choi6301_hw4

2024-09-29

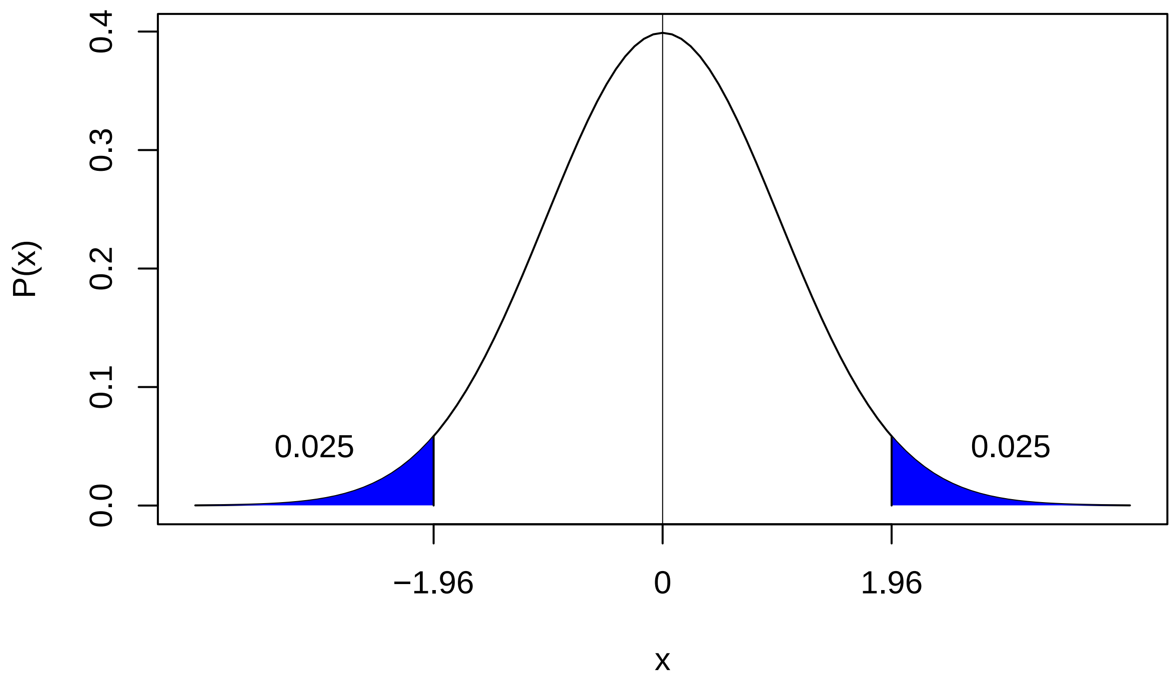
total 86

Question2

- Part A

  - Problem statement: we want to prove logging actually increases the percentage of seedling lost in the time span studied.
  -
$$H_0 : distributionofunlogged(U) = distibutionoflogged(L)$$
$$distributionofunlogged(U) < distibutionoflogged(L)$$
$$\alpha \ = 0.05$$
  - Critical Value(left_sided): -1.96
  - value of Test Statistic: z = -3.2814
  - p-value: 0.0001
  - Conclusion: The data provide convincing evidence that logging the burned trees enhances forest recovery after "logged(L)" rather than the "unlogged(U)" method (one-sided, normal approximation with p-value=0.0005, from the rank-sum test). A range of plausible values for how much smaller the "logged(L)" distribution is than the "unlogged(U)" is [-41.2, -18.8]times.(95% confidence interval based on a rank-sum test) with a point-estimate of -28.4 times.

```
crit.value <- qt(0.90, 15, lower.tail=T)
shade(100000, 0.05, 0, t_calc=NULL, sides='both')
```




```
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q1.jpg")
```

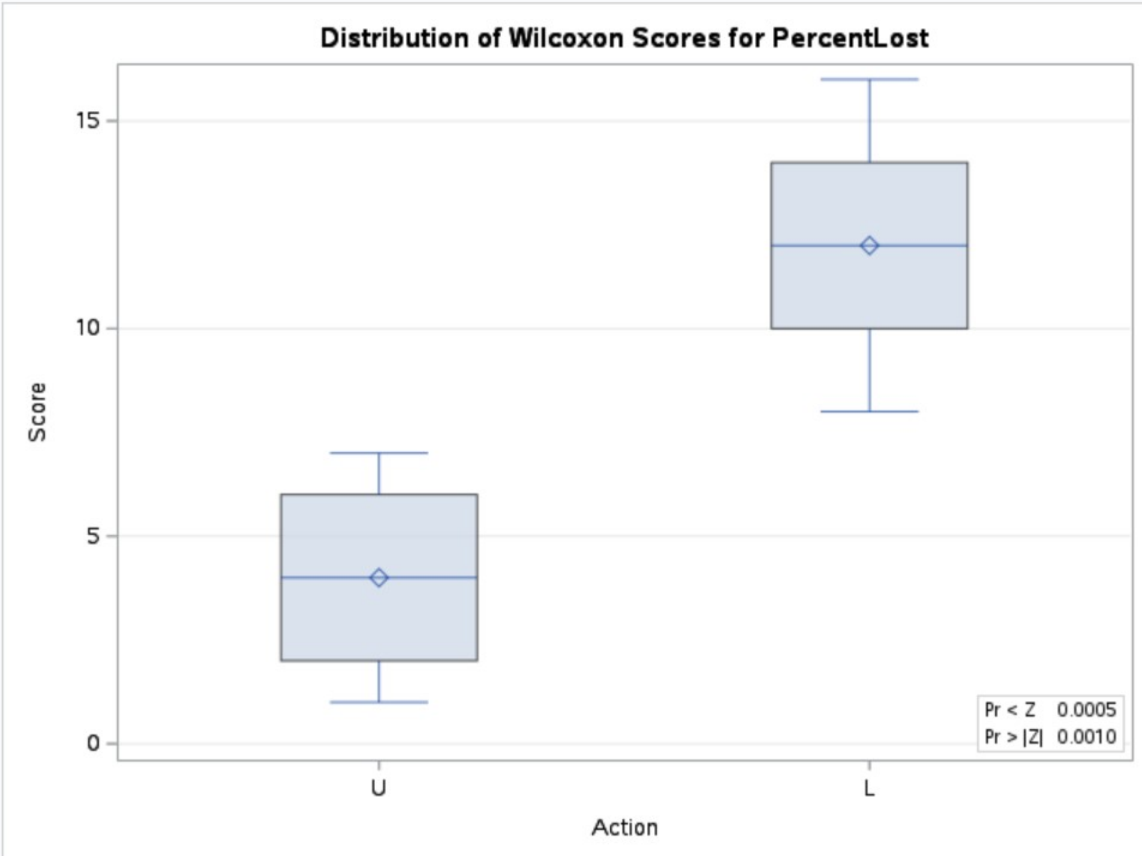
The NPAR1WAY Procedure

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
U	7	28.0	59.50	9.447222	4.0
L	9	108.0	76.50	9.447222	12.0

Wilcoxon Two-Sample Test							
				t Approximation		Exact	
Statistic (S)	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|	Pr <= S	Pr >= |S-Mean|
28.0000	-3.2814	0.0005	0.0010	0.0025	0.0050	<.0001	0.0002
Z includes a continuity correction of 0.5.							

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
11.1176	1	0.0009

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q1_2.jpg")
```




```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q1_3.jpg")
```

Hodges-Lehmann Estimation				
Location Shift (U - L) -28.4000				
Type	90% Confidence Limits		Interval Midpoint	Asymptotic Standard Error
Asymptotic (Moses)	-42.3000	-18.5000	-30.4000	7.2347
Exact	-41.2000	-18.8000	-30.0000	

- Part B

```r
##Training Data
Logg <- read.csv('C:/Users/choih/OneDrive/Desktop/Logged6301_2024.csv')
Logg$Action <- factor(Logg$Action)

##Note that your grouping variable MUST be a factor
```

```r
##Exact Rank Sum Test
wilcox_test( PercentLost ~ Action, data=Logg, alternative='greater', conf.level=0.90, distribution='exa
```

```
##
##  Exact Wilcoxon-Mann-Whitney Test
##
## data:  PercentLost by Action (L, U)
## Z = 3.3343, p-value = 8.741e-05
## alternative hypothesis: true mu is greater than 0
```

```r
##Normal Approximation to the Rank Sum Test
wilcox_test(PercentLost ~ Action, data=Logg, alternative='greater', conf.level=0.90, distribution='appr
```

```
##
##  Approximative Wilcoxon-Mann-Whitney Test
##
## data:  PercentLost by Action (L, U)
## Z = 3.3343, p-value < 1e-04
## alternative hypothesis: true mu is greater than 0
```

```r
##Exact Rank Sum Test w/ confidence interval
wilcox_test(PercentLost  ~ Action, data=Logg, alternative='two.sided', conf.int=T, conf.level=0.90, dis
```

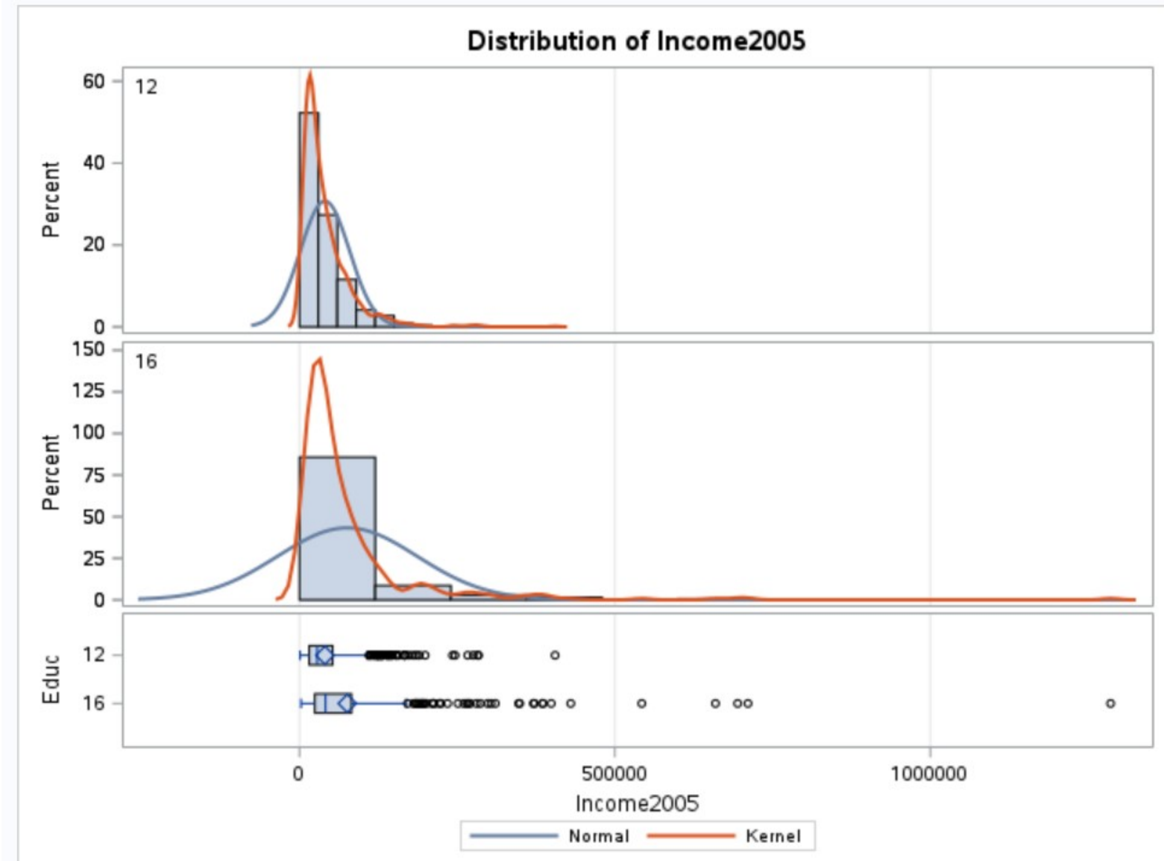
```
##
##  Approximative Wilcoxon-Mann-Whitney Test
##
## data:  PercentLost by Action (L, U)
## Z = 3.3343, p-value = 1e-04
## alternative hypothesis: true mu is not equal to 0
## 90 percent confidence interval:
##  18.8 41.2
## sample estimates:
## difference in location
##                   28.4
```

Question3

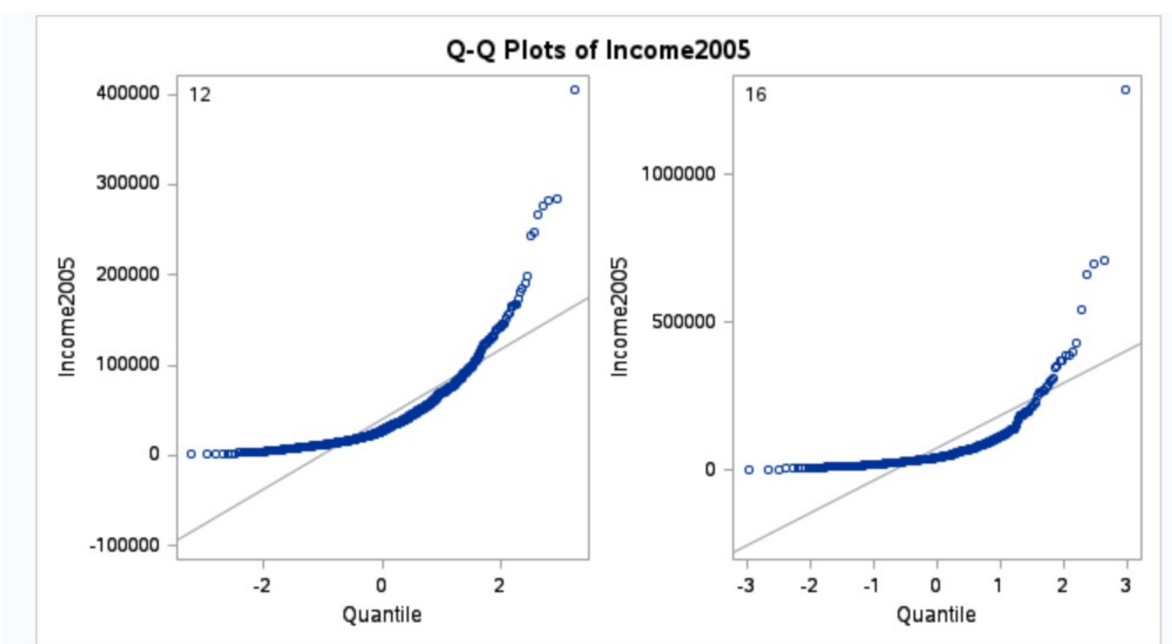
- Part A

  - Normality: There is enough evidence from the histograms and QQ plots of drastic departures from normality. We will assume that the samples sizes are large enough for CLT to hold.
  - Equal standard deviations: There is enough evidence suggest drastic differences in the population standard deviations, thus we will assume that the standard deviations are not equal
  - Independence: We will assume that the observations are independent both between and within groups.
  - Decision: The two sample t-test and confidence intervals are not appropriate to use for these data.

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q2.jpg")
```




```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q2_2.jpg")
```



- Part B

  - Problem statement: we want to prove from this data set is educated college people(16 years) makes more income than educated high school people(12 years).
  -
$$H_0 : \mu_{12} = \mu_{16}$$
$$H_A : \mu_{12} < \mu_{16}$$
$$\alpha \ = 0.05$$
  - Critical Value: -1.646
  - value of Test Statistic: -6.32
  - p-value: 2.2e-16
  - Conclusion: There is strong evidence to suggest that the mean income of the high school educated people group is less than the mean income of the college educated people group(p=2.2e-16). A95% confidence interval for the difference is $[-\infty$,-26278.67].

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q3.jpg")
```

The TTEST Procedure

Variable: Income2005

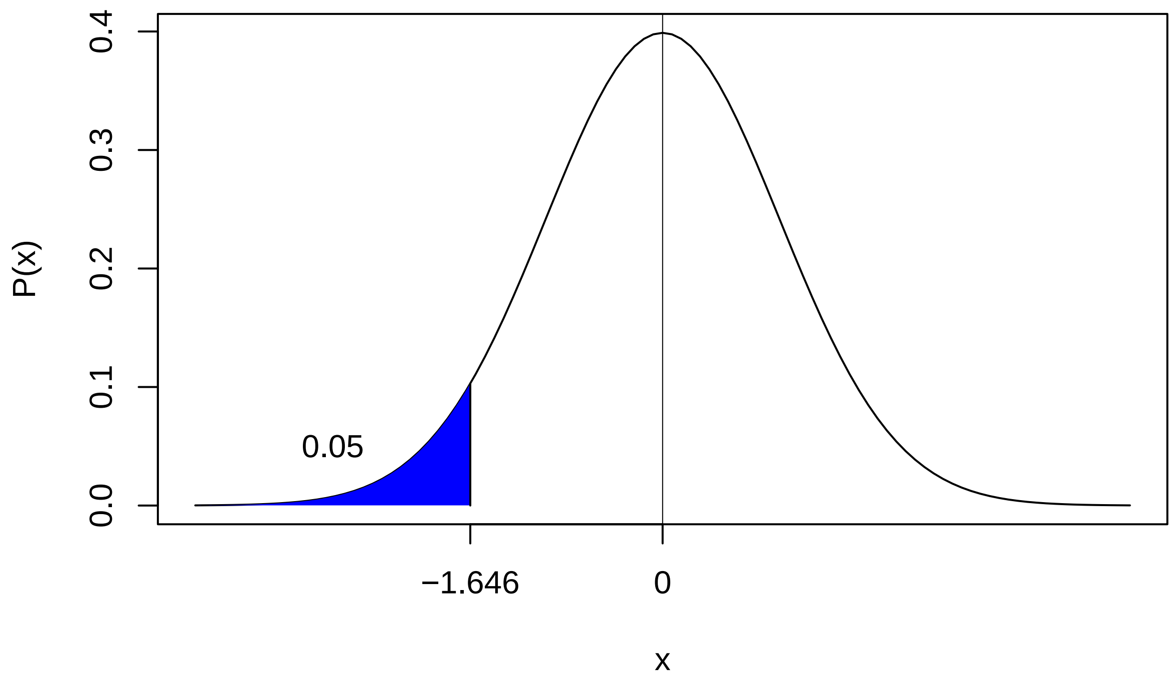
Educ	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
12		1020	40297.0	38943.8	1219.4	1041.4	405216
16		406	75841.5	110574	5487.7	2478.0	1285898
Diff (1-2)	Pooled		-35544.5	67547.4	3963.7		
Diff (1-2)	Satterthwaite		-35544.5		5621.5		

Educ	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12		40297.0	37904.2	42689.8	38943.8	37324.1	40711.5
16		75841.5	65053.6	86629.5	110574	103456	118752
Diff (1-2)	Pooled	-35544.5	-Infty	-29020.5	67547.4	65155.3	70123.1
Diff (1-2)	Satterthwaite	-35544.5	-Infty	-26278.7			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	1424	-8.97	<.0001
Satterthwaite	Unequal	445.55	-6.32	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	405	1019	8.06	<.0001

```r
crit.value <- qt(0.95, 1425, lower.tail=T)
shade(1425, 0.05, 0, t_calc=NULL, sides='left')
```



- Part C
  - 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 subject of this sample.
- Part D

```
education <- read.csv("C:/Users/choih/OneDrive/Desktop/EducationData6301.csv")
education$Educ <- factor(education$Educ)
par(mfrow=c(2,2))


t.test(Income2005 ~ Educ, data=education, alternative='less')
```

```
##
##  Welch Two Sample t-test
##
## data:  Income2005 by Educ
## t = -6.3229, df = 445.55, p-value = 3.122e-10
## alternative hypothesis: true difference in means between group 12 and group 16 is less than 0
## 95 percent confidence interval:
##       -Inf -26278.67
## sample estimates:
```

```
## mean in group 12 mean in group 16
##         40296.99         75841.53
```

- Part E
    - Compared to log transformed and Welch's analysis I think Welch's analysis is more appropriate. Since we have enough sample size to invoke the CLT. It is robust to different standard deviations even when the sample size is not equal.

Question4

- Part A

- the data provide convincing evidence that trauma patients could have higher metabolic expenditures than other reasons patients(p=0.0006).

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q4_1.jpg")
```

a) Rank transformations for the data

1	NT	18.8
2	NT	20
3	NT	20.1
4	NT	20.9
5	NT	20.9
6	NT	21.4
7	T	22
8	NT	22.7
9	NT	22.9
10	T	23
11	T	24.5
12	T	25.8
13	T	30
14	T	37.6
15	T	38.5

78

82

b) Calculate the rank-sum

$$\bar{R} = 8$$

$$S_R = \sqrt{\frac{(1-8)^2 \cdots (15-8)^2}{14}} = 4.472$$

$$\text{Mean}(NT) = 8 \times 8 = 64$$

$$\text{Mean}(T) = 7 \times 8 = 56$$

$$SD(T) = 4.472 \sqrt{\frac{56}{7+8}} = 8.6407$$

$$Z_{NT} = \frac{38 - 64 + 0.5}{8.6407} = -2.9511$$

$$Z_T = \frac{82 - 56 - 0.5}{8.6407} = 2.9511$$

```
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q4_2.jpg")
```

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Metapolic Classified by Variable trauma					
trauma	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
NT	8	38.0	64.0	8.633269	4.750000
T	7	82.0	56.0	8.633269	11.714286
Average scores were used for ties.					

Wilcoxon Two-Sample Test							
				t Approximation		Exact	
Statistic (S)	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|	Pr >= S	Pr >= |S-Mean|
82.0000	2.9537	0.0016	0.0031	0.0052	0.0105	0.0006	0.0012
Z includes a continuity correction of 0.5.							

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
9.0698	1	0.0026

- Part B

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q4_2.jpg")
```

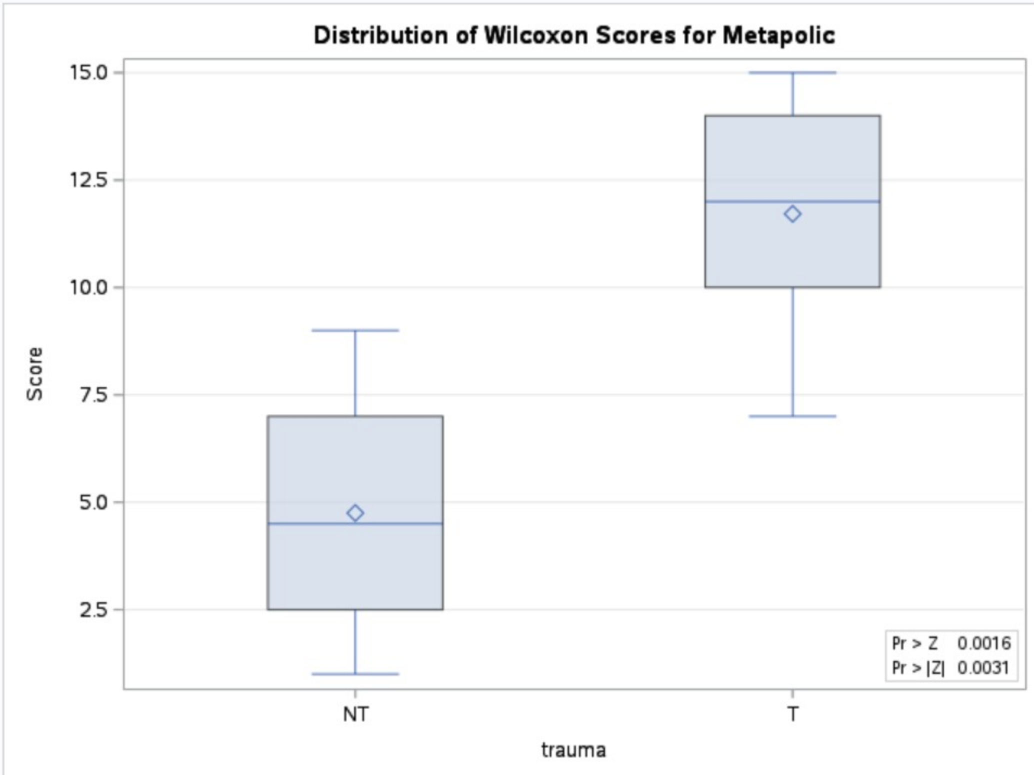
The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable Metapolic Classified by Variable trauma					
trauma	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
NT	8	38.0	64.0	8.633269	4.750000
T	7	82.0	56.0	8.633269	11.714286
Average scores were used for ties.					

Wilcoxon Two-Sample Test							
				t Approximation		Exact	
Statistic (S)	Z	Pr > Z	Pr > |Z|	Pr > Z	Pr > |Z|	Pr >= S	Pr >= |S-Mean|
82.0000	2.9537	0.0016	0.0031	0.0052	0.0105	0.0006	0.0012
Z includes a continuity correction of 0.5.							

Kruskal-Wallis Test		
Chi-Square	DF	Pr > ChiSq
9.0698	1	0.0026

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q4_3.jpg")
```




```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q4_4.jpg")
```

Hodges-Lehmann Estimation				
Location Shift (T - NT) 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	

The TTEST Procedure

Variable: Metapolic

trauma	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
NT		8	20.9625	1.3794	0.4877	18.8000	22.9000
T		7	28.7714	6.8354	2.5835	22.0000	38.5000
Diff (1-2)	Pooled		-7.8089	4.7528	2.4598		
Diff (1-2)	Satterthwaite		-7.8089		2.6292		

trauma	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
NT		20.9625	19.8093	22.1157	1.3794	0.9120	2.8074
T		28.7714	22.4498	35.0931	6.8354	4.4047	15.0520
Diff (1-2)	Pooled	-7.8089	-13.1230	-2.4949	4.7528	3.4455	7.6569
Diff (1-2)	Satterthwaite	-7.8089	-14.1398	-1.4781			

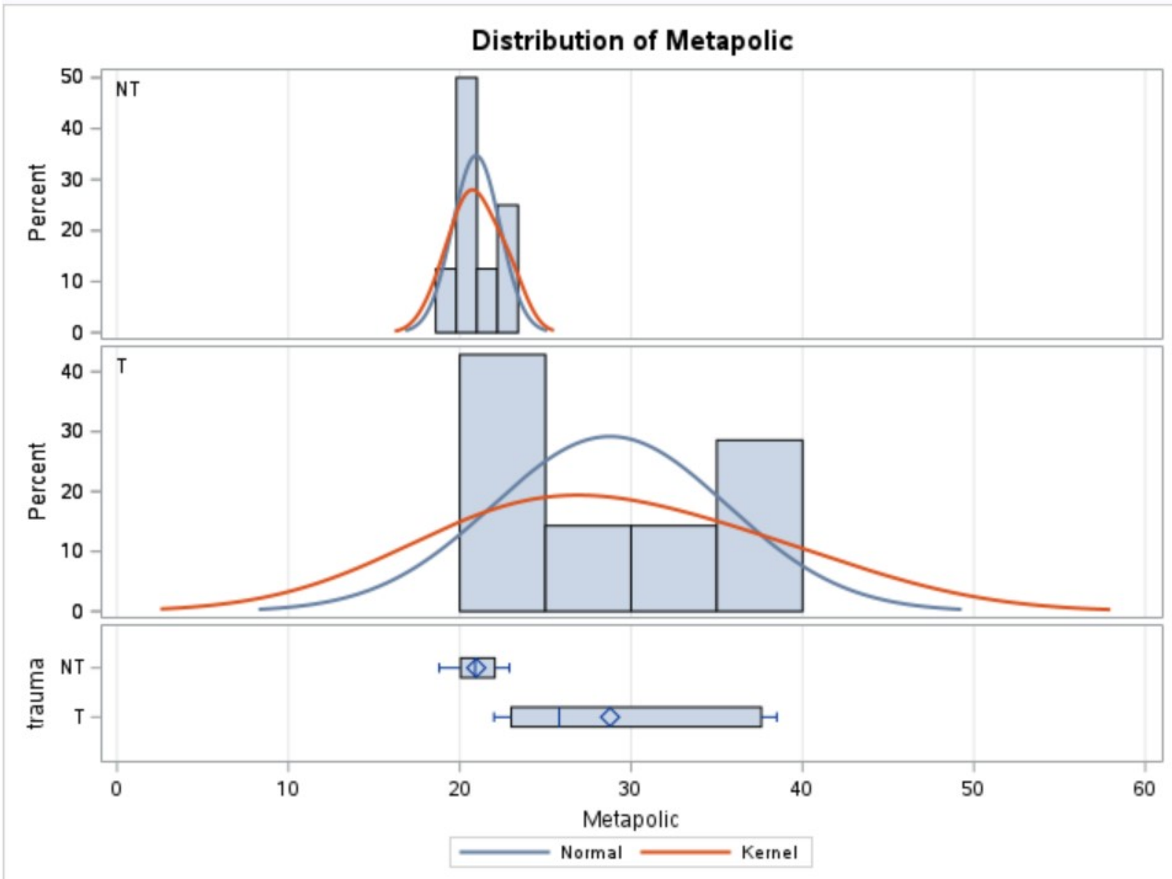
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	13	-3.17	0.0073
Satterthwaite	Unequal	6.4282	-2.97	0.0230

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6	7	24.56	0.0005

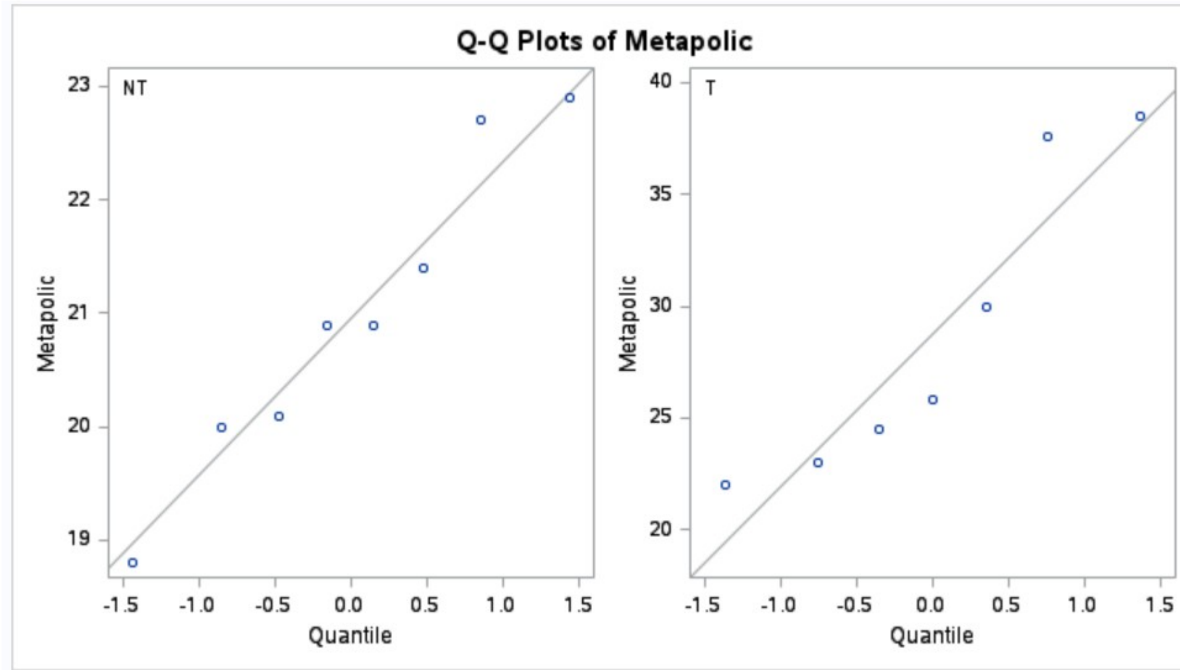
- Part C
  - Problem statement: We want to prove from this data set is the trauma patients has more higher metabolic than none trauma patients.
  - $$H_0 : \mu_{\text{None trauma}} = \mu_{\text{Trauma}}$$
    $$H_A : \mu_{\text{None trauma}} < \mu_{\text{Trauma}}$$
    $$\alpha \ = 0.1$$
  - Test Statistic: $= \pm\ 2.9511$
  - p-value: 0.00016
  - Conclusion: the data provide convincing evidence that trauma patients could have higher metabolic expenditures than other reasons patients(p=0.00016). A range of plausible values

for how much higher the “trauma patients” distribution is than the “None trauma patients” us [2.1000, 15.6000] with a point-estimate of 5.3000.

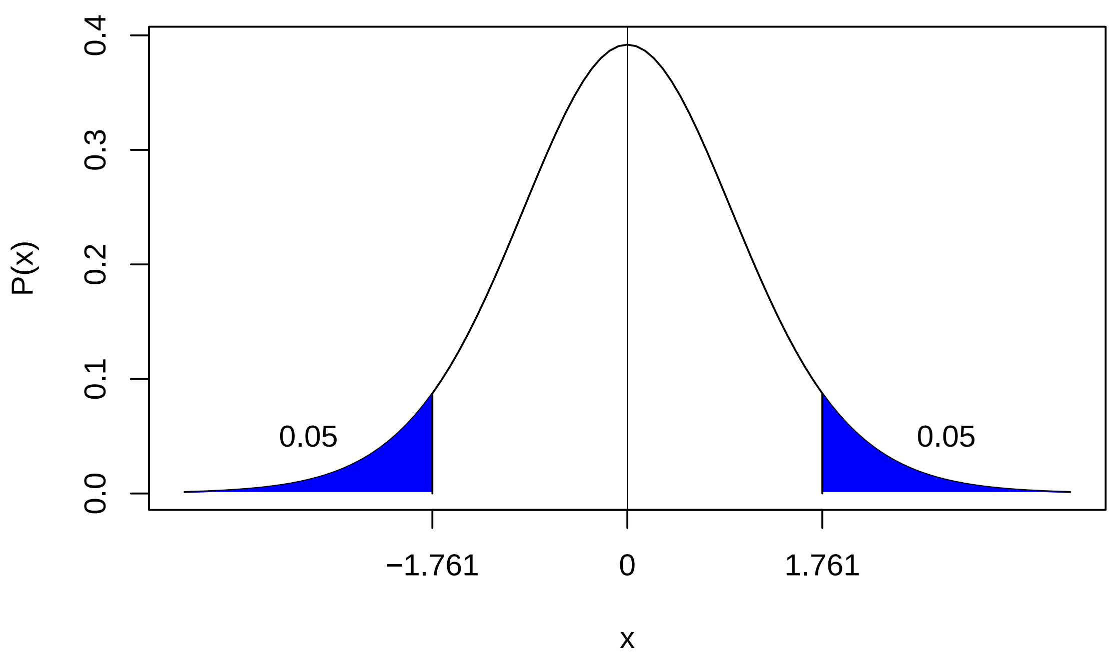
```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q4_5.jpg")
```




```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q4_6.jpg")
```




```
shade(14, 0.1, 0, t_calc=NULL, sides='both')
```



Question5

- Part A

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q5.jpg")
```

R	Child	before	After	|D|	S
3	1	85	75	10	+
6	2	70	50	20	+
3	3	40	40	-10	-
7	4	65	65	25	+
9	5	80	80	60	+
3	6	75	75	10	+
5	7	55	55	15	+
1	8	20	20	-5	-
8	9	70	70	40	+

$K = 7$

$S = 41$

$$S = \frac{9(9+1)}{4} = 22.5$$

$$SD(S) = \sqrt{\frac{(90)(19)}{24}} = 8.441$$

$$Z = \frac{S - mean(S) - 0.5}{SD(S)}$$

$$Z = \frac{41 - 22.5 - 0.5}{8.441} = 2.1324$$

$$p\text{-value} = 0.0313/2 = 0.01565$$

- Part B

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q5_2.jpg")
```

```sas
data child;
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 child2;
set child;
diff = before - after;
run;

proc univariate data = child2;
var diff;
run;
```

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q5_3.jpg")
```

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

```
before <- c(85,70,40,65,80,75,55,20,70)
after <- c(75,50,50,40,20,65,40,25,30)

wilcoxsign_test(before ~ after, distribution = "exact", alternative = "greater")
```

```
##
##  Exact Wilcoxon-Pratt Signed-Rank Test
##
## data:  y by x (pos, neg)
##   stratified by block
## Z = 2.1994, p-value = 0.01562
## alternative hypothesis: true mu is greater than 0
```

- Part C
  - Problem statement: We want to prove the fact from this data set is yoga treatment affects autism children that reduce the time to puzzling
  - $$H_0:$$

    The median difference in yoga treatment between before and after is zero

    $$H_A:$$

The median difference in yoga treatment between before and after is greater than zero

  - Test Statistic: z = 2.1324
  - p-value: 0.01565
  - Conclusion: There is strong evidence that the median difference in yoga treatment between “before” and “after” is greater than 0(normal approximation sign test one-sided p = 0.0313). this mean yoga treatment was effective in reducing the time.

- Part D

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q5_4.jpg")
```

```
data child_p;
input before after @@;
datalines;
85 75 70 50 40 50 65 40 80 20
75 65 55 40 20 25 70 30
;
run;

proc ttest data= child_p alpha= 0.01 side=L;
paired before*after;

run;

```

```r
knitr::include_graphics("C:/Users/choih/OneDrive/Desktop/hw4q5_5.jpg")
```

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	99% CL Mean		Std Dev	99% CL Std Dev	
18.3333	-Infty	39.2367	21.6506	13.0692	52.8140

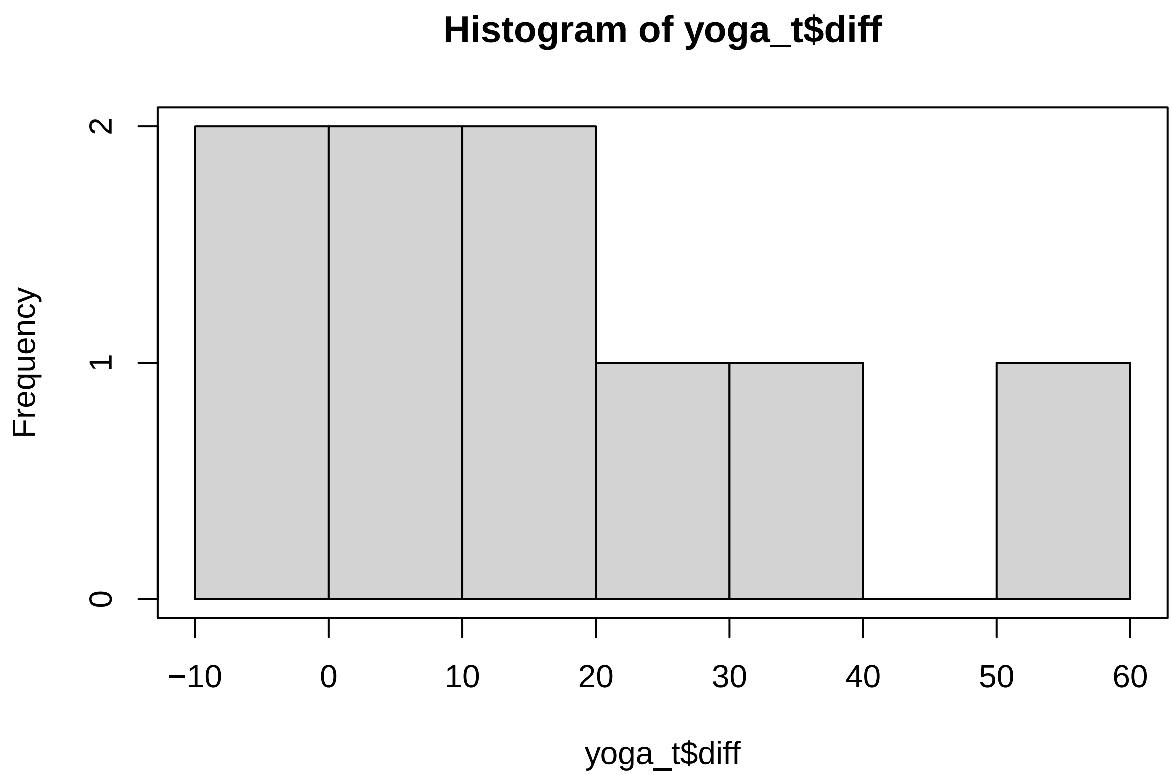
DF	t Value	Pr < t
8	2.54	0.9827

- Part E

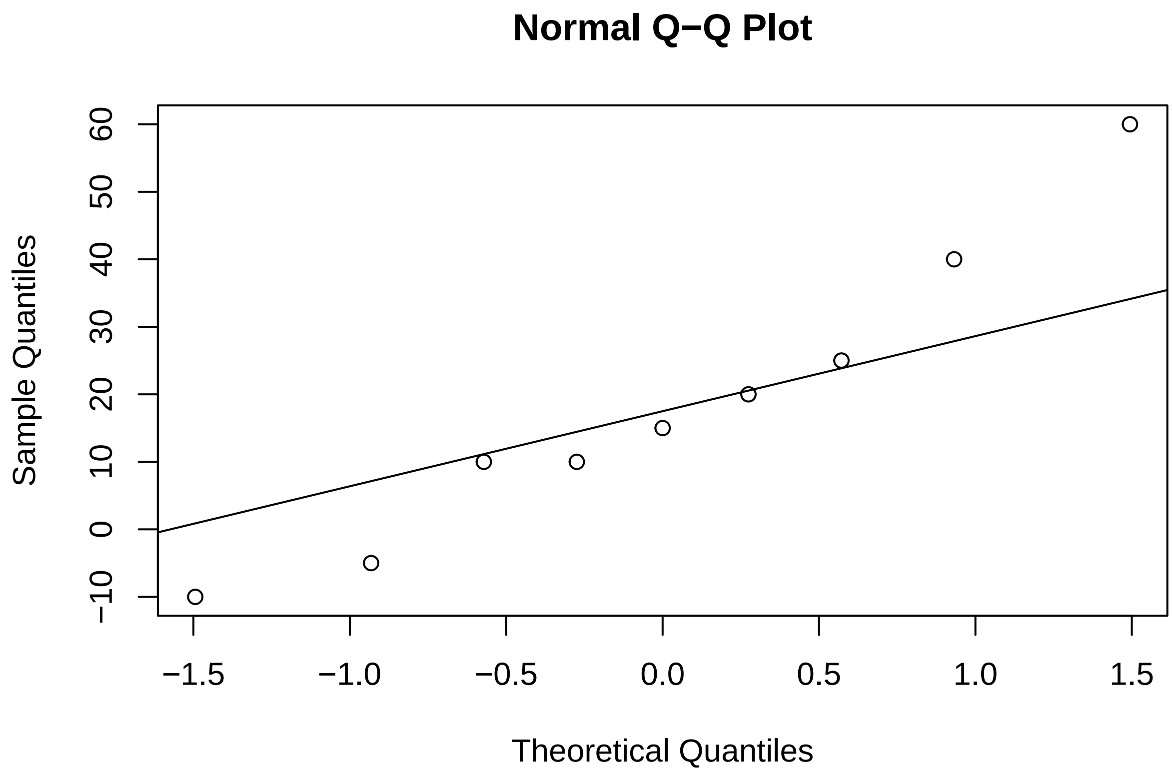
```r
yoga_t <- read.csv("C:/Users/choih/OneDrive/Desktop/yoga_t.csv")
yoga_t$diff <- with(yoga_t, before-after)

t.test(yoga_t$diff, alternative='less')
```

```
##
##  One Sample t-test
##
## data:  yoga_t$diff
## t = 2.5403, df = 8, p-value = 0.9827
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##      -Inf 31.75347
## sample estimates:
## mean of x
##  18.33333
```

```
hist(yoga_t$diff)
box()
```




```
qqnorm(yoga_t$diff)
qqline(yoga_t$diff)
```



- Part F
- I think the sign test is most appropriate for this data. it is one sample and the normality is not good because the sample size is too small for hold to CLT. Also if we see the histogram and Q-Q plot , those distribution symmetric is not looks good. This all reason why I said the Sign test is good for this study