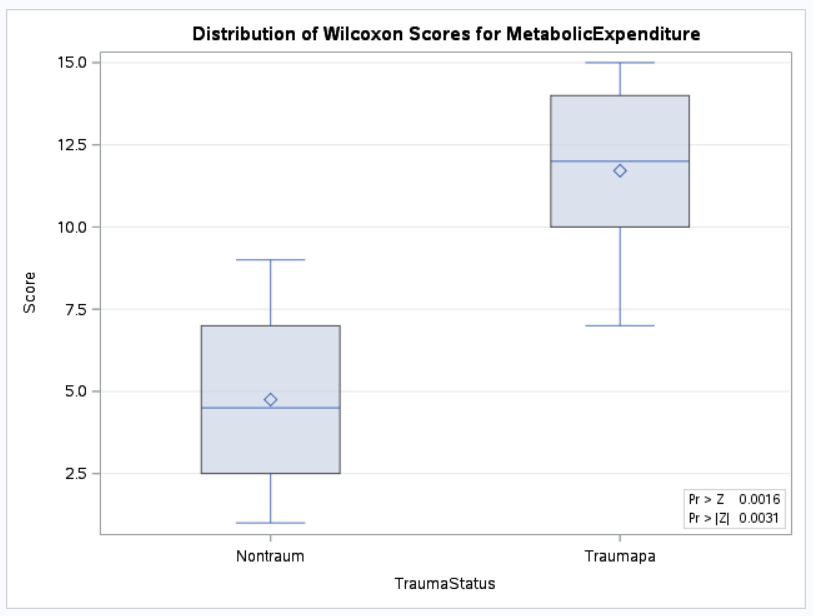
***The following plots indicate that the distributions are not normal. There are no extreme or unrealistic outliers.***

***The rank sum test eliminates the importance of the population distributions altogether and is very resistant to outliers.***

***After replacing the value 20 with 200 the conclusion is still the same (one sided P-value .0161)***







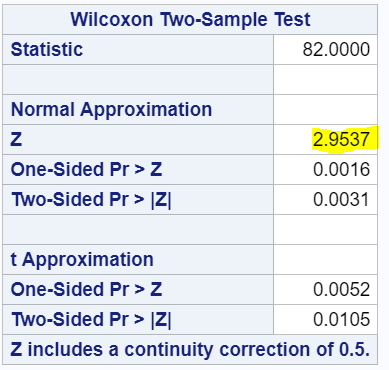
* + 1. Show all 6 steps of the hypothesis test for the rank sum test of the trauma data. Use the critical values, test statistics, p-values, etc. obtained above. Add a **confidence interval from the Hodges-Lehmann procedure (from SAS).**

***Step 1:***

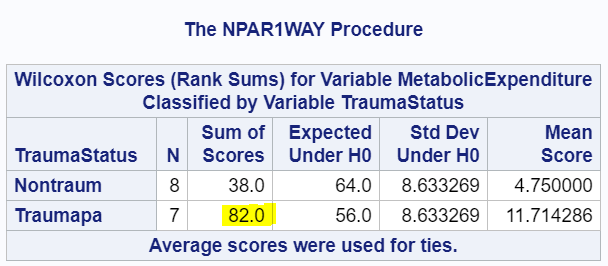
***: The metabolic expenditure in trauma patients is equal to the metabolic expenditure of non-trauma patients.***

***: The metabolic expenditure in trauma patients is greater than the metabolic expenditure of non-trauma patients.***

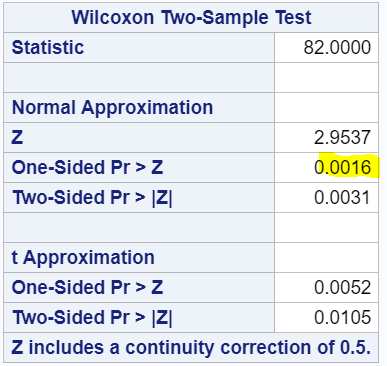
***Step 2: : Normal Approximation Critical Value z = 2.9537***

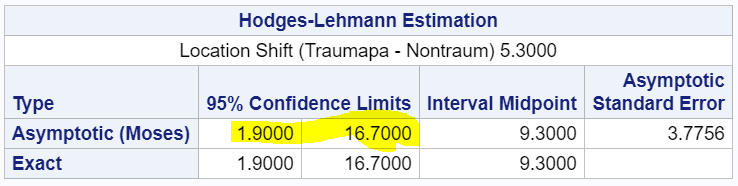


***Step 3:***



***Step 4:***





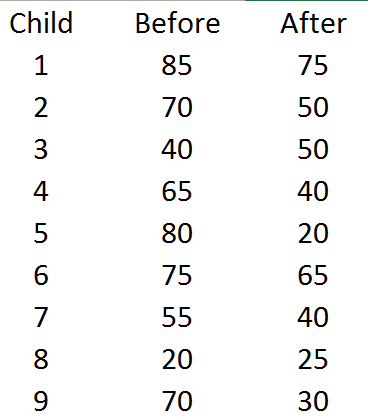
***Step 5: Reject***

***Step 6: The evidence suggests that metabolic expenditure in trauma patients is greater metabolic expenditure in non-trauma patients (one-sided, normal approximation w/ Continuity Correction p-value = 0.0016, 95% confidence interval [1.9, 16.7] from the rank-sum test).***

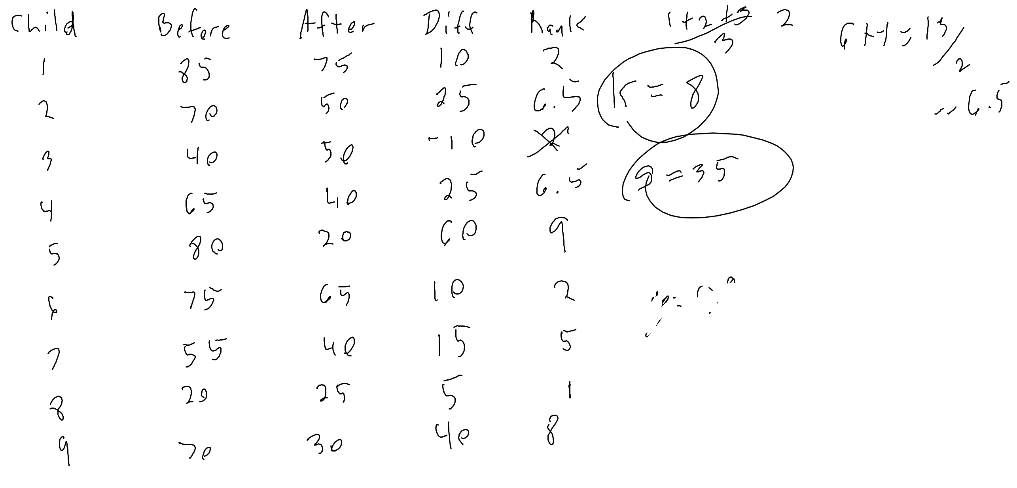
* + 1. Also include a scope of inference statement.

***We cannot make causal inference and must limit our conclusions to differences of the groups in this study.***

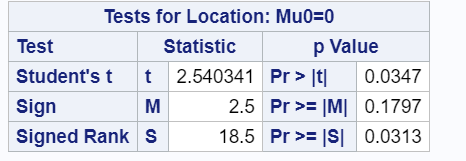
1. A study was performed to test a new treatment for autism in children. In order to test the new method, parents of children with autism were asked to volunteer for the study in which 9 parents volunteered their children for the study. The children were each asked to complete a 20 piece puzzle. The 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:



a. 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).



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



c. 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***

***: Children having yoga exercise complete puzzle in same amount of time as with our having yoga exercise***

***:Children having yoga exercise complete puzzle in different amount of time as with our having yoga exercise.***

***Step 2: : Normal Approximation Critical Value s = 18.5***

***Step 3:***

***Step 4:***

***Step 5: Reject***

***Step 6: The evidence suggests that there is a 95% chance that children having yoga exercise complete puzzle in different amount of time as with our having yoga exercise.***

***p-value = 0.0347 form singed rank test.***

d. 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.

**data** Children;

input Child Before After;

datalines;

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

;

**data** Children2;

set Children;

diff= Before-After;

**run**;

**proc** **print** data=children2;

**proc** **univariate** data=Children2;

var diff;

**run**;

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

## Two-sample test. Borrowed from Internet from Hollander & Wolfe (1973), 69f.  
x <- c(85 ,70 ,40 ,65 ,80 ,75 ,55 ,20 ,70)  
y <- c(75 ,50 ,50 ,40 ,20 ,65 ,40 ,25 ,30)  
wilcox.test(x, y, alternative = "g")

## Warning in wilcox.test.default(x, y, alternative = "g"): **cannot compute  
## exact p-value with ties**  
## Wilcoxon rank sum test with continuity correction  
## W = 61.5, p-value = 0.03448  
## alternative hypothesis: true location shift is greater than 0

wilcox.test(x, y, alternative = "greater",  
 exact = FALSE, correct = FALSE)  
## W = 61.5, p-value = 0.03123  
## alternative hypothesis: true location shift is greater than 0

wilcox.test(rnorm(9), rnorm(9, 2), conf.int = TRUE)

## W = 1, p-value = 8.227e-05  
## alternative hypothesis: true location shift is not equal to 0  
## 95 percent confidence interval:  
## -3.797948 -1.136325  
## sample estimates:  
## difference in location   
## -2.361511

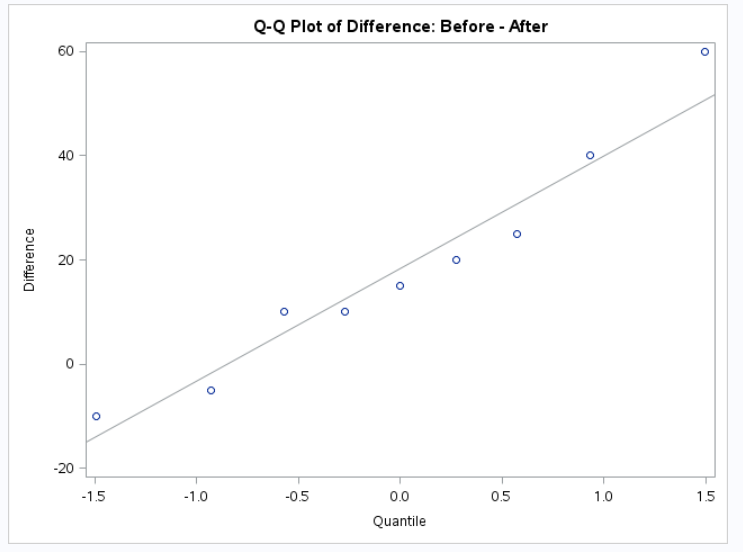
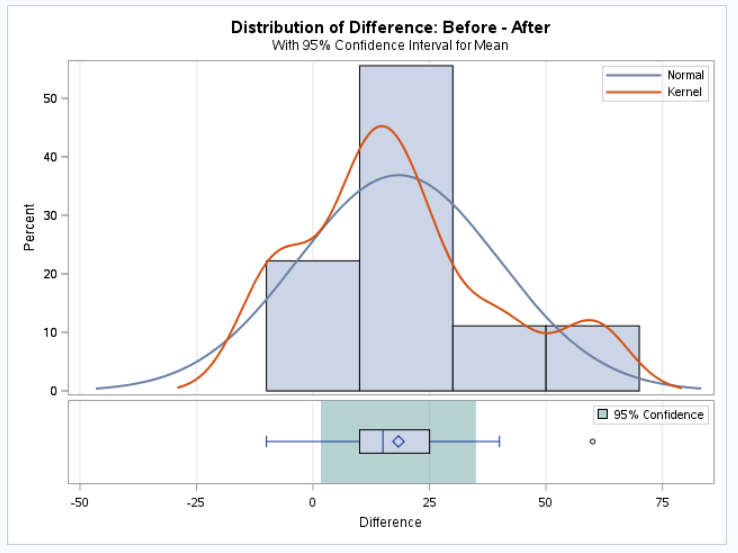
f. Use your data from above to construct a “complete analysis” of the test that you feel is most appropriate to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle. This is simply formatting your results. You should be able to cut and paste most of the work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

***Test the claim that Children having yoga exercise complete puzzles in a different amount of time than children not having yoga exercise***

***Independence:***

***I know nothing about the sampling; I’m critically low on time in completing this assignment therefore this is observational.***

***Looks like a roughly normal distribution but big SD. I would like to transform the data but don’t have time.***

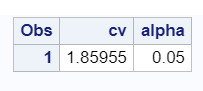


***Step 1***

***: Children having yoga exercise complete puzzle in same amount of time as with our having yoga exercise***

***:Children having yoga exercise complete puzzle in different amount of time as with our having yoga exercise.***

***Step 2: Critical Value =1.85 alpha .05 df=8***



***Step 3:***

**data** Children;

input Before After @@;

datalines;

85 75 70 50 40 50 65 40

80 20 75 65 55 40 20 25

70 30;

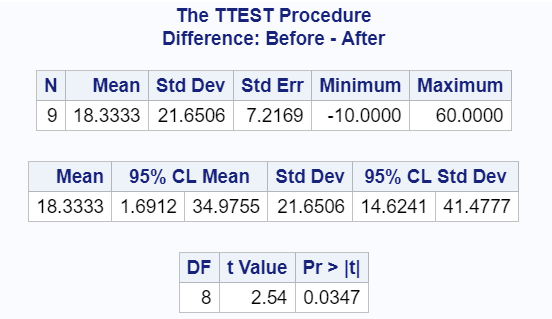
ods graphics on;

**proc** **ttest**;

paired Before\*After;

**run**;

ods graphics off;



***Step 4:***

***Step 5: Reject***

***Step 6: The evidence suggests that there is a 95% chance that children having yoga exercise complete puzzle in different amount of time as with our having yoga exercise. Perhaps if I take Yoga I will be able to get my homework done in a reasonable amount of time.***

***p-value = 0.0347 from paired t-test. Inference should be drawn only about the data in this study.***