Kyle Thomas HW 4

1.

Done

2A.

Problem

Does logging an area after a forest fire increase the number of lost seedlings, thus hampering forest recovery?

Check Assumptions

The sample sizes are quite small for both the logged and unlogged areas. Since there are so few data points, it is hard to gain any insight from a histogram of the data or a QQ-plot of the data. As a result, it a rank-sum test will be conducted on the data.

Hypothesis

- H0: The distribution of seedling loss in logged plots are the same as the seedling loss in unlogged plots.
- HA: The distribution of seedling loss in logged plots are higher than the seedling loss in unlogged plots.

Analysis

A rank sum test on the data was conducted to determine if there was a difference in seedling loss between the logged and unlogged plots using SAS. The procedure NPAR1WAY was used for this analysis. The results of the test are shown **Figure 1**. The sum of the ranks of the logged data was equal to 100, and the sum of the unlogged data was equal to 36. This provides evidence that logged plots of land experienced more seedling loss than unlogged plots of land.

Figure 1

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 (S)	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.	

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	

Conclusion

The Wilcoxon Rank Sum Test produced a one-sided p-value of 0.0075 providing strong evidence against the null hypothesis. As a result of this analysis I reject the null hypothesis. There is strong evidence that logging forests after a forest fire causes increased loss of tree seedlings which hiders' forest recovery.

Scope of Inference

Since this is a randomized experiment in which tracks of land that were impacted by a forest fire were randomly assigned a treatment (logging vs non-logging) we can draw casual inference. There is sufficient evidence that logging a plot of forest after a forest fire increase the loss of tree seedling, which hinders forest recovery (Wilcoxon Rank Sum Test one-sided p-value of 0.0075).

Code Box 1

The code used to come to this conclusion is shown in below.

```
/*HW4 Problem 2A*/
```

```
/* Get the data into SAS */
FILENAME REFFILE '/folders/myshortcuts/SMU/MSDS6371 - Stats/Unit
4/Logging.csv';
PROC IMPORT DATAFILE=REFFILE
     DBMS=CSV
     OUT=WORK.IMPORT;
     GETNAMES=YES;
RUN;
/*make the data an easy name*/
data work.logging;
     set work.import;
run;

/*check assumptions of the data*/
proc univariate data=work.logging;
     class action;
     histogram PercentLost;
     qqplot  PercentLost;
run;

/*Analysis of logging data using a rank sum test*/
proc npar1way data=work.logging wilcoxon hl alpha=0.05;
     var PercentLost;
     class Action;
     exact wilcoxon hl /mc;
run;
```

2B.

The same analysis was conducted in R. The output is shown in **Figure 2**, and the code used to generate the output is shown in **Code Block 2.** The two-sided p-value and two-sided confidence intervals match between R and SAS.

Figure 2

```
        Wilcoxon rank sum test with continuity correction

data:  logged and unlogged
W = 55, p-value = 0.01491
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 10.79996 65.10000
sample estimates:
difference in location
              33.39998
```

Code Block 2

```
####STATs HW4###
#Question 2B#
setwd("C:\\Users\\kthomas\\Documents\\SMU\\MSDS6371 - Stats\\Unit 4")
logging = read.csv("Logging.csv")
logged = logging$PercentLost[logging$Action == "L"]
unlogged = logging$PercentLost[logging$Action == "U"]

wilcox.test(logged, unlogged, correct = TRUE, exact = FALSE, conf.int
= TRUE)
```

3A – 3C

Problem

Does completing 16 years of education (E16) have an impact on the income of a person when compared to completing only 12 years of education (E12)? A Welch's t-Test will be used in order to investigate this question.

Assumptions

In order to perform the Welch's t-Test, the data must be independent and take the form of a normal distribution. The data does not suggest that incomes are normally distributed. Rather, income for both groups appears to be skewed. The data for E16 has a strong right skew, and the data for E12 has moderate right skewed. The QQ-plots show that the data does not follow a normal distribution. The histograms are shown in **Figure 3** and the QQ-Plots are shown in **Figure 4**.

However, the central limit theorem states that averages based on large samples will be normally distributed even if the underlying population is not normally distributed. However, it is also known that

the t-tools are greatly impacted by skewness when if the sample sizes are not similar. In this case, the data is heavily skewed and the sample sizes are quite different (there are more than twice as many samples in the E12 group than there are in the E16 group).

Figure 3

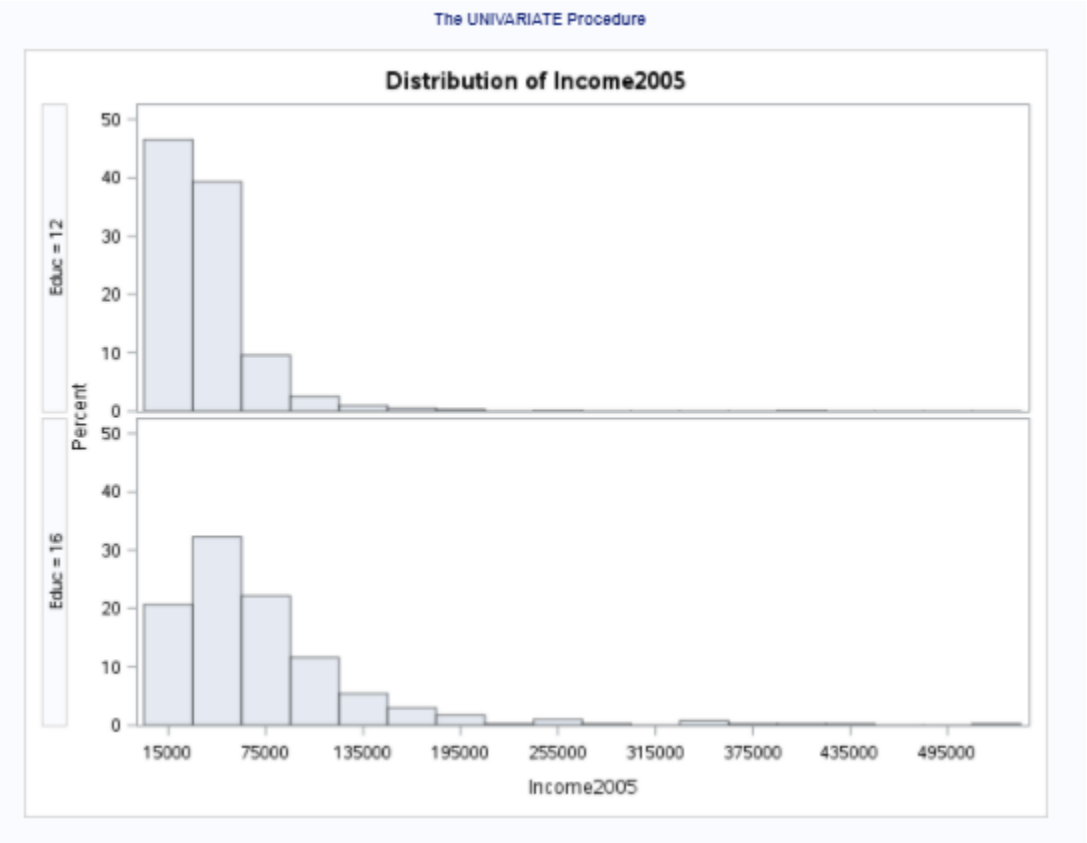


Figure 3: Histograms for the income variable. Income is plotted on the X axis. The top graph shows E12, and the bottom graph shows E16

Figure 4

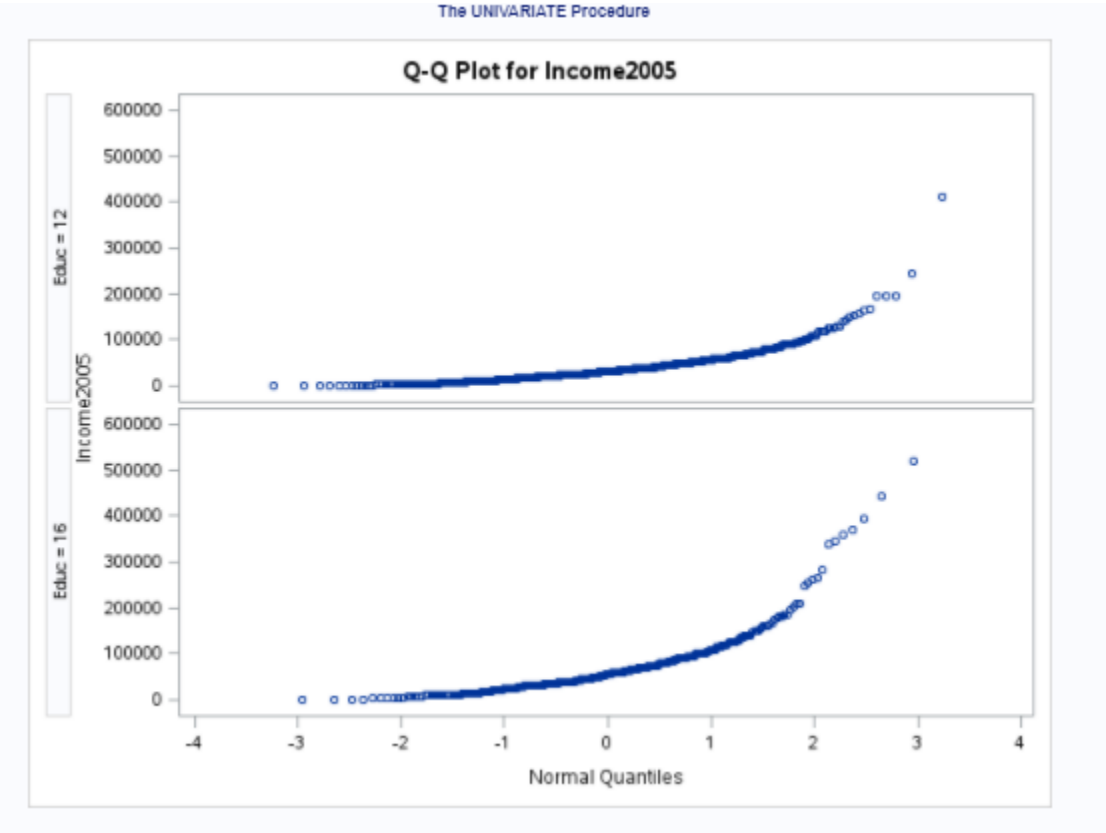


Figure 4: QQ-Plot for income variable. Income is plotted on the X axis. The top graph shows E12, and the bottom graph shows E16

Hypothesis

Educational achievement has an impact on the amount of money earned. I hypothesize that the income level of those with 16 years of education will earn more money, on average, than those who's highest level of education attainment was 12 years of education.

- **H0:** There is no difference in income between those who have complete 16 years of education (E16) and those who have complete 12 years of education (E12): E16 – E12 <= 0:
- **HA:** The mean income of those with 16 years of education (E16) is higher than the mean income of those with 12 years of education (E12): E16 – E12 > 0

Find Critical Value

- Assuming that alpha = 0.05
- Using a two tailed test
- Degrees of freedom = 1426 – 2 = 1424
- Critical value for two tailed alpha of 0.05 with 1424 degrees of freedom: 1.646

Analysis

A two sided, two sample t-test was conducted on the income data by education attainment using PROC TTEST in SAS. The alpha for this test was 0.05. T t-value returned by SAS using the Satterthwaite adjustment was 9.98 with a corresponding p-value less than 0.0001.

Conclusion

Based on the results of this t-test there is sufficient evidence to reject the null hypothesis (p-value less than 0.001). This suggests that the evidence shows that the income of these whom have completed 16 years of education (E16) is higher than the mean income of those whom completed 12 years of education (E12). The true difference in earning potential of those with 16 years of education and 12 years of education is $26,610.4 and $39,653.8 (95% confidence interval)

Score of Inference

Since this is an observation study no casual inferences can be drawn. However, since the samples were picked at random, we can draw inferences to the populations in the study.

The code used to conduct this analysis is below:

```
/* Welch's two sample t test on Education Data from HW3 */
/*get the data*/
filename EduData '/folders/myshortcuts/SMU/MSDS6371 - Stats/Unit
3/Homework/EducationData.csv';

proc import datafile=EduData
     DBMS = CSV
     OUT = work.EduData;
     GETNAMES = Yes;
run;

proc sort data=work.edudata;
     by descending Educ;
run;

proc univariate data=work.edudata;
     class Educ;
     var income2005;
     histogram income2005;
     QQplot income2005;
run;

/* conduct the t-test */
proc ttest data=work.edudata sides=2 order=data;
     class Educ;
     var income2005;
run;
```

3D.

Conduct the same analysis in R. The output of the Welch's t-test is shown in **Figure 5** and the code used to generate this test is shown in the code block below Figure 5. Overall, the p-values, and confidence intervals agree. Their signage is different because in R analysis, E16 was subtracted from E12 making the difference in means negative. In SAS I specified that E12 should be subtracted from E16 to make the output more logical.

Figure 5

```
        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
```

Code Block: Welch's t-Test in R

```
#Question 3D#
EduData = read.csv("EducationData.csv")
t.test(Income2005 ~ Educ, data=EduData, var.equal=FALSE, conf.level =
0.95)
```

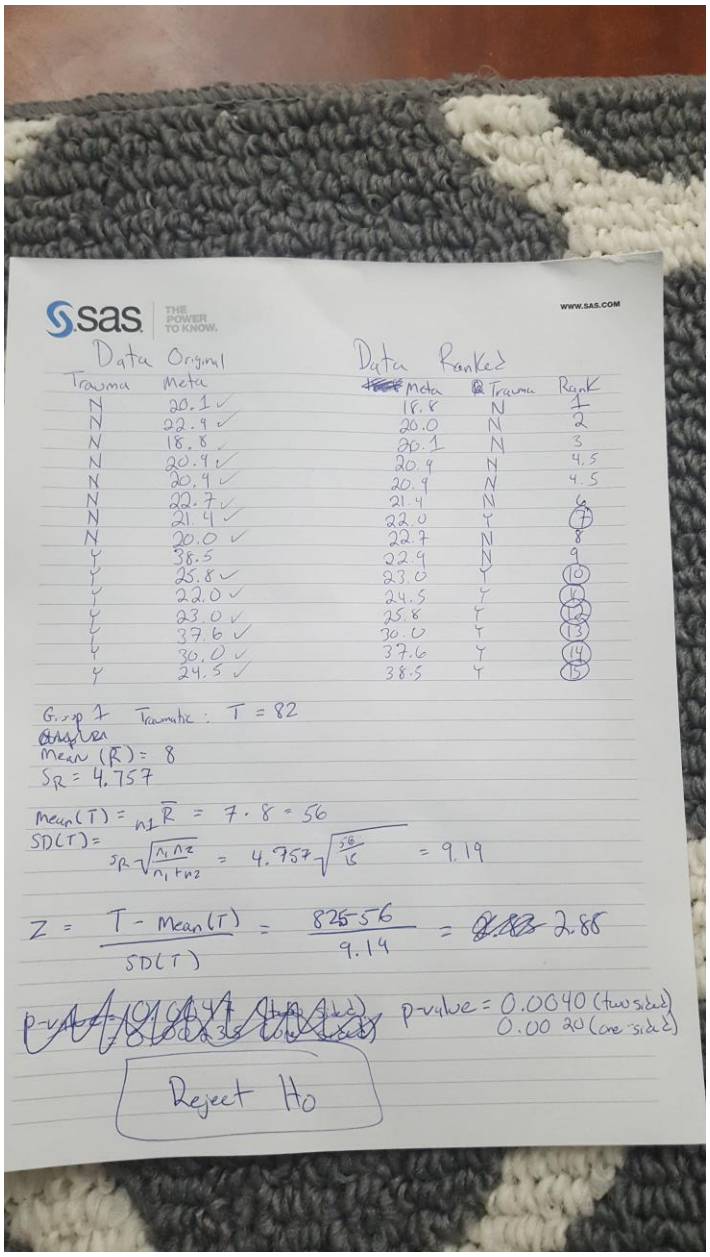
3E.

I would prefer to run the log transformed analysis over this analysis because the log transformed data addresses the skewness of the data. Since the sample sizes are quite different, I know that skewness in the data can have a large impact on the t-test. However, logging the data will provide a different inference as it is looking at a multiplicative impact. Nevertheless, the logical outcome of the multiplicative impact lends evidence to the alternative hypothesis.

4A.

Figure 7 shows my results of the rank sum test by calculated by hand.

Figure 7



4B.

My answer is a little different from SAS. I had a Z score of 2.88 compared to the Z-score created by SAS of 2.95. Additionally, my p-values are different. I had a one sided p-value of 0.004 and SAS had a one sided p-value of 0.0052.

I think these differneces come down to the fact that I had significant drop in my precision (I only carried two decimals) and the fact that I asked SAS to calculate the exact p-value using a Monte Carlo

simulation. While there are small differences in the answers, the results are the same. In either instances the Z-scores and p-values are far past the traditional values needed to reject the null hypothesis. My SAS output is shown in **Figure 8** and the code used to generate the analysis is shown in the code block below.

Figure 8

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable meta Classified by Variable tram

tram	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
N	8	38.0	64.0	8.633269	4.750000
Y	7	82.0	56.0	8.633269	11.714286

Average scores were used for ties.

Wilcoxon Two-Sample Test	
Statistic (S)	82.0000
Normal Approximation	
Z	2.9537
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.	

Hodges-Lehmann Estimation

Location Shift (Y - N) 5.3000

Type	95% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	1.9000	16.7000	9.3000	3.7756
Exact	1.9000	16.7000	9.3000	

```
/*HW 4 Question 4B */
data work.trauma;
     input tram $ meta;
datalines;
N 20.1
N 22.9
N 18.8
N 20.9
N 20.9
N 22.7
N 21.4
```

```
N 20.0
Y 38.5
Y 25.8
Y 22.0
Y 23.0
Y 37.6
Y 30.0
Y 24.5
;
run;

proc npar1way data = work.trauma wilcoxon hl;
      var meta;
      class tram;
      exact wilcoxon hl /mc;
run;

data crit;
      critValue = quantile("T", 0.95, 14);
run;


proc univariate data=work.trauma;
      class tram;
      var meta;
      histogram meta;
      QQPLOT meta;
run;
```

4C.

Problem

Do traumatic injuries cause an increase in caloric intake of patients?

Assumptions

Since there are so few samples taken, we cannot be certain that the data is normal. The histograms show that the data for the trauma patients is pretty flat, and does not appear to be normally distributed. This means that the assumptions of the t-test are not met. As a result, a rank sum test will be conducted on the data in order to gain evidence to answer the question presented in the problem. The histograms of the caloric expenditures is shown in **Figure 9** and QQ-plots of the caloric expenditures of the patients is shown in **Figure 10**.

Figure 9

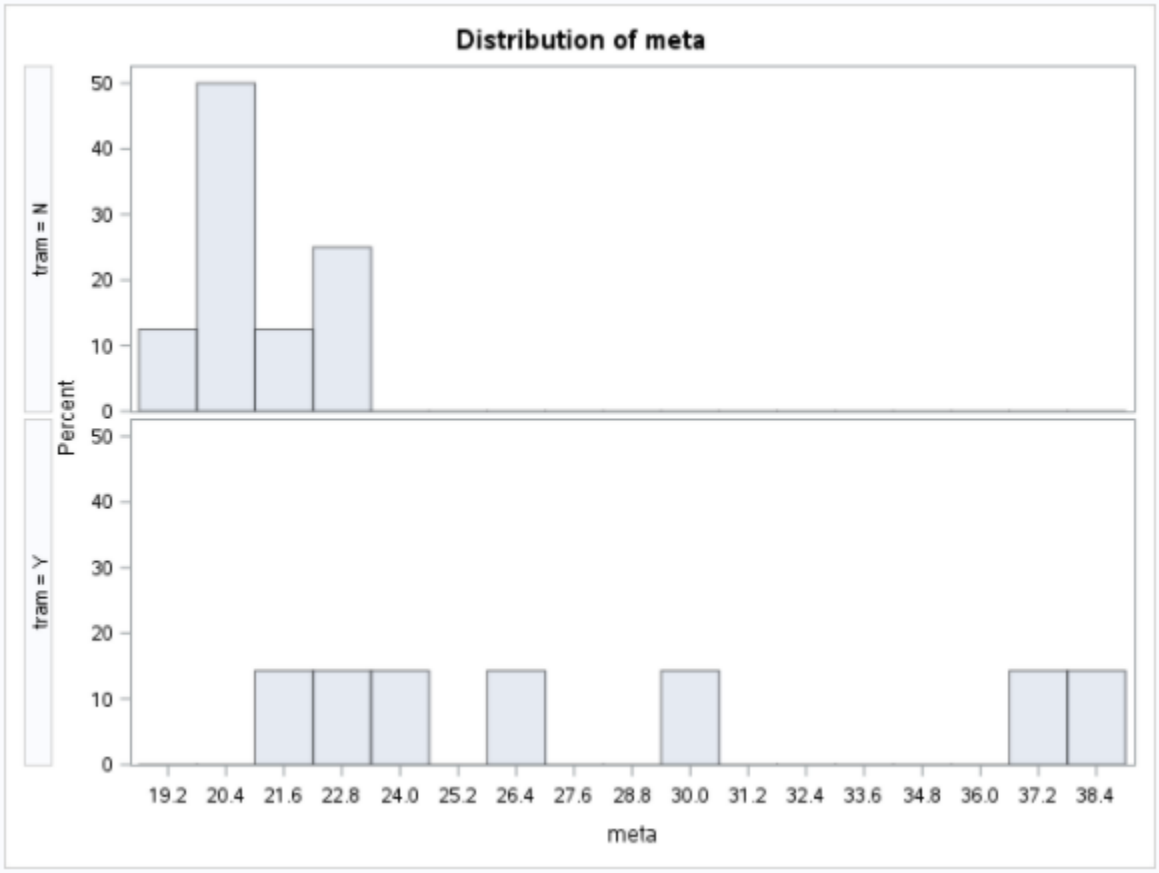
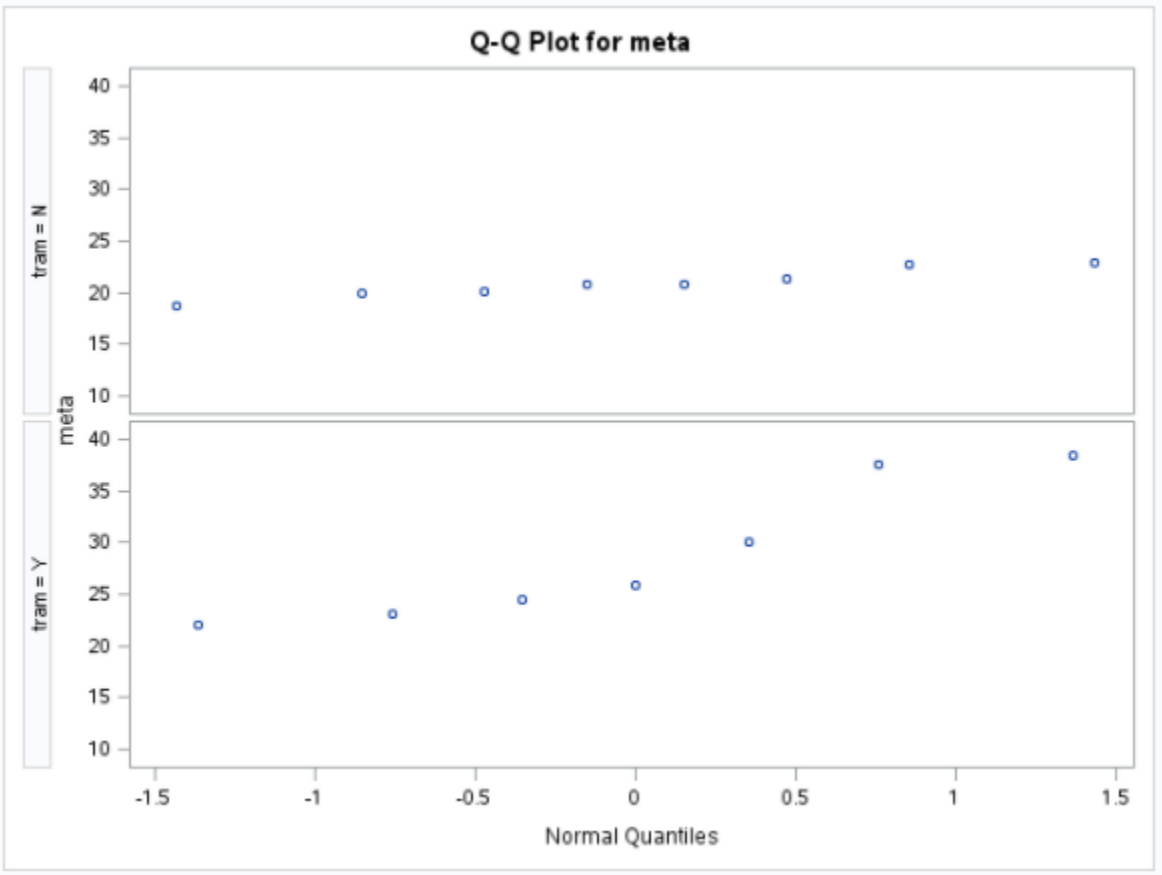

Distribution of meta
tram = N
tram = Y
Percent
0
10
20
30
40
50
19.2
20.4
21.6
22.8
24.0
25.2
26.4
27.6
28.8
30.0
31.2
32.4
33.6
34.8
36.0
37.2
38.4
meta

Figure 10



Hypothesis

Caloric intake for trauma patients is higher than non-trauma patients.

- H0: The distribution of caloric intake trauma patients are the same as the caloric intake of non-trauma patients.
- HA: The distribution of caloric intake trauma patients is not the same as the caloric intake of non-trauma patients.

Find Critical Value

- Assuming that alpha = 0.05
- Using a two tailed test
- Degrees of freedom = 14
- Critical value for two tailed alpha of 0.05 with 14 degrees of freedom: 1.76

Analysis

A rank sum test produced a Z-score of 2.96 and an associated two-sided p-value of 0.0031 (using the Normal Approximation method).

Conclusion

There is strong evidence that the body of a trauma patient consumes different amounts of calories than the body of a non-trauma patient (two-sided p-value of 0.0031 using a rank sum test). The true difference in means between a trauma patient and a non-trauma patient is 1.9 kCal/kg/day to 16.7 kcal/kg/day (95% confidence interval).

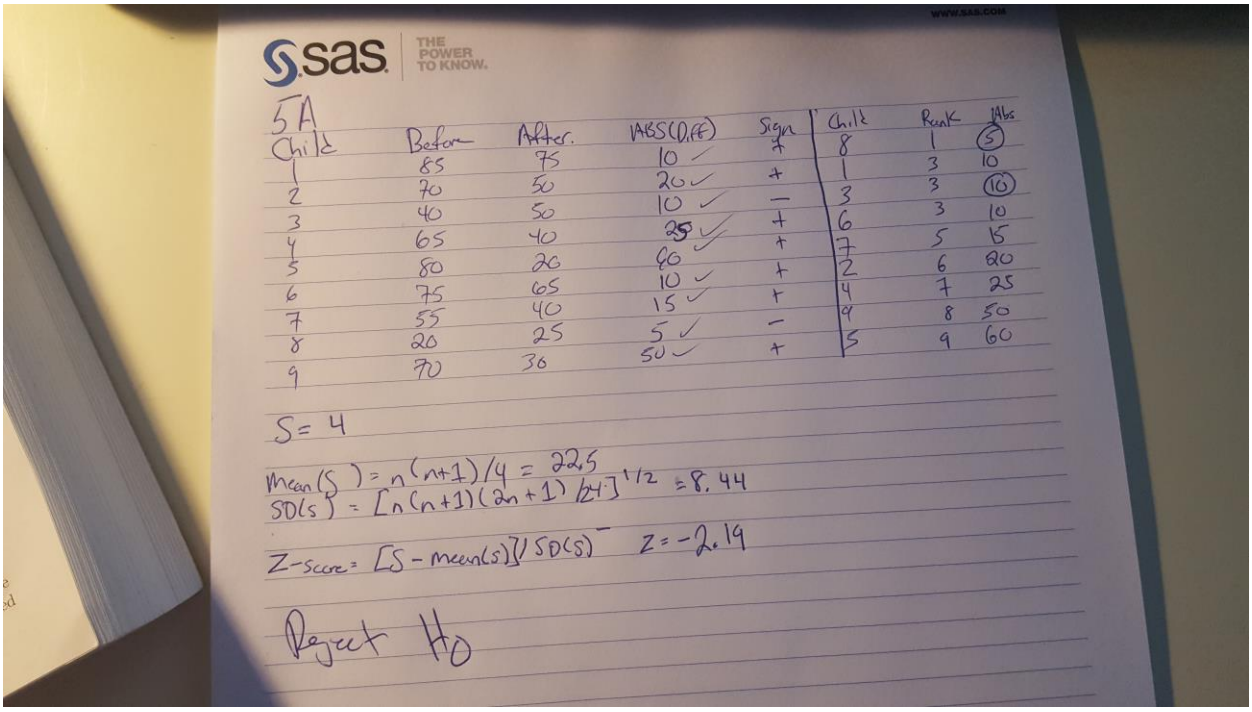
Scope of Inference

Since this was an observational study, no casual inference can be made (i.e. it cannot be stated that the traumatic injuries caused the higher consumption of calories). However, we can be fairly certain that the difference between being a trauma patient and a non-trauma patient is the result of random process (and a rather unfortunate one at that). As a result, we can make inferences to the populations that the patients come from. That is to say, we can expect trauma patients to consume more calories than non-trauma patients.

Kyle Thomas HW 4

5A.

The results of the sign rank test below: Reject the null hypothesis. There is evidence that yoga reduced the time required to solve the puzzle after a treatment of yoga (p-value of 0.0285).



5A

Child	Before	After	ABS(Diff)	Sign
1	85	75	10	+
2	70	50	20	+
3	40	50	10	−
4	65	40	25	+
5	80	20	60	+
6	75	65	10	+
7	55	40	15	+
8	20	25	5	−
9	70	30	50	+

Child	Rank	Abs
8	1	5
1	3	10
3	3	10
6	3	10
7	5	15
2	6	20
4	7	25
9	8	50
5	9	60

S = 4

Mean(S) = n(n+1)/4 = 22.5

SD(S) = [n(n+1)(2n+1)/24]^1/2 = 8.44

Z-score = [S − mean(S)]/SD(S) Z = −2.14

Reject H_0

5B.

Again my results are close to what I would expect and are likely caused by rounding errors that SAS and R would not experience. Also, I calculated a Z-score and SAS calculated a t-score (which would be more conservative). Also, when the population is less than 20 SAS uses the binomial distribution to calculate the p-values for increased accuracy.

My Results: Z = 1.9, p-score (two sided) 0.0285

SAS: Results and Code shown below

The UNIVARIATE Procedure
Variable: diff

Moments			
N	9	Sum Weights	9
Mean	18.3333333	Sum Observations	165
Std Deviation	21.6506351	Variance	468.75
Skewness	0.7310904	Kurtosis	0.5328254
Uncorrected SS	6775	Corrected SS	3750
Coeff Variation	118.094373	Std Error Mean	7.21687836

Basic Statistical Measures			
Location		Variability	
Mean	18.33333	Std Deviation	21.65064
Median	15.00000	Variance	468.75000
Mode	10.00000	Range	70.00000
		Interquartile Range	15.00000

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

```
/*HW 5B */
data work.yoga;
      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
run;

data work.yoga2;
      set work.yoga;
      diff = before - after;
```

```
run;

proc univariate data=work.yoga2;
      var diff;
run;
```

R: Results and code for R analysis shown below:

```
        Wilcoxon signed rank test with continuity correction

data:  yoga$Before and yoga$After
V = 41, p-value = 0.03236
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
  2.499951 37.500051
sample estimates:
(pseudo)median
      17.49993
```

```
#Question 5B#
yoga = read.csv("yoga2.csv")
wilcox.test(yoga$Before, yoga$After, correct = TRUE, exact = FALSE,
conf.int = TRUE, paired = TRUE)
```

5C.

Problem

Does a treatment of yoga increase the ability of children with autism to solve problems?

Testing Assumptions

A visual investigation of the data shows that the before data is slightly left skewed, but a QQ-plot of the data reveals that it is fairly normal. The after treatment times also appear normal. However, there is so little data that a parametric test could result in misleading results. A sign-rank test is being used to test the problem.

Hypothesis

Yoga increases the ability of children with autism to solve problems.

- H0: There is no difference in the distribution of times it takes to solve a puzzle before or after a treatment of yoga.
- HA: The distribution of the times of children treated with yoga are less than the times before a treatment of yoga.

Analysis

SAS software was used to conduct a signed rank test on the difference in times before and after the treatment of yoga using PROC UNIVARIATE. There is sufficient evidence to reject the null hypothesis (p-value of 0.0313).

Conclusion

There is strong evidence to suggest that a treatment of yoga increases the problem solving ability in children with autism.

Score of inference

Since this is an observational study with no randomization processes no casual inference can be made, nor can inferences to the larger population be made. However, there is evidence that the treatment of yoga helped these children and this study could warrant further investigations.

5D.

Problem

Does a treatment of yoga increase the ability of children with autism to solve problems?

Testing Assumptions

A visual investigation of the data shows that the before data is slightly left skewed, but a QQ-plot of the data reveals that it is fairly normal. The after treatment times also appear normal. However, there is so little data that a parametric test could result in misleading results. A paired t-test will be used to investigate.

Hypothesis

Yoga increases the ability of children with autism to solve problems.

- H0: There is no difference in the distribution of times it takes to solve a puzzle before or after a treatment of yoga.
- HA: The distribution of the times of children treated with yoga are less than the times before a treatment of yoga.

Find the Critical Value

- Assuming that alpha = 0.05
- Using a two tailed test
- Degrees of freedom = 8
- Critical value for two tailed alpha of 0.05 with 8 degrees of freedom: 1.859

Analysis

SAS software was used to conduct a paired t-test using PROC TTEST. This procedure resulted in a t-value of 2.54 which exceeds the critical value of 1.859 which was needed to reject the null hypothesis. Furthermore, the p-value associated with this t-value is 0.0347. This indicates that the result would only be expected by random change 347 out of 10,000 tries. There is sufficient evidence to reject the null hypothesis (p-value of 0.0347).

Conclusion

There is strong evidence to suggest that a treatment of yoga increases the problem solving ability in children with autism (p-value of 0.0347).

Score of inference

Since this is an observational study with no randomization processes no casual inference can be made, nor can inferences to the larger population be made. However, there is evidence that the treatment of yoga helped these children and this study could warrant further investigations.

Code Used to Reach This Conclusion

```
/*5D same but now with paird t test */
proc ttest data=work.yoga2;
      paired before*after;
run;
```

5E.

My results from R are not matching SAS. I think it has to do with the degrees of freedom shown below. In R my DF = 15.718, and in SAS my DF = 8. There is a large difference here and I believe the DF should be 8. I tried running the R code several times and could not correct this problem. If you have insight on how to correct this problem please let me know.

```
        Welch Two Sample t-test

data:  yoga$Before and yoga$After
t = 1.9927, df = 15.718, p-value = 0.06396
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.198881 37.865547
sample estimates:
mean of x mean of y
 62.22222  43.88889
```

```
t.test(yoga$Before,yoga$After, paird=TRUE)
```

5F

Problem

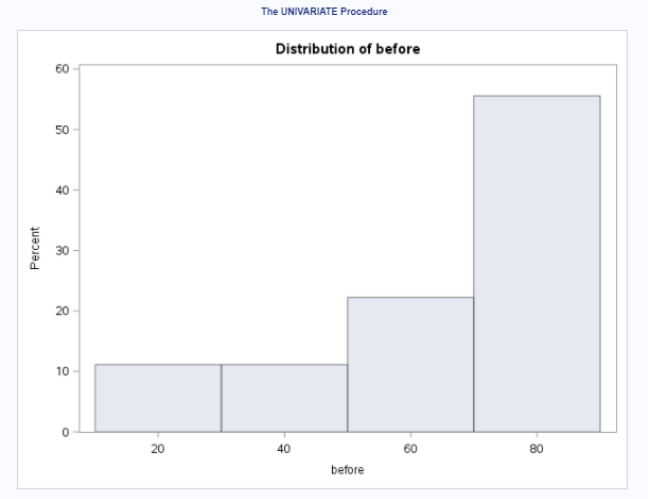
Does a treatment of yoga increase the ability of children with autism to solve problems?

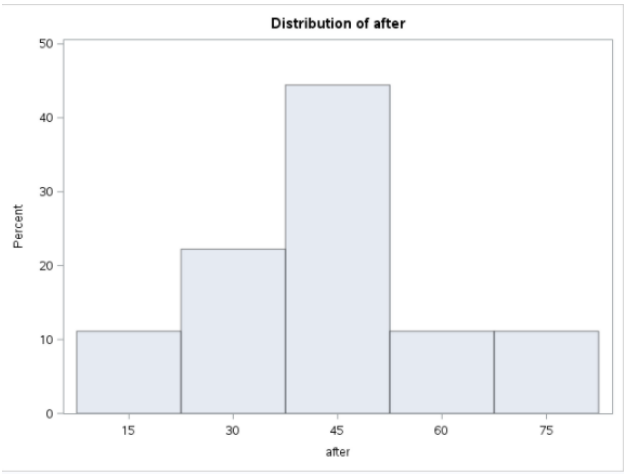
Testing Assumptions

A visual investigation of the data shows that the before data is slightly left skewed and the after data is fairly normal. QQ-plots of the data reveal that both the before and after times are fairly normal as well. The histograms and QQ-plots are shown for the before and after treatments, respectively.

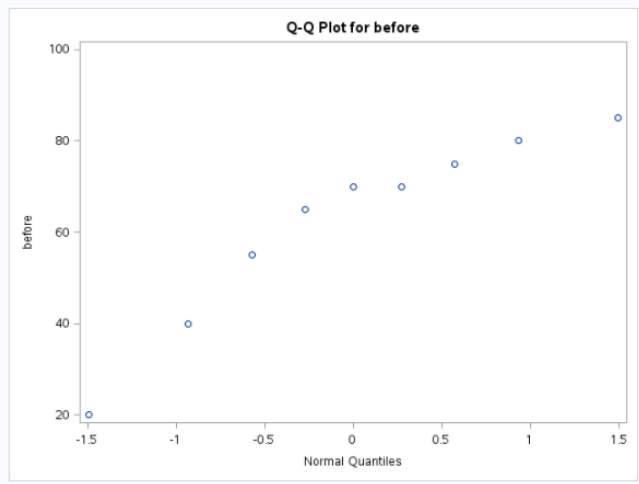
However, there is so little data that a parametric test could result in misleading results. A sign-rank test is being used to test the problem.

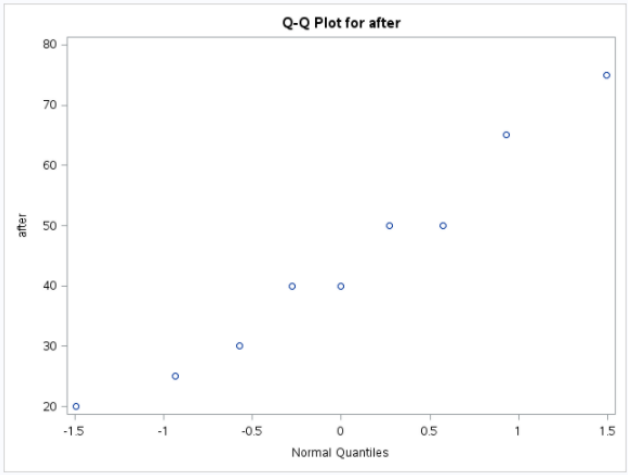
Histograms: Before and After Yoga Treatment





QQ-Plots: Before and After Yoga Treatment





Kyle Thomas HW 4

Hypothesis

Yoga increases the ability of children with autism to solve problems.

- H0: There is no difference in the distribution of times it takes to solve a puzzle before or after a treatment of yoga.
- HA: The distribution of the times of children treated with yoga are less than the times before a treatment of yoga.

Analysis

SAS software was used to conduct a signed rank test on the difference in times before and after the treatment of yoga using PROC UNIVARIATE. There is sufficient evidence to reject the null hypothesis (p-value of 0.0313).

Conclusion

There is strong evidence to suggest that a treatment of yoga increases the problem solving ability in children with autism (p-value of 0.0313). This indicates that the odds of seeing this result are slightly more than 3 percent.

Score of inference

Since this is an observational study with no randomization processes no casual inference can be made, nor can inferences to the larger population be made. However, there is evidence that the treatment of yoga helped these children and this study could warrant further investigations.