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.

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

## Problem

The Timber industry argues that logging burned trees enhances forest recovery and decreases the percentage of seedlings lost.

The EPA argues that not logging burned trees enhances forest recovery and decreases the percentage of seedlings lost.

## Identify the Null Hypothesis

***H*0: *The distribution of Logging Scores = The distribution of the Not Logging Method Scores***

***H*1:*The distribution of Logging Scores > The distribution of the Not Logging Method Scores***

## Assumptions:

Since a rank sum test is non-parametric, we do not have to test to see if the data is normally distributed. However, other assumptions apply:

1. All observations are independent - We will assume independence as we do not have a lot of specifics.
2. The Y values are ordinal: Here the y values are ordinal since they represent the percentage lost

## Code:

proc npar1way data = logging WILCOXON ALPHA = .05;

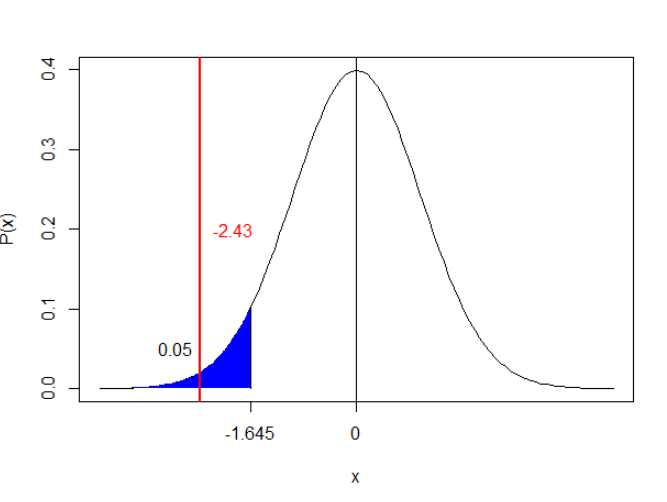
class Action;

var PercentLost;

exact HL;

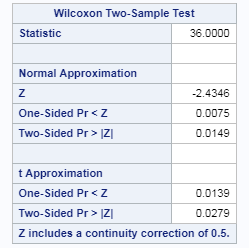
run;

## Critical Value:



## P- Value:

.0075

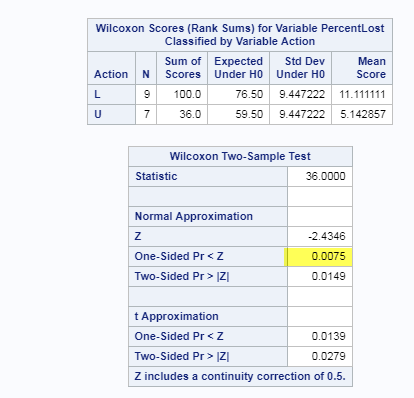


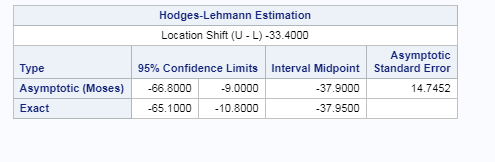
## Null Hypothesis:

We reject the null hypothesis

## Results:

There is evidence (alpha = .05) to suggest that the distribution of scores from the Logging method is greater than the distribution of the Not Logging method. (Normal approx. to rank-sum test p-value = .0075 w/ Continuity correction of .5. A range of plausible values for how much smaller the logging distribution is than the non-logging distribution is [-65.1,-10.8] (95% CI based on rank-sum test) with a point-estimate of -37.9.





## Scope of Inference:

This is an observational study given that there was no experimental design here because of no randomization of what area was logged vs not. Also, the results of this experiment only applies to this particular forest as these type of trees may behave differently over others in other forests or other land conditions.

* 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.

Code:

a = c(-20.1,-8.1,13.3,18.1,23.6,34.2,56.1)

b = c(18.2,40.8,43.2,45,46.7,53.1,75.5,85.4,85.6)

wilcox.test(U,L, alt = "two.sided", mu = 0, conf.int = T, conf.level = .95, exact = T, correct = T)

Output:

Wilcoxon rank sum test

data: U and L

W = 8, p-value = 0.01154

alternative hypothesis: true location shift is not equal to 0

95 percent confidence interval:

-65.1 -10.8

sample estimates:

difference in location

-33.4

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.
   2. Show all 6 steps (including a thoughtful, thorough, yet non-technical conclusion. Include a confidence interval.
   3. Include a scope of inference at the end. (You may copy and paste this from a previous HW if you like.)
   4. 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.
   5. 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.)

## Problem:

A survey was conducted for subjects who had 12 or 16 years of education. We would like to know if the income of those with 16 years of education exceed those with 12 years of education.

## Assumptions

Since a rank sum test is non-parametric, we do not have to test to see if the data is normally distributed. However, other assumptions apply:

1. All observations are independent - We will assume independence given patients can’t be both.
2. The Y values are ordinal

## Step 1: Identify the Null Hypothesis

***H*0: *The distribution of Logging Scores = The distribution of the Not Logging Method Scores***

***H*1:*The distribution of Logging Scores > The distribution of the Not Logging Method Scores***

## Step 2: Draw, Shade, and fine the critical value

Df = 473.85 for pooled

data test1;

alpha = .05;

p = 1-alpha; /\*/2;\*/

df = 473.85;

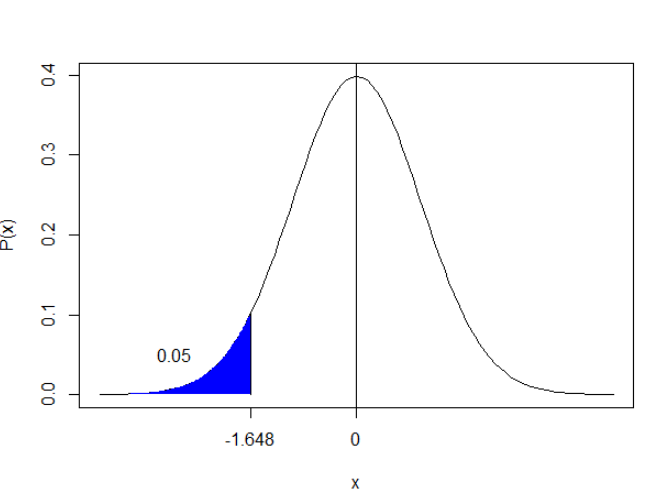
CritVal = TINV(p, df);

proc print data = test1;

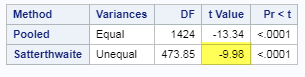
run;

Critical value = 1.64808

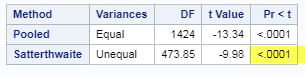
DF: 473.85



## Step 3: Find the Test Statistic (t-value)



## Step 4: Find the p-value



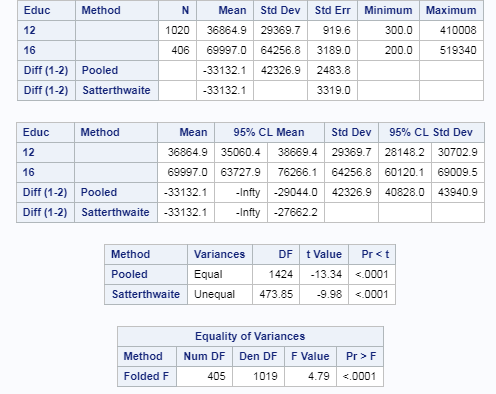
P-value (.0001)

## Step 5: Reject H0 or Accept Ha

We reject the null hypothesis.

## Step 6: Conclusion

There is sufficient evidence to suggest that the mean income is greater for individuals with 16 years of educational experience vs those with 12 years educational experience (p-value .0001 from a two sample Welch test). The estimated effect is -33132.1 dollars (95% confidence interval of [-infinity, -27662.2].



## Scope of Inference

This is an observational study since we cannot randomize the income between the subjects and it appears that this survey was a voluntary survey so we might not have all incomes. Also, we can only infer the results to be true for those who participated in the survey with ages between 41-49.

## Verify in R

**Code:**

setwd("C:/Users/Marin Family/Desktop/Statistical Foundations for Data Science/Unit3")

EducationData <- read.csv("EducationData.csv")

t.test(Income2005 ~ Educ, data = EducationData, var.equal = FALSE, conf.level = .95)

**Output:**

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:

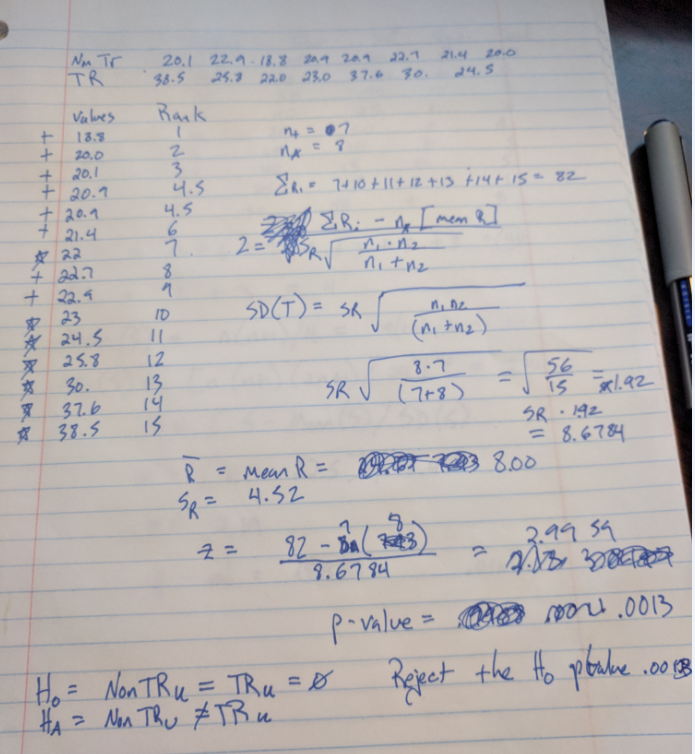
mean in group 12 mean in group 16

36864.90 69996.97

## Preference over Log Transform

I prefer the Welch test over the Log Transformation given that it is a non-parametric test. For the log transformation, you have to test for normality and if it doesn’t pass, then you can transform the data, but still have to check for normality again. Also, it is possible that the log transform may not be appropriate like the Nasa O-Ring experiment. As long as the data is independent and y is ordinal, this test works.

* 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.)



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

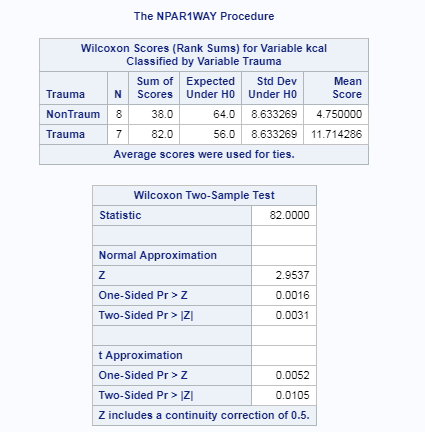
proc npar1way data = Trauma WILCOXON ALPHA = .05;

class Trauma;

var kcal;

exact HL;

run;



No, the p-value is slightly different. .0016 in SAS vs .0011 by hand. Yes, a continuity correction is used of .5.

* 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.
     2. 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.
     3. 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).
     4. Also include a scope of inference statement.

## Problem

Are metabolic expenditures different between NonTrauma patients and Trauma patients?

## Assumptions

Since a rank sum test is non-parametric, we do not have to test to see if the data is normally distributed. However, other assumptions apply:

1. All observations are independent - We will assume independence as we do not have a lot of specifics.

The Y values are ordinal

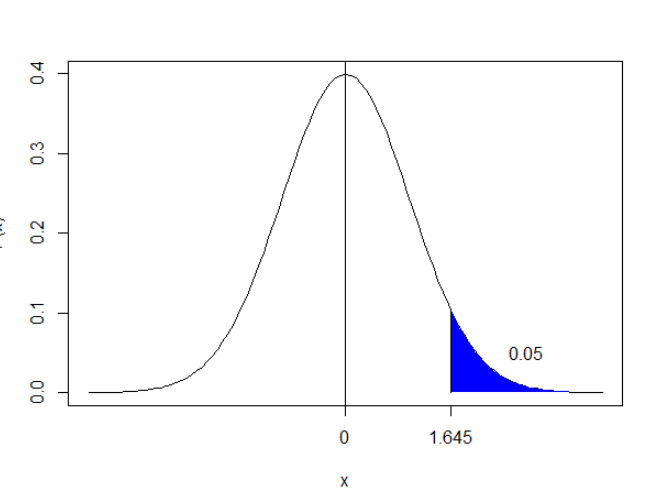
## Step 1: Identify the H0 and Ha

***H*0: *The distribution of Trauma Expenditures = The distribution of the Non-Trauma Expenditures***

***H*1:*The distribution of Trauma Expenditures < The distribution of the Non-Trauma Expenditures***

## Step 2: Find the Critical Value

1.645



## Step 3: Find the t-value

Code:

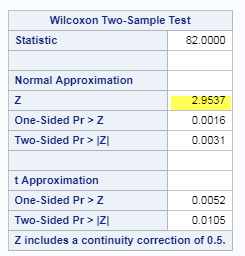
proc npar1way data = Trauma WILCOXON ALPHA = .05;

class Trauma;

var kcal;

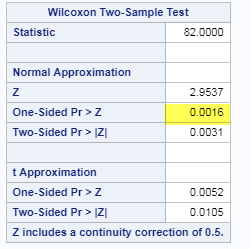
exact HL;

run;



Z = 2.9537

## Step 4: Find the p-value



p-value = .0016

## Step 5: Fail to reject H0

Fail to reject the null hypothesis

## Step 6: Conclusion

There is sufficient evidence at the alpha of .5 level of significance (p-value = .0016) to suggest that the distribution of scores between trauma patients are greater than non-trauma patients with a 95% confidence interval of [1.9,16.7].

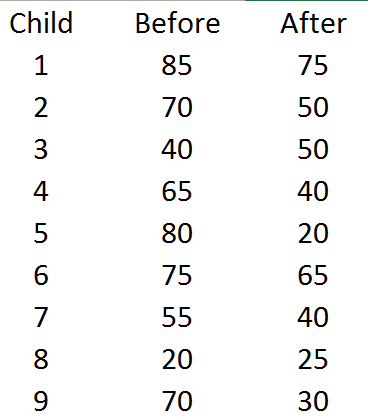
## Confidence Interval

## 

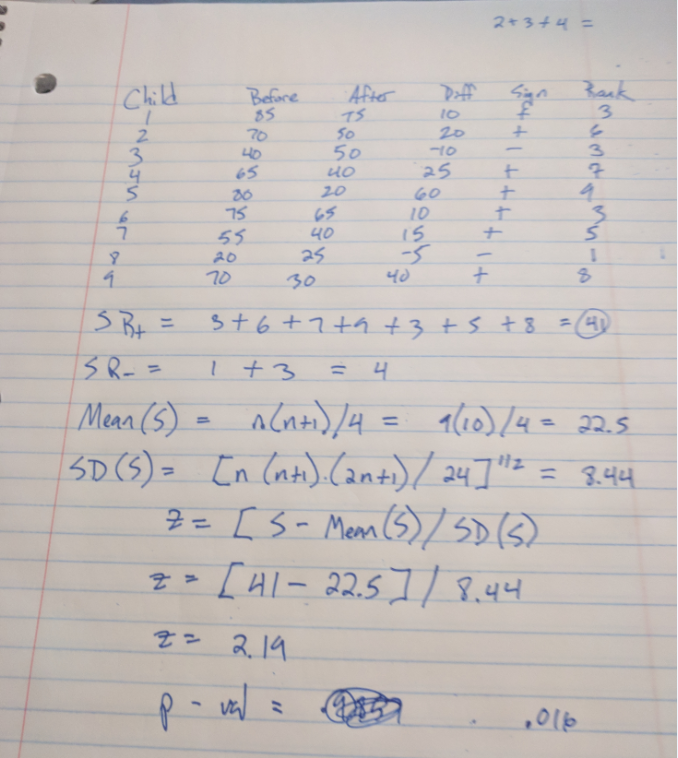
## Scope of Inference

There is no randomization here and no experimental design. This is an observational study from patients coming into the hospital. These results only apply to the sample that was observed.

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:



* 1. 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).



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

Code:

*data Work.Child2;*

*set Work.Child;*

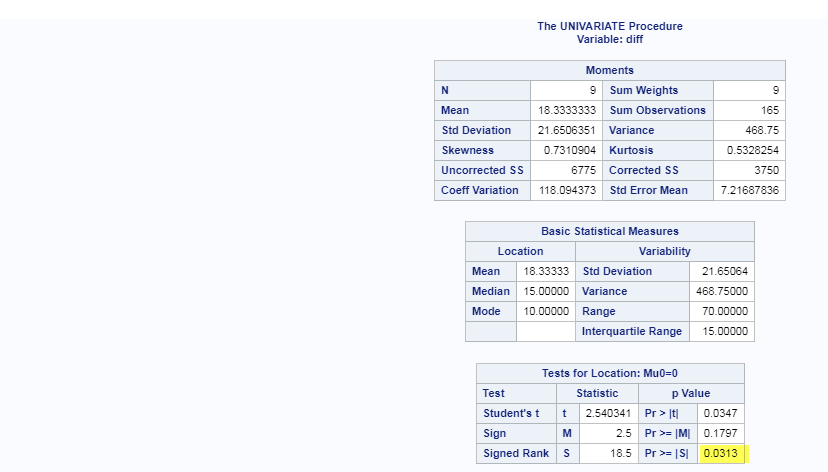
*diff = Before - After;*

*run;*

*proc univariate data = Child2;*

*var diff;*

*run;*



2 Sided p- value = .0313. One side is .0313/2 = .01565

R:

wilcox.test(data1$Before, data1$After, paired = TRUE)

Wilcoxon signed rank test with continuity correction

data: data1$Before and data1$After

V = 41, p-value = 0.03236

alternative hypothesis: true location shift is not equal to 0

That is 2 sided p-value. 1-sided p-value = .1618

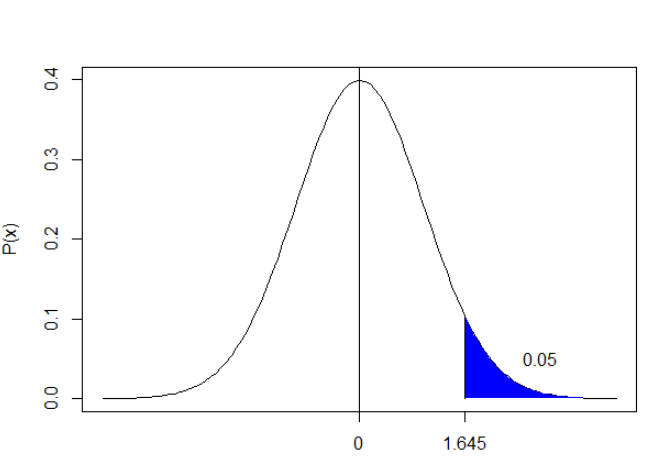
* 1. Conduct the six step hypothesis test using your calculations from above to test the claim that the yoga treatment was effective in reducing the time to finish the puzzle.

## Step 1: Declare H0 and Ha

H0: The median difference in time to complete puzzle before and after yoga is zero

HA: The median difference in time to complete puzzle before and after yoga is positive

## Step 2: Find critical value



## Step 3: Find t-value/z

2.19

## Step 4: Find P-value

.016

## Step 5: Reject or Fail to Reject

Reject the H0

## Step 6: Conclusion

There is strong evidence that the median time to complete a puzzle after yoga is less than the median time to complete a puzzle before yoga with a Wilcoxn signed rank test one-sided p-value of .016.

## Step 7: Scope of Inference

This study is an observational study and the results only apply to the group that participated.

* 1. 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.

**Code:**

*proc ttest data = Work.Child Alpha = .05 Side = U;*

*Paired Before\*After;*

*run;*

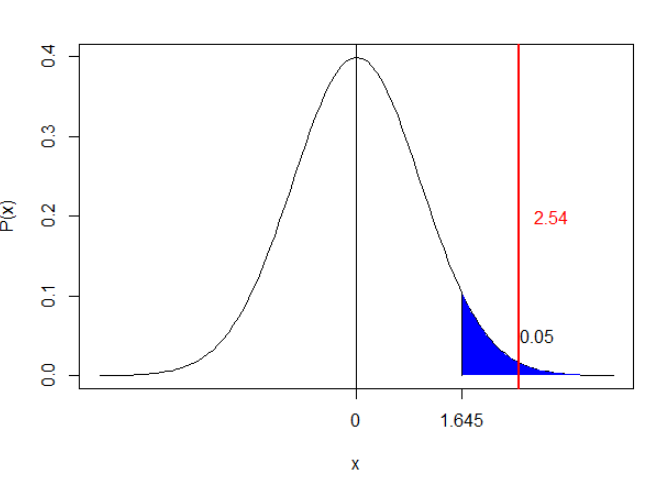
## Step 1: Declare H0 and Ha

H0: uYoganPuzzle – uNoYogaPuzzle = 0

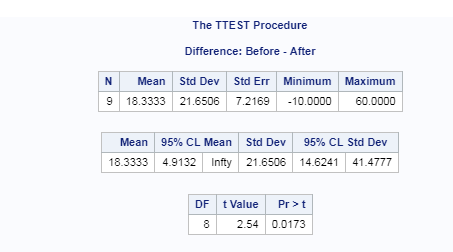
Ha: uYogaNPuzzle – uNoYogaNPuzzle > 0

## Step 2: Find critical value

1.645 with alpha = .05



## Step 3: Find t-value/z



T-Value = 2.54

## Step 4: Find P-value

P-value = .0173

## Step 5: Reject or Fail to Reject

Reject the H0

## Step 6: Conclusion with confidence interval

There is evidence to suggest that on average, the time on average was less to complete a similar puzzle after doing yoga with a p-value of .0173. A 95% confidence interval for the mean difference in scores is (14.62,41.47)

## Step 7: Scope of Inference

This study is an observational as the parents volunteered for their kids to be experimented on, but there is no randomization. The study only applies to the subjects tested.

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

t.test(data1$Before, data1$After, paired = TRUE, mu = 0, )

Paired t-test

data: data1$Before and data1$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

Results above are two-sided.

* 1. 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:

Does the treatment of yoga help children with autism complete puzzles faster?

**Experiment:**

Parents volunteered their children to complete a puzzle before yoga and a similar puzzle after 20 mins of yoga.

## Assumptions:

**All observations are indepdent**

**The y Values are ordinal.**

**Both assumptions are met.**

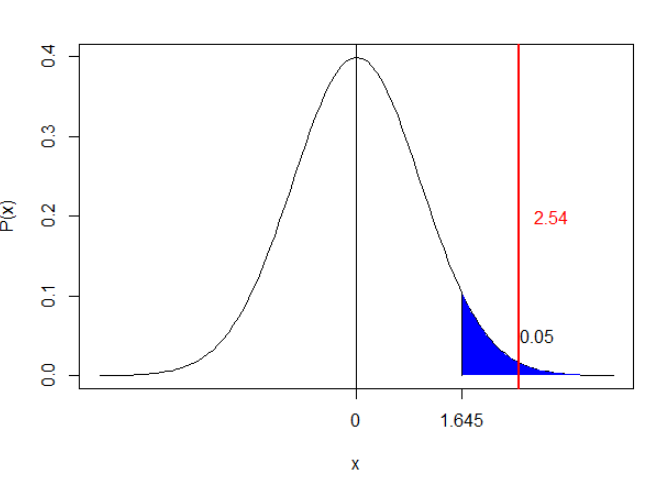
## Step 1: Declare H0 and Ha

H0: uYoganPuzzle – uNoYogaPuzzle = 0

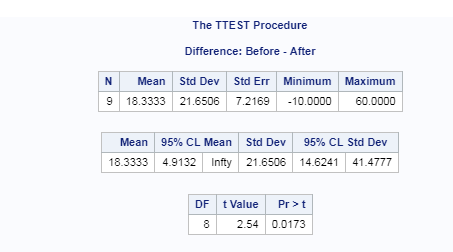
Ha: uYogaNPuzzle – uNoYogaNPuzzle > 0

## Step 2: Find critical value

1.645 with alpha = .05



## Step 3: Find t-value/z



T-Value = 2.54

## Step 4: Find P-value

P-value = .0173

## Step 5: Reject or Fail to Reject

Reject the H0

## Step 6: Conclusion with confidence interval

There is evidence to suggest that on average, the time on average was less to complete a similar puzzle after doing yoga with a p-value of .0173. A 95% confidence interval for the mean difference in scores is (14.62,41.47)

## Step 7: Scope of Inference

This study is an observational as the parents volunteered for their kids to be experimented on, but there is no randomization. The study only applies to the subjects tested.

**Though needed, I unfortunately didn’t have time for this.**

**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:

* 1. 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.
  2. 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!). |  |