Name: James Tsai
Section: MSDS6371-401
Date: 9/27/15

Question 2.

Problem 20 form the text.

Part a) Determine the rank transformations for the data.

Expenditures	NorT	Order	Rank
18.8	N	1	1
20	N	2	2
20.1	N	3	3
20.9	N	4	4.5
20.9	N	5	4.5
21.4	N	6	6
22	T	7	7
22.7	N	8	8
22.9	N	9	9
23	T	10	10
24.5	T	11	11
25.8	T	12	12
30	T	13	13
37.6	T	14	14
38.5	T	15	15

$n_2 = 8$
$n_1 = 7$

Part b) Calculate the rank-sum statistic by hand (taking the trauma patients to be group 1).
$T_1 = 7+10+11+12+13+14+15 = 82$

Rank-sum statistic = 82

Part c) Mimc the procedures used in Display 4.5 and Display 4.7 to computer the Z-statistic.
Average of the ranks = R = 120/15 = 8
Standard deviation of the ranks = $S_R = 4.472$
Mean $(T_1) = (n_1 R) = 7x8 = 56$
SD $(T_1) = S_R$ x sqrt($n_1 n_2$ / (n_1+n_2)) =
4.472 x sqrt((7x8)/7+8)) = 4.472 x 1.932 = 8.641
Rank-sum statistic with continuity correction = $T_1 - 0.5 = 81.5$

$Z = [T_1 - \text{Mean}(T_1)]/SD(T_1) = (81.5 - 56) / 8.641 = 2.951$

Z-statistic = 2.951

Part d) Find the one-sided p-value as the proportion of a standard normal distribution larger than the observed Z-statistic.

p-value = 0.0016

Question 3.

Problem 21 form the text.

SAS code:

```
DATA METABOLIC_EXP;
INPUT EXPENDITURE PTYPE $;
DATALINES;
20.1    NT
22.9    NT
18.8    NT
20.9    NT
20.9    NT
22.7    NT
21.4    NT
20.0    NT
38.5    T
25.8    T
22.0    T
23.0    T
37.6    T
30.0    T
24.5    T
;

PROC NPAR1WAY DATA=METABOLIC_EXP WILCOXON;
CLASS PTYPE;
VAR EXPENDITURE;
EXACT;
RUN;
```

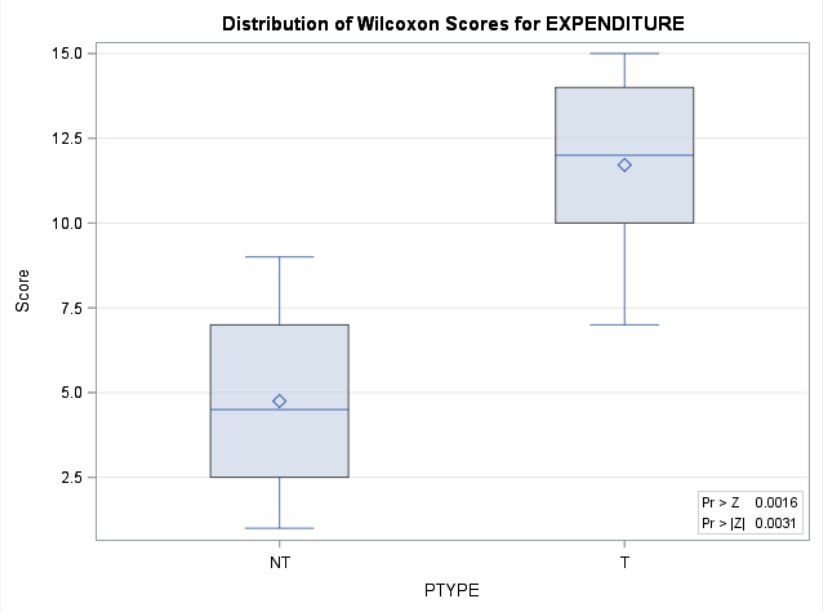
The SAS System

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable EXPENDITURE Classified by Variable PTYPE					
PTYPE	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	
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
Exact Test	
One-Sided Pr >= S	0.0006
Two-Sided Pr >= |S - Mean|	0.0012
Z includes a continuity correction of 0.5.	

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



The statistical package SAS uses a continuity correction. If no continuity correction were used, the Z-statistic would equal 3.0089 and not 2.9537.

Question 4.

Per Professor Sadler, skip this question.

Question 5. Write up a complete analysis using the information you have gained from problems 2, 3, and 4.

The following data are metabolic expenditures for eight patients admitted to a hospital for reasons other than trauma and for seven patients admitted for multiple fractures (trauma).

Metabolic Expenditures (kcal/kg/day)
Nontrauma patients: 20.1 22.9 18.8 20.9 20.9 22.7 21.4 20.0
Trauma patients: 38.5 25.8 22.0 23.0 37.6 30.0 24.5

Part a) State the problem.

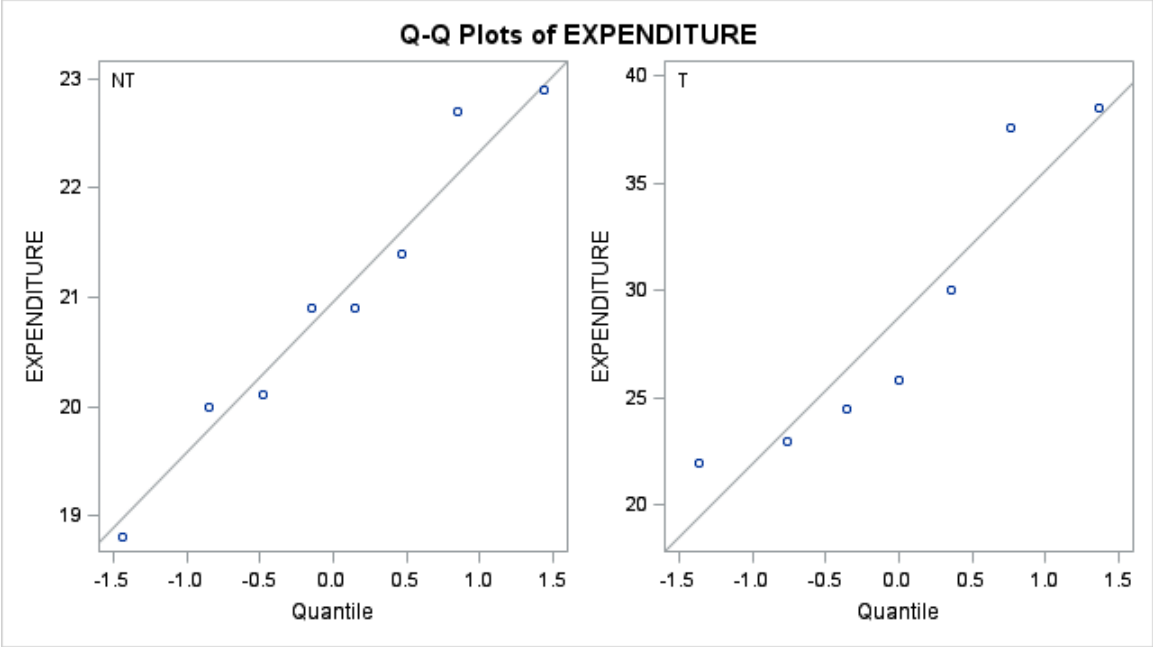
We are testing the claim that trauma patients have higher average metabolic expenditures than non-trauma patients.

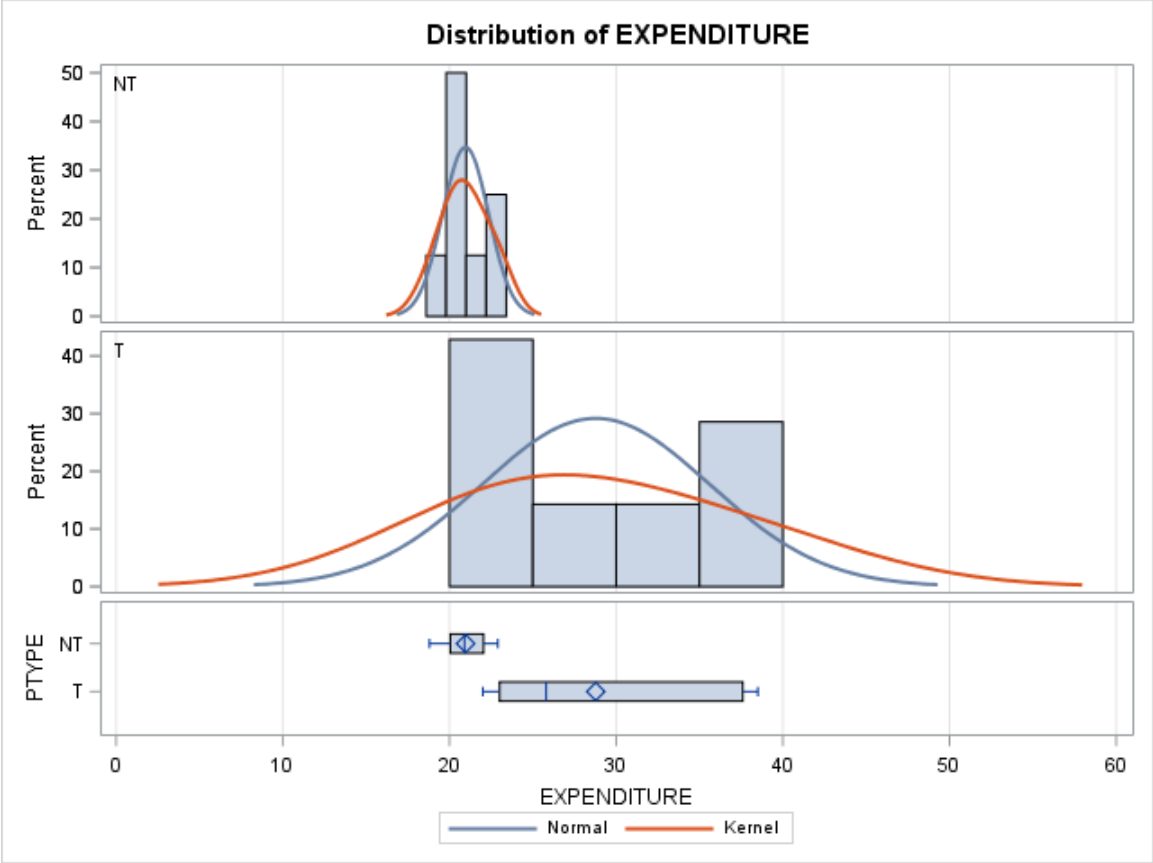
Part b) State the assumptions you are making, why you are making them. Justify your decisions. Print out any histograms, qq plots, box plots, etc. that you use in your justification.

Before we begin any test, we must first examine the characteristics of the data sets to determine the best approach. The first step is to identify if there is any evidence against normality in the two sets of data. After examining the Q-Q plots below, we can state there is no evidence against the normality of the data sets; therefore we can continue with the assumption that the data sets are normal. Next, we examine the variance in the data sets to see if they are equal. From the histogram and the box-plot of the patient types, we can see that the variances are not equal. The trauma patients have a much wider distribution and the variance is much larger than the non-trauma patients. Furthermore, from the boxplot, we can tell that the mean is higher than the median for the trauma patients, and the distribution is right-skewed, whereas the mean and median for the non-trauma patients are approximately equal, making the non-trauma distribution not skewed. Since the variances differ significantly, we rule out the student t-test.

Since we cannot use the student t-test, we can utilize the Wilcoxon rank-sum test, which is a more widely applicable nonparametric test. It is important to note that when we use the Wilcoxon rank-sum test, we are assuming the observations from both groups must be independent from each other. Furthermore, we make no

assumptions about the distribution of the data sets.





Part c) Show all 6 steps of the hypothesis test for the rank sum test of the Trauma data. Use p-values, confidence limits etc. obtained from questions 2, 3 and 4 above.

Step 1: Set up H_0 and H_A.

Since the rank sum is an ordinal test (i.e. a ranking of the data points), we state the null hypothesis in terms of medians.

H_0: $Median_{NT} >= Median_T$
H_A: $Median_{NT} < Median_T$

Step 2: Identify alpha and critical value.

Our hypothesis test is with a significance level (alpha) of 0.05.

From the normal distribution, our critical value for Z(1-alpha) is 1.96.

Step 3: Identify the test Statistic.

Our test statistic (Z-statistic with continuity correction) is 2.951.

Step 4. Find the p-value.

Our p-value is 0.0016.

Step 5. Reject H_0 if the p-value is less than the significance level. Fail to reject if H_0 is not.

Reject H_0 since 0.0016 < .05.

Step 6. Conclusion.

There is strong evidence at alpha = .05 to support the claim that trauma patients have higher median metabolic expenditures than non-trauma patients.

Question 6. Conduct either a two-sample t-test or a Welch's two-sample t-test on the Trauma data used above.

Part a) State the assumptions / reasons you chose the test you did. Be sure and back your answer up with what you know about theory as well as with histograms, box plots, qq plots etc.

Based on the observations from Question 5, part b, we know that there is no evidence against normality for the two data sets. To summarize, we concluded we couldn't use a student t-test due to the significant difference in the variances of the two data sets. Therefore, a two-sample student t-test would not be appropriate since it fails to meet one of the three criteria necessary to run a student t-test. A Welch's two-sample t-test would be appropriate as it is more reliable when the two samples have unequal variances and unequal sample sizes. Therefore, we can justify using the Welch's two sample t-test since it fulfills all 3 criteria:

- Assumptions of normality
- Unequal variances
- Independent or unpaired observations

SAS code:

```
DATA METABOLIC_EXP;
INPUT EXPENDITURE PTYPE $;
DATALINES;
38.5    T
25.8    T
22.0    T
23.0    T
37.6    T
30.0    T
24.5    T
20.1    NT
22.9    NT
18.8    NT
20.9    NT
20.9    NT
22.7    NT
21.4    NT
20.0    NT
;

PROC TTEST DATA=METABOLIC_EXP ORDER=DATA SIDES=1;
CLASS PTYPE;
VAR EXPENDITURE;
RUN;
```

The SAS System

The TTEST Procedure

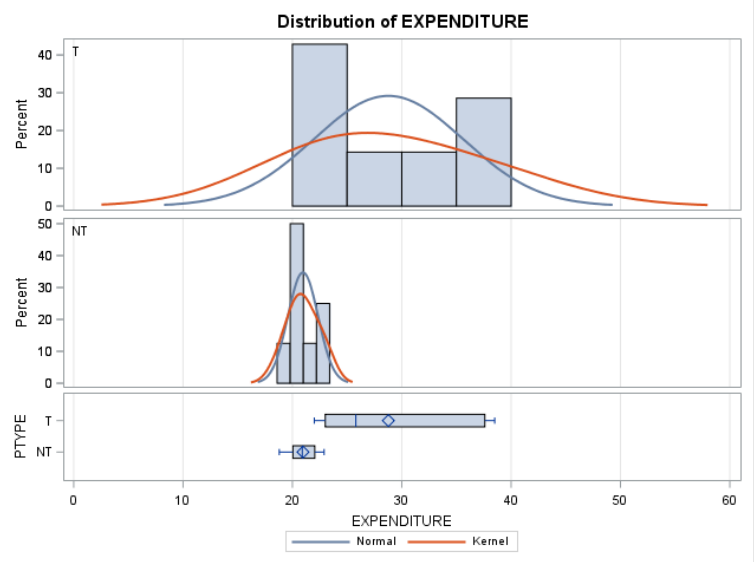
Variable: EXPENDITURE

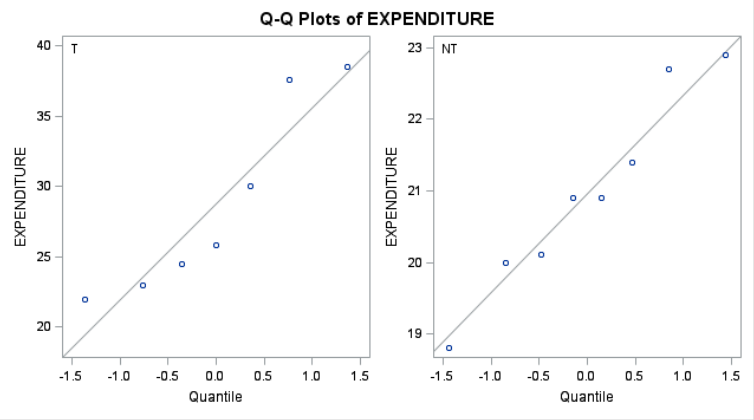
PTYPE	N	Mean	Std Dev	Std Err	Minimum	Maximum
T	7	28.7714	6.8354	2.5835	22.0000	38.5000
NT	8	20.9625	1.3794	0.4877	18.8000	22.9000
Diff (1-2)		7.8089	4.7528	2.4598		

PTYPE	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
T		28.7714	22.4498	35.0931	6.8354	4.4047	15.0520
NT		20.9625	19.8093	22.1157	1.3794	0.9120	2.8074
Diff (1-2)	Pooled	7.8089	3.4528	Infty	4.7528	3.4455	7.6569
Diff (1-2)	Satterthwaite	7.8089	2.7603	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	13	3.17	0.0037
Satterthwaite	Unequal	6.4282	2.97	0.0115

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





Part b) Show all 6 steps (including a thoughtful, thorough yet non-technical conclusion).

Step 1: Set up H_0 and H_A.

H_0: $u_{NT} >= u_T$
H_A: $u_{NT} < u_T$

Step 2: Identify alpha and critical value.

Our hypothesis test is with a significance level (alpha) of 0.05.

DF = 6.4282

Critical value = $t_{6.4282}(.95) = 1.92$.

Step 3: Identify the test Statistic.

Our test statistic is 2.97.

Step 4. Find the p-value.

Our p-value is 0.0115.

Step 5. Reject H_0 if the p-value is less than the significance level. Fail to reject if H_0 is not.

Reject H_0 since 0.0115 < .05.

Step 6. Conclusion.

There is strong evidence at alpha = .05 to support the claim that trauma patients have higher average metabolic expenditures than non-trauma patients.

Bonus.

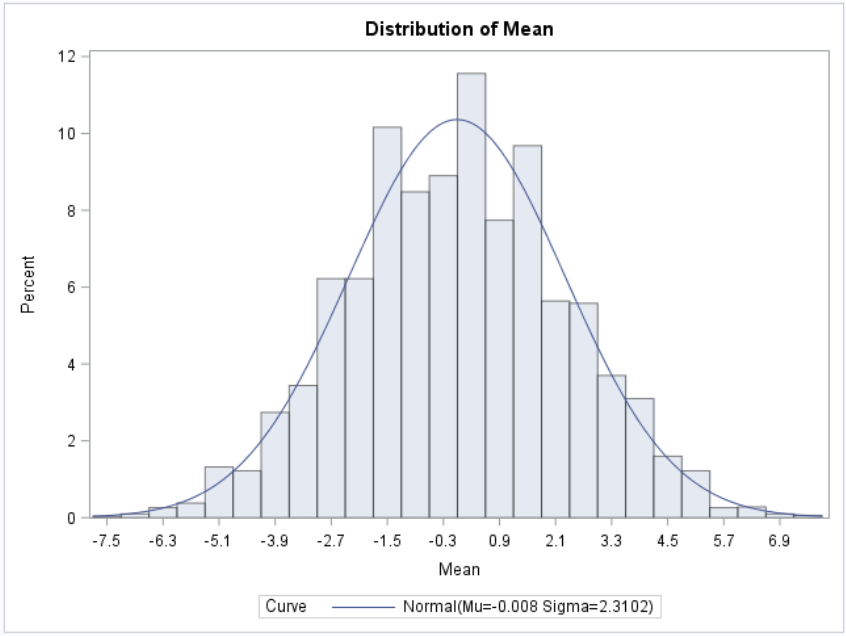
Part a) Build the permutation distribution for the rank sum statistic used in question 2 and 3. Use 5000 permutations. Use SAS to fit / overlay a normal curve to the resulting histogram. Compare the mean and standard deviation of this normal curve that was fit to the permutation / randomization distribution to the mu and sigma you found in question 2.

SAS code:

```
DATA ME;
INPUT PTYPE EXPENDITURE;
DATALINES;
0       1
0       2
0       3
0       4.5
0       4.5
0       6
1       7
0       8
0       9
1       10
1       11
1       12
1       13
1       14
1       15
;

proc iml;
use ME;
read all var{PTYPE EXPENDITURE} into x;
p = t(ranperm(x[,2],5000));
paf = x[,1]||p;
create newds from paf;
append from paf;
quit;

ods output conflimits=diff;
proc ttest data=newds plots=none;
class col1;
var col2 - col5001;
run;

proc univariate data=diff;
where method = "Satterthwaite";
var mean;
histogram mean / normal (mu=est sigma=est color=red);
run;
```

The SAS System

The UNIVARIATE Procedure



The SAS System

The UNIVARIATE Procedure
Fitted Normal Distribution for Mean

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	-0.00806
Std Dev	Sigma	2.310168

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.02079297	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.23041954	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	1.34947587	Pr > A-Sq	<0.005

Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	-5.22321	-5.38232
5.0	-3.75000	-3.80795
10.0	-2.94643	-2.96866
25.0	-1.60714	-1.56625
50.0	0.00000	-0.00806
75.0	1.60714	1.55012
90.0	3.08036	2.95254
95.0	3.88393	3.79183
99.0	5.15625	5.36619

The mean mu that was obtained from this permutation test is -0.00806 and standard deviation sigma is 2.3101. This tells us with a 95% confidence level under the null hypothesis that the difference of the mu between the two groups is between -4.64 to 4.60 (2 standard deviations). From question 2, the theoretical "null hypothesis" mean and standard deviation are 56 and 8.641 for group T_1, and 64 and 8.641 for group T_2. If we take the difference of means, we obtain a value of 8 which is outside the our 95% confidence level, therefore we reject the H_0.

Part b) Compare the one sided p-value found in this permutation distribution with the one found in question 2.

The one-sided p-value found in question 2 is 0.0016. The one-sided p-value found in this permutation distribution are <0.010 or <0.005 depending on the goodness-of-fit test used. While the p-values differ slightly, they are consistent with the rejection of our null hypothesis since our critical value is 0.005.