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;

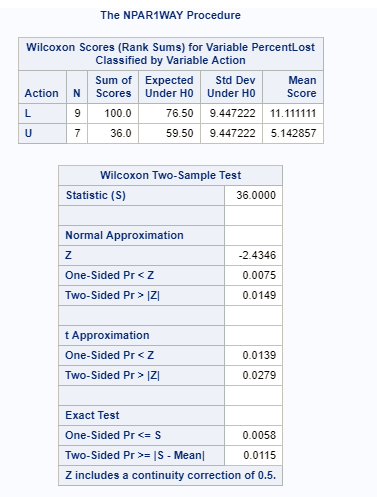
proc npar1way data = loggingdata wilcoxon;

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).

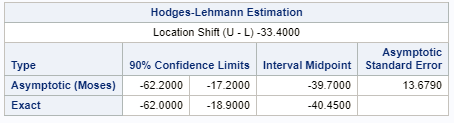
proc npar1way data = loggingdata wilcoxon alpha = 0.1;

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;



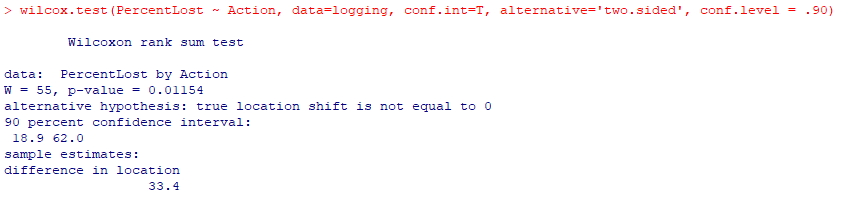
### Part B (5 points total)

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).

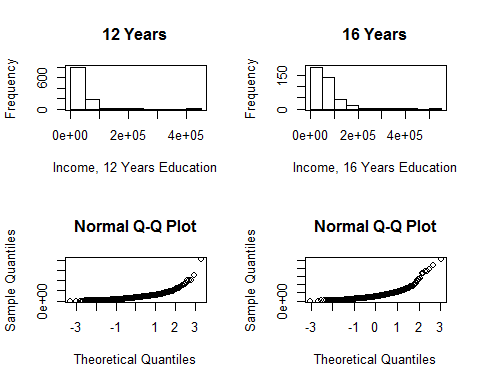
### Part A (5 points)

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:

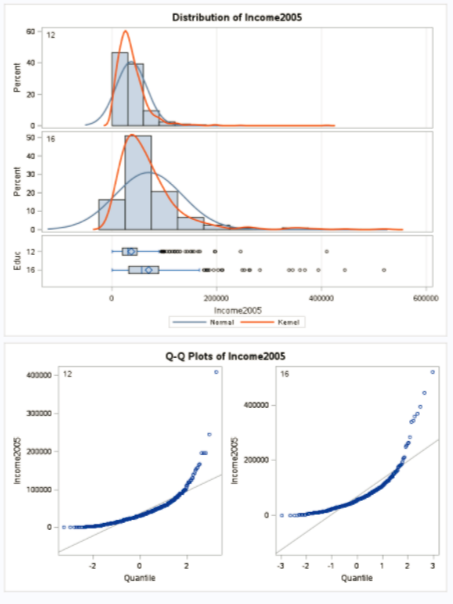
\*To check t-test assumptions on original data;

proc ttest data = education sides = l;

class educ;

var income2005;

run;



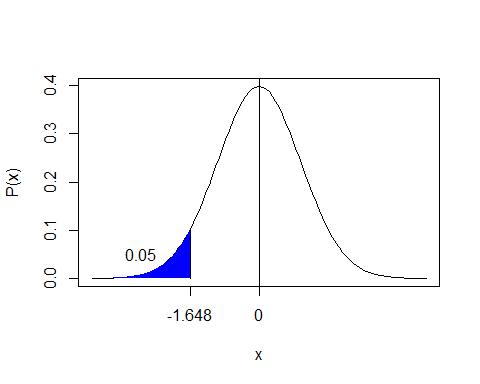
### Part B (5 points)

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:**

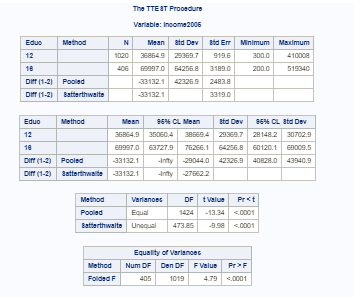
\*To perform t-test;

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.**

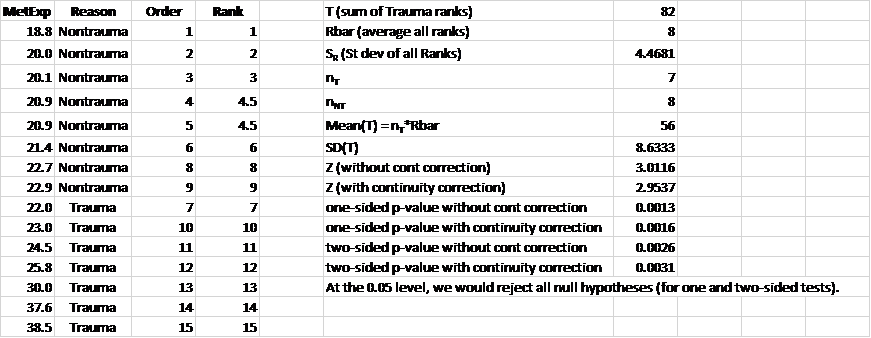
*Note: any reasonable argument can receive full credit.*

## 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;

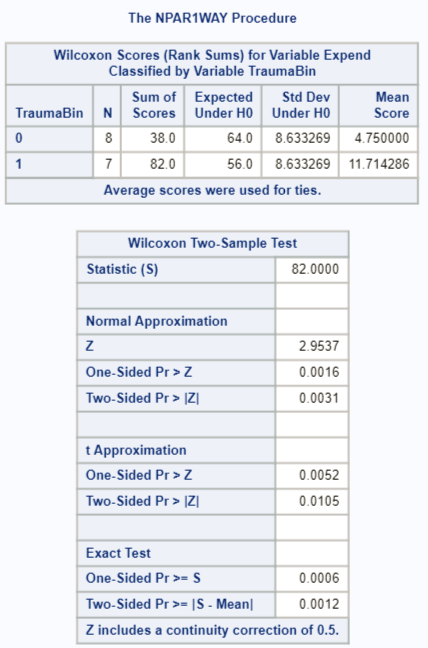
proc npar1way data=Metabolic WILCOXON;

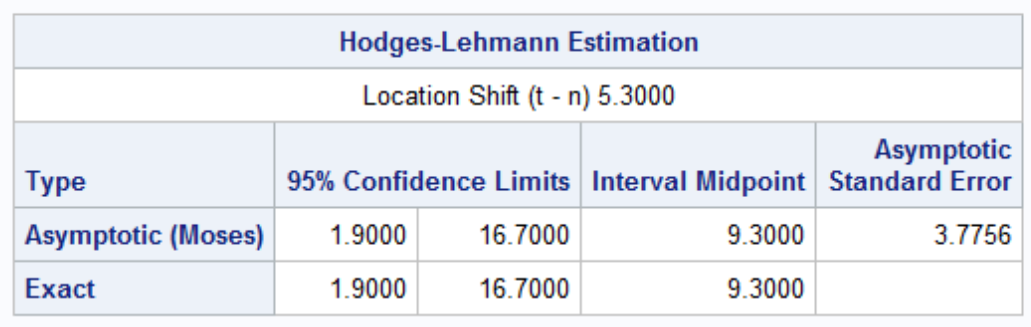
class Traumabin;

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;

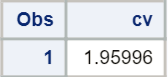
data mycritval;

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;

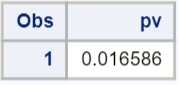
data pvalue;

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;

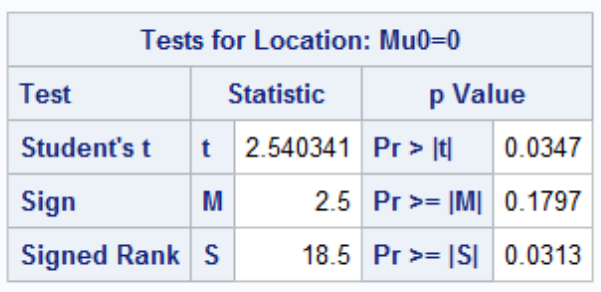
run;

\*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 )**

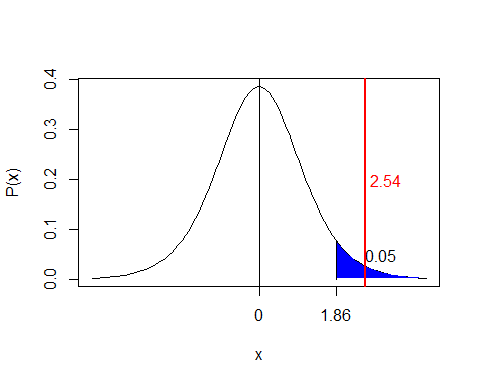
**Step 6 - Conclusion (1 point): There is evidence that the median difference in time to finish the puzzle was greater than 0 ( from a one-sided signed rank test). There is evidence that yoga is associated with shorter puzzle solution times for the individuals who participated. Because the participants 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**

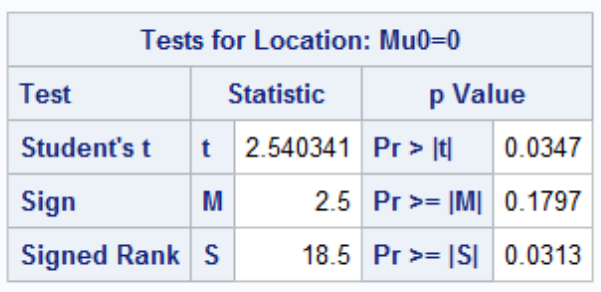
**Step 6 - Conclusion (1 point): There is evidence at the level of significance () that mean difference in time to finish the puzzle was greater than 0.**

**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.*

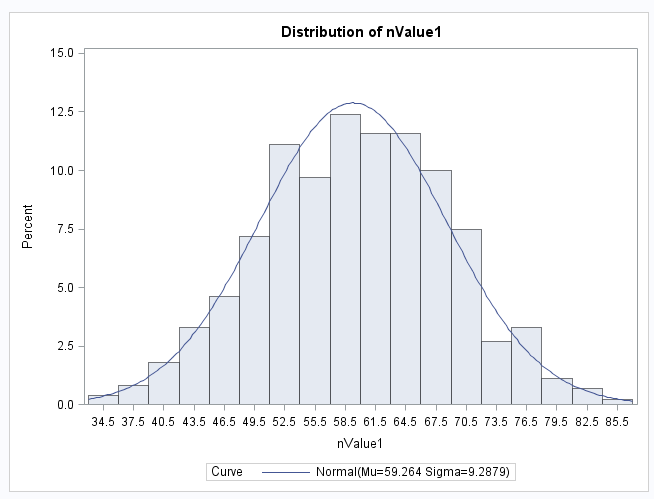
## Bonus (+5 points total)

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.**