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).

```
proc import datafile = '/home/chec0/New Folder/Logging.csv'
 out = logging
 dbms = CSV
 ;

proc print data = logging;
run;

proc npar1way data=logging wilcoxon;
var PercentLost;
class Action;
exact HL;
run;
```

CV
-1.64592439

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable PercentLost Classified by Variable Action

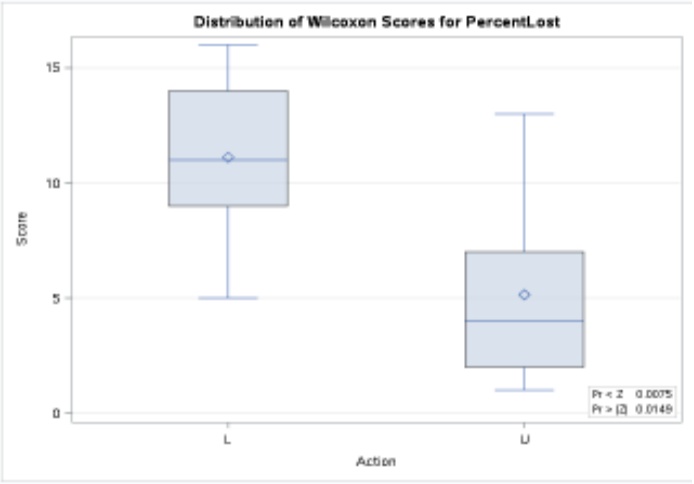
Action	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
L	9	100.0	76.50	9.447222	11.111111
U	7	36.0	59.50	9.447222	5.142857

Wilcoxon Two-Sample Test

Statistic	36.0000
Normal Approximation	
Z	-2.4346
One-Sided Pr < Z	0.0075
Two-Sided Pr > |Z|	0.0149
t Approximation	
One-Sided Pr < Z	0.0139
Two-Sided Pr > |Z|	0.0279
Z includes a continuity correction of 0.5.	

Kruskal-Wallis Test

Chi-Square	6.1877
DF	1
Pr > Chi-Square	0.0129



Hodges-Lehmann Estimation

Location Shift (U - L): -33.4000

Type	95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	-66.8000	-9.0000	-37.9000	14.7452
Exact	-65.1000	-10.8000	-37.9500	

Step 1:

H_0: The distribution of the percent of seedlings in the logged plots is equal to that of the unlogged plots

H_a: The distribution of the percent of seedlings in the logged plots is less than that of the unlogged plots.

Step 2:

$\overline{R} = 8.5 \;\; s_R = 30.2812$

Mean(T) for L group = 9 *8.5 = 76.5 & Mean(T) for U group = 7*8.5= 59.5

SD(T)= Z= 9.4472

Step3: **Z-test = -2.4346**

Step4: **P-value = .0075(1-sided) & .0149(2-sided)**

Step5: **Reject H_0**

Step6: **There is sufficient evidence to suggest that the distribution of the tree seedlings in the logged plots is less than that of the unlogged plots. (the p-value=.0075 from a one sided rank sum test). A 95% confidence interval based off rank**

sum test for the increase in median percent seedlings in favor of the unlogged plots is (10.8%, 65.1%)

Scope of inference: 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 only the 16 plots.

b. 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.

```
> L = c(45, 53.1, 40.8, 75.5, 46.7, 85.4, 85.6, 18.2)
> U = c(23.6, 13.3, 34.2, 18.1, 56.1, -8.1, -20.1)
> wilcox.test(L,U, correct=FALSE)

        Wilcoxon rank sum test

data:  L and U
W = 49, p-value = 0.01399
alternative hypothesis: true location shift is not equal to 0
```

3. 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).
   a. **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: Test the claim that the distribution of incomes for those with 16 years of education($\mu_{1)}$. exceeds the distribution for those with 12 years of education($\mu_{2)}$.

   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 histogram and box plot below indicate strong evidence of inequality of variance between the two populations. With a large sample size the assumption will be robust and will assume independence and run Welch's t test.

The TTEST Procedure

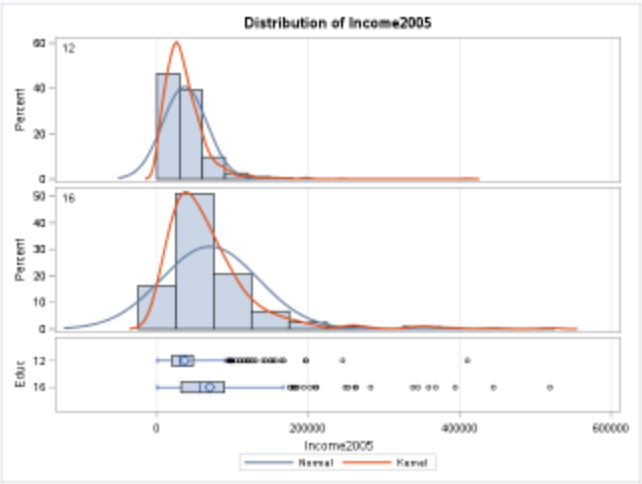
Variable: Income2005

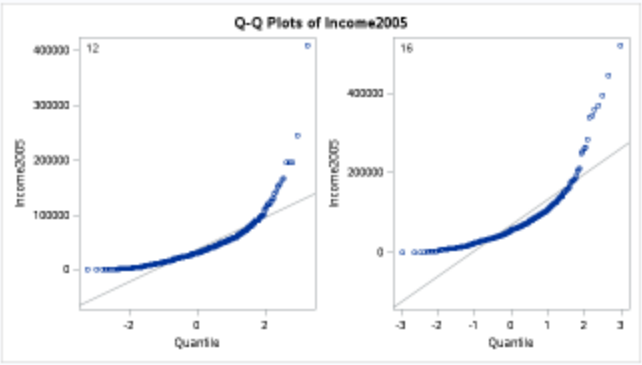
Educ	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
12		1020	36864.9	29369.7	919.6	300.0	410000
16		406	69997.0	64256.8	3189.0	200.0	519340
Diff (1-2)	Pooled		-33132.1	42326.9	2483.8		
Diff (1-2)	Satterthwaite		-33132.1		3319.8		

Educ	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12		36864.9	35060.4	38669.4	29369.7	28148.2	30702.9
16		69997.0	63727.9	76266.1	64256.8	60120.1	69009.5
Diff (1-2)	Pooled	-33132.1	-38004.3	-28259.8	42326.9	40628.8	43940.9
Diff (1-2)	Satterthwaite	-33132.1	-39653.8	-26610.4			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1424	-13.34	<.0001
Satterthwaite	Unequal	473.65	-9.98	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	405	1019	4.79	<.0001





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

Step1: H_0: $\mu_1 = \mu_2$ H_a: $\mu_1 > \mu_2$

Step2: CV=-1.645

Step3: Value of Test Statistic (Satterwaite): $t = -9.98$

Step4: p-value : $p < 0.0001$

Step5: Reject H_o

Step6: There is overwhelming evidence at the alpha = 0.05 level of significance from sample 1 sided Welch's t test($p < 0.0001$) that the mean income in 2005 for people with 12 years of education is less than mean income for those that had 16 years of

education. A 95% confidence interval for the difference is [$26610, $39653] in favor of those with college degrees.

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

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.

d. **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.

```
data <- read.csv(file="C:/Users/che/Desktop/R/EducationData.csv", header=TRUE, sep=",")
t.test(Income2005 ~ Educ, data=data,
    var.equal=FALSE,
    conf.level=0.95)

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
95 percent confidence interval:
 -39653.77 -26610.39
sample estimates:
mean in group 12 mean in group 16
      36864.90         69996.97
```

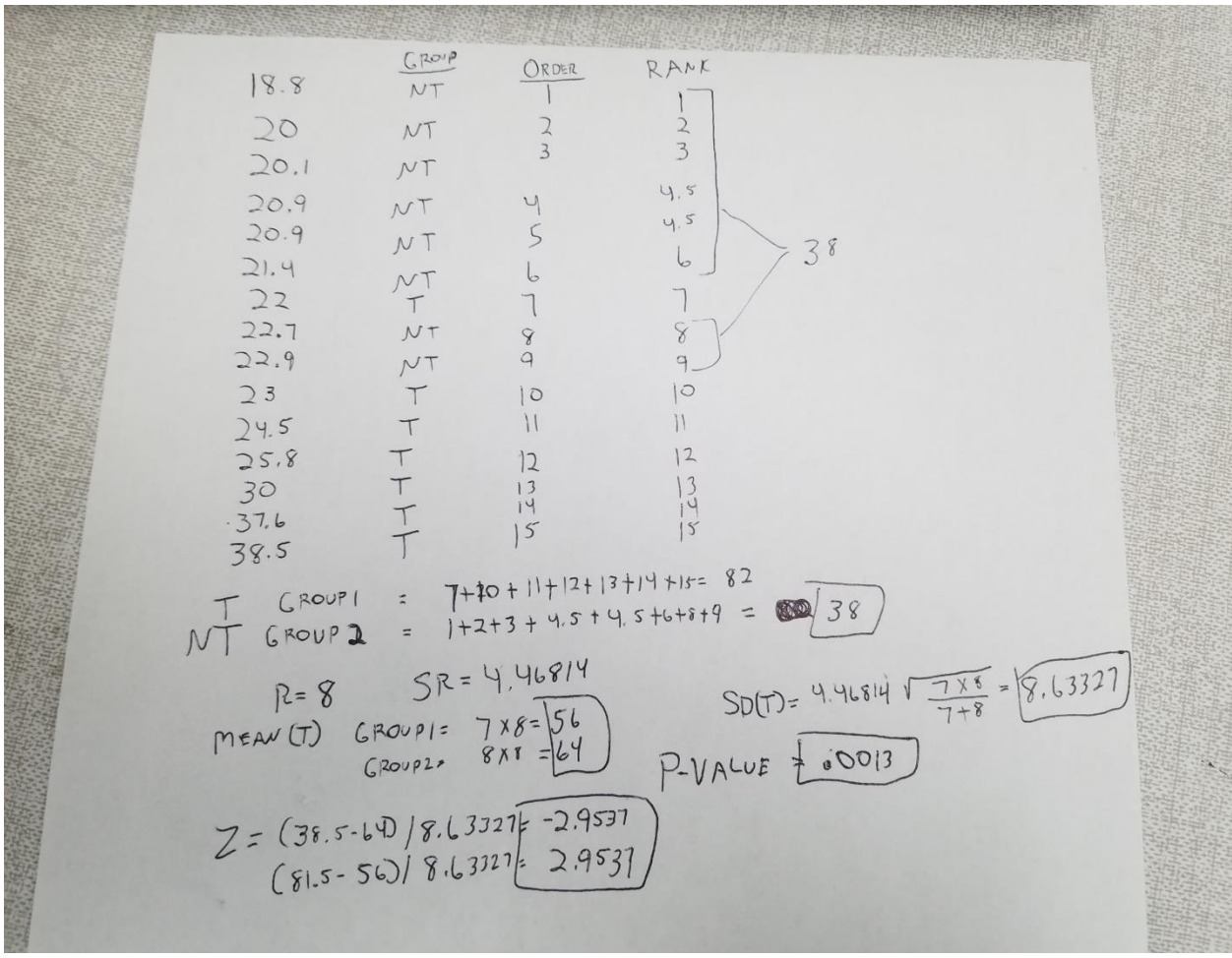
e. **Would you prefer to run the log transformed analysis you ran in HW3, or do you feel this analysis is more appropriate? Why or Why not? (Make mention of the assumptions as well as the parameters that each test provides inference on. As you know, they are different.)**

The original distributions of income were right-skewed, the median may be a more valuable measure of center. So for this reason a log transformation with this test analysis may be preferred.

4.

a. 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.)



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

Sum of Ranks SD(T) and Z statistic are the same

P-value is .0016 in SAS and .0013 by hand

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable number Classified by Variable group

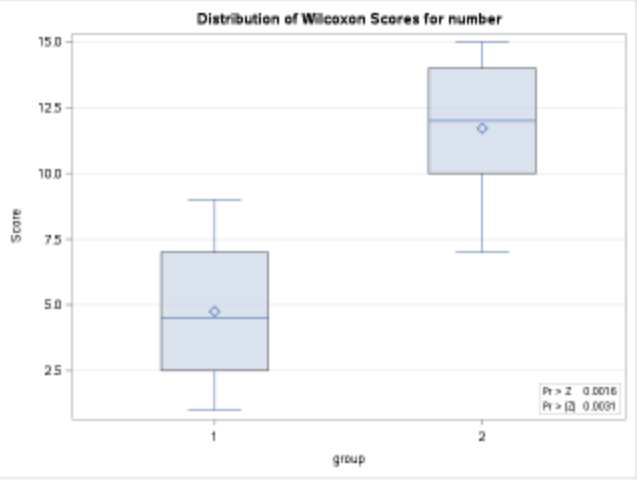
group	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
1	8	36.0	64.0	8.633268	4.750000
2	7	82.0	56.0	8.633268	11.714286

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic	82.0000
Normal Approximation	
Z	2.9637
One-Sided Pr > Z	0.0016
Two-Sided Pr > |Z|	0.0031
t Approximation	
One-Sided Pr > Z	0.0052
Two-Sided Pr > |Z|	0.0105

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test	
Chi-Square	8.8090
DF	1
Pr > Chi-Square	0.0030



Hodges-Lehmann Estimation

Location Shift (2 - 1) 5.3000

Type	95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	1.9000	10.7000	6.3000	2.7756
Exact	1.9000	10.7000	6.3000	

c. 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.

Test the claim that the mean of non-trauma patients exceeds the mean trauma patients

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).

Step1:

H_0: The distribution of the "trauma(group1)" metabolic expenditures is the same as the distribution of the "non-trauma(group2)" metabolic expenditures.

H_a: The distribution of the "trauma(group1)" metabolic expenditures is different from the distribution of the "non-trauma(group2)" metabolic expenditures

Step2:

$\bar{R} = 8\ s_R = 4.46814$

Mean(T) for group1 = 7*8 = 56 & Mean(T) for group2 = 8*8.5= 64

SD(T)=4.46814 √(7*8)/(7+8)=8.63327

Step3:

Z test for group2= (38.5-64)/8.63327=-2.9537

for group1= (81.5-56)/8.63327= 2.9537

Step4: **P-value = .0016(1-sided) & .0031(2-sided)**

Step5: **Reject the H_0**

Step6: **There is evidence at the alpha = 0.05 level of significance from Welch's t test($p = 0.0016$) that the mean metabolic expenditures of group2(Non-Trauma) is less than mean metabolic expenditures of group2(Trauma). A 95% confidence interval for the difference is [1.9000, 16.7000] in favor of those with Trauma**

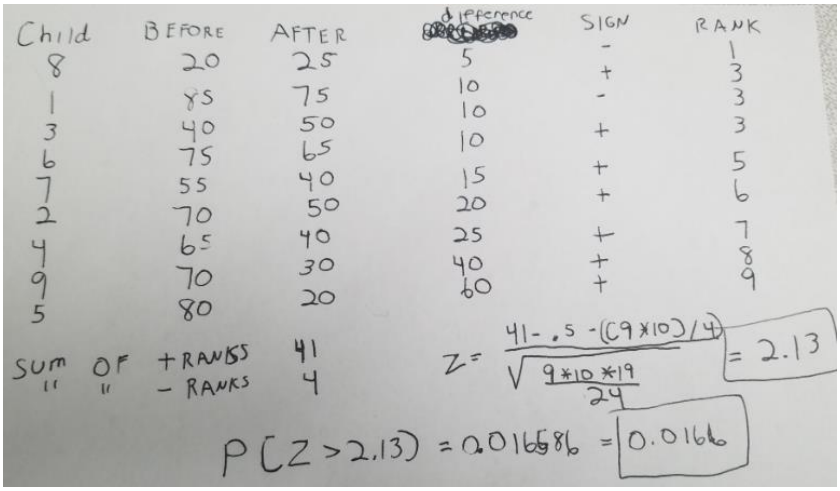
iv. Also include a scope of inference statement.

Since the metabolic expenditures were not randomized to receive either the trauma or non-trauma , ono causation can be implied here. Since the metabolic expenditures were randomly selected, this inference can be generalized to only the 15 patients.

5. 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

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.

SAS CODE and Output

```
data autism;

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 autism2; set Child;

diff = Before - After;

run;
```

proc univariate data=autism2;

var diff;

run;

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.540341	Pr > |t|	0.0347
Sign	M	2.5	Pr >= |M|	0.1797
Signed Rank	S	18.5	Pr >= |S|	0.0313

R Code and Output

before <- c(85, 70, 40, 65, 80, 75, 55, 20, 70)

after <- c(75, 50, 50, 40, 20, 65, 40, 25, 30)

##Sign Test

binom.test(6, 9, alternative='greater')

##Signed Rank Test

wilcox.test(before, after, paired=T, alternative='greater')

```
        Wilcoxon signed rank test with continuity correction

data:  before and after
V = 41, p-value = 0.01618
alternative hypothesis: true location shift is greater than 0
```

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.

Step1:

H_0: The median difference in time to finish a puzzle before and after yoga is zero

H_A: The median difference in the time to finish a puzzle after yoga is greater than zero (if one-sided).

Step2: **Critical Value 1.96 one sided**

Step3: **zstat = 2.13**

Step4: **P-value =0.0166**

Step5: **Reject the H_0**

Step6: **There is sufficient evidence to suggest that the median time to finish a puzzle for individuals before treatment is greater than the median time to finish a puzzle for individuals after they do yoga treatment.(p-value 0.0166 from one sided Signed Rank Test). There is evidence that yoga is associated with shorter puzzle solution times.**

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 autism;

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

;

proc ttest data= autism;

        paired Before*After;

run;
```

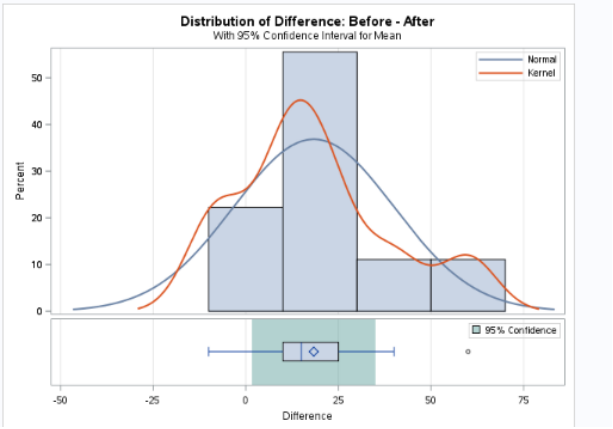
The TTEST Procedure

Difference: Before - After

N	Mean	Std Dev	Std Err	Minimum	Maximum
9	18.3333	21.6506	7.2169	-10.0000	60.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
18.3333	1.6912	34.9755	21.6506	14.6241	41.4777

DF	t Value	Pr > |t|
8	2.54	0.0347



Step1:

H_0: The median difference in time to finish a puzzle before and after yoga is zero

H_A: The median difference in the time to finish a puzzle before and after yoga is greater than zero (if one-sided).

Step2: **Critical value 1.96 for one sided**

Step3: **zstat =2.13**

Step4: take half of 0.0347 and **p-value is .01735 <.05**

Step5: **Reject the H_0**

Step6: **There is sufficient evidence to suggest that the median time to finish a puzzle for individuals before treatment is greater than the median time to finish a puzzle for individuals after they do yoga.(p-value 0.01735 from one sided Paired T Test). There is evidence that yoga is associated with shorter puzzle solution times.**

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

Input = ("

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

")


Data = read.table(textConnection(Input),header=TRUE)


t.test(Data$Before,

    Data$After,

    paired=TRUE,

    conf.level=0.95)
```

```
        Paired t-test

data:  Data$Before and Data$After
t = 2.5403, df = 8, p-value = 0.03469
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.691182 34.975485
sample estimates:
mean of the differences
               18.33333
```

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.

Problem: Test the claim that the yoga treatment was effective in reducing the time for autism patients in finishing a puzzle.

Assumptions: Data are normally distributed, independent between observations with a small sample size. I will use the Signed Rank Test.

Step1:

H_0: The median difference in time to finish a puzzle before and after yoga is zero

H_A: The median difference in the time to finish a puzzle after yoga is greater than zero (if one-sided).

Step2: **Critical Value 1.96 one sided**

Step3: **zstat = 2.13**

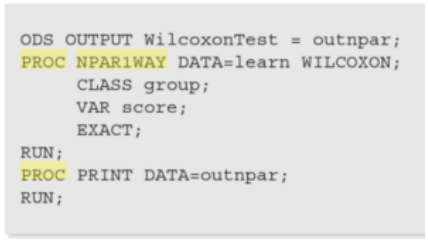
Step4: **P-value =0.0166**

Step5: **Reject the H_0**

Step6: **There is sufficient evidence to suggest that the median time to finish a puzzle for individuals before treatment is greater than the median time to finish a puzzle for individuals after they do yoga treatment.(p-value 0.0166 from one sided Signed Rank Test). There is evidence that yoga is associated with shorter puzzle solution times.**

BONUS (1 pt on 20 pt scale, 5pts on 100 point scale, etc.) This one is challenging and involves hard core SAS coding! Using our permutation test SAS code that we have used in prior HWs, do the following:

a. 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 in earlier in the homework.
b. 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!).	

```
ODS OUTPUT WilcoxonTest = outnpar;
PROC NPAR1WAY DATA=learn WILCOXON;
     CLASS group;
     VAR score;
     EXACT;
RUN;
PROC PRINT DATA=outnpar;
RUN;
```