Sean Kennedy Statistical Foundations for Data Science: Homework 4

**Question 2:**

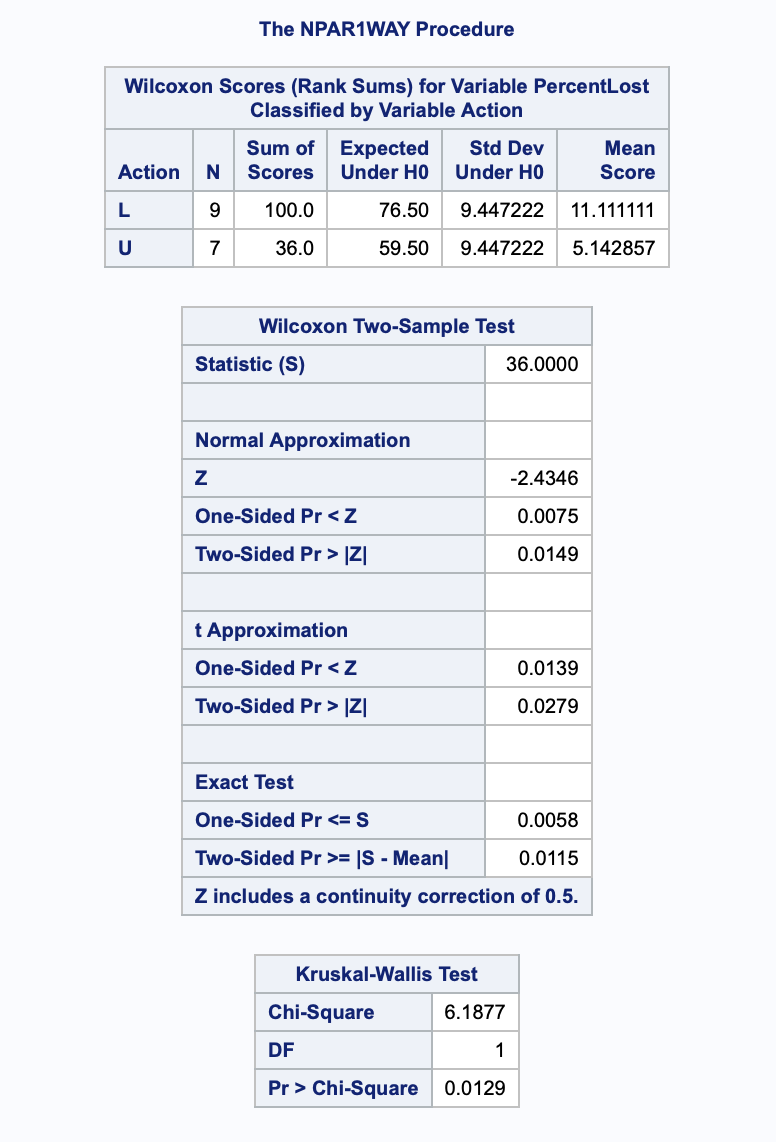
1. 6 Step Analysis on Logging.csv using Wilcoxon Ranksum Test in SAS:

* **Identify H0 and Ha:**

H0: logged distribution = unlogged distribution (logging has no effect)

Ha: logged distribution > unlogged distribution (logging has negative effect)

* **Critical Values/Draw Shade:**

****

* **Critical Value** = -1.753 (left sided alpha=0.05)
* **Test Statistic** = min of group ranksums = 36

Z-score = -2.436

* **P-value** = 0.0075 (using 1-sided normal approximation)

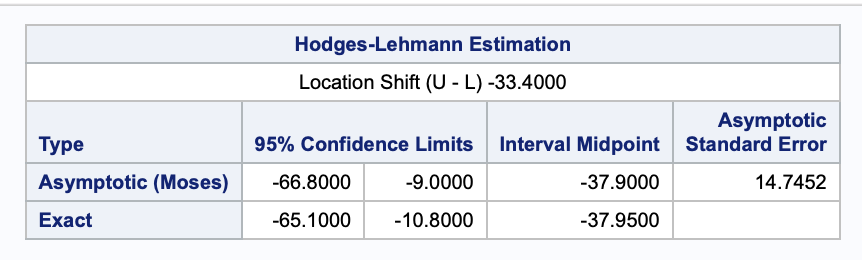
Reject H0

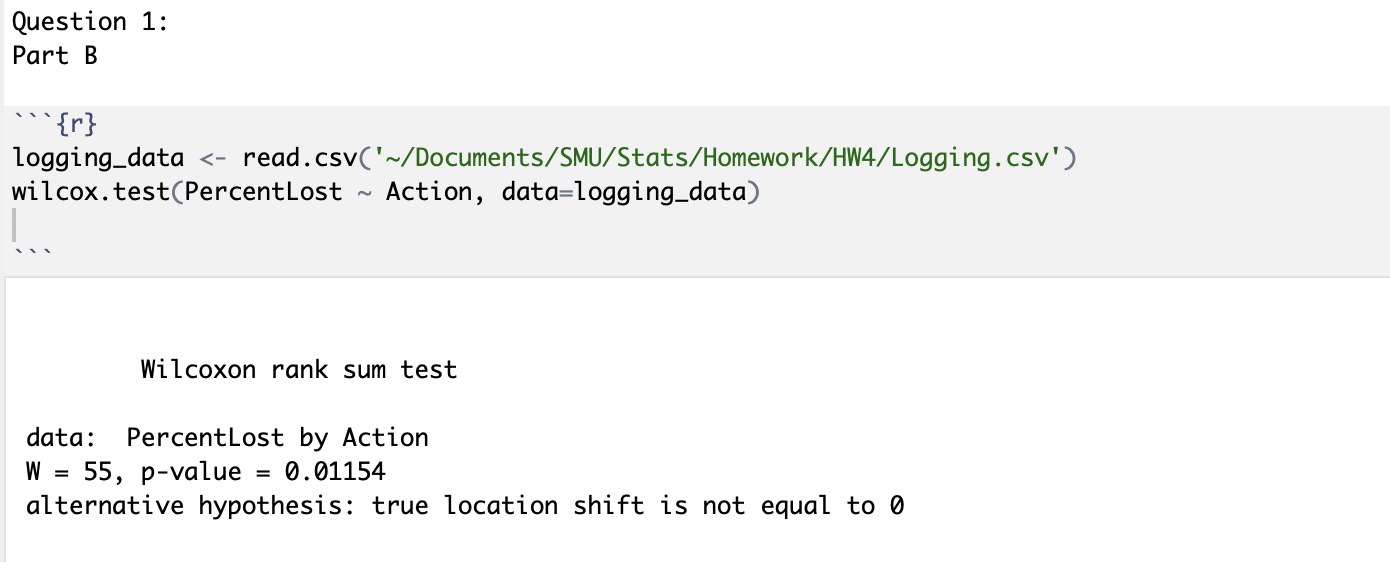
* **Conclusions/Scope**

Low p-value causes us to reject the null hypothesis that the logged and unlogged distributions are equal. We can assume a normal distribution to be appropriate give the close results given by the exact test (p-value 0.0058). We find strong evidence for the alternative hypothesis (logged distribution shows greater percent loss).

Given the 95% CL given by the HL estimation – the point estimate of percent loss due to logging is -37.95% with an interval of [-65.1%, -10.8%].

Due to the nature of this observational study – direct causal inference cannot be made as to the effect of logging on seedling loss – but there is strong evidence that a link exists between logging and the number of seedlings observed in the next year.

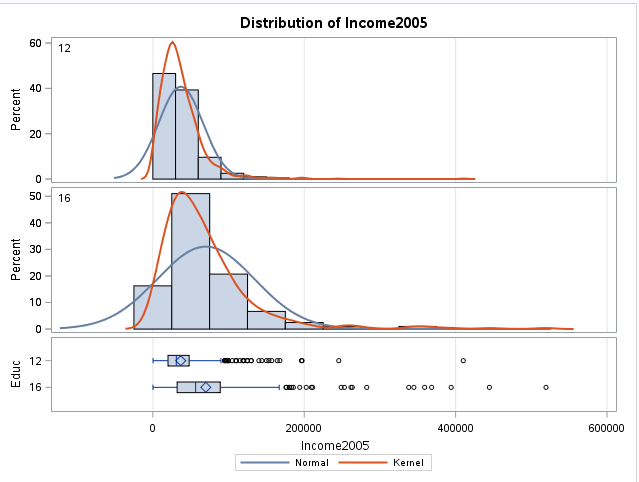


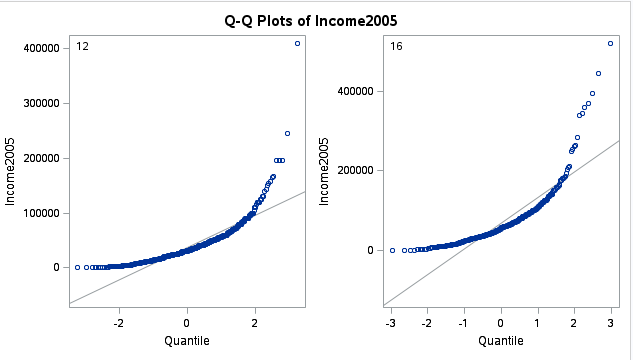


**Question 3:**

**Problem:** We wish to test the claim that adding an additional 4 years of education has a profound impact on the salary distribution of employees. To do so – 2 random(?) samples of data from NLSY79 were selected. All subjects were between 41 and 49 years old, had paying jobs in 2005 and had either 12 or 14 years of education at the time of the study. Their incomes for 2005 were recorded.

1. Assess validity of standard t-test:





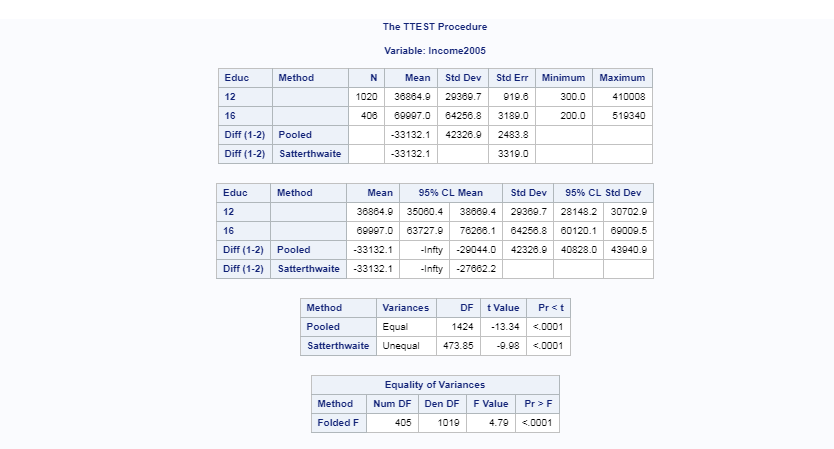
Both samples appear to be normal and have large samples sizes. Inspection of histograms, box plots and Q-Q plots indicate a large difference in standard deviation between the two samples. Since we cannot assume the same standard deviation, but because the sample sizes are large enough to invoke the CLT – we will conduct a one sided welch’s t-test on the data to test our hypothesis.

1. 6 Step Analysis on Education data using Welch’s T-Test Approximation in SAS:

* **Identify H0 and Ha:**

H0: Mean(college) - Mean(high school)=0 (education has no effect)

Ha: Mean(high school) < Mean(college) (education has positive effect)



* **Critical Values:**

At alpha = .05 (95% CL) a one-sided t-statistic df = 473.85

Critical t = -1.648

T-statistic = -9.98

* **P-Value:** <.0001(1 sided)

Reject H0

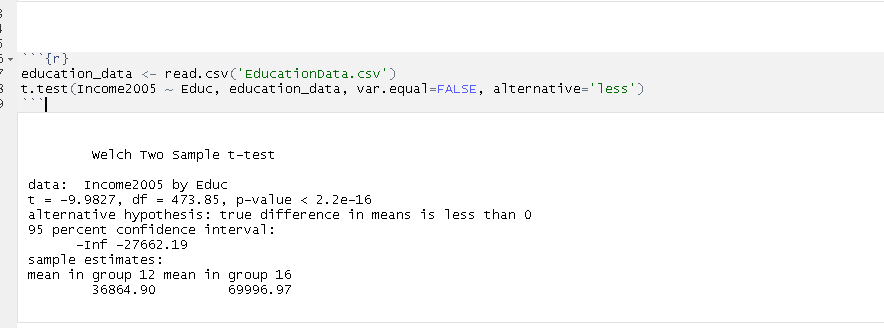
1. **Conclusions/Scope:**

It is estimated that the average difference in income of the 12 year group was -33,132 dollars less than that of the group with 16 years of education (one-sided p-value = <.0001) at a 95% confidence interval of [-inf, -27,662]. Due to the extremely low p-value, there is very strong evidence against the null hypothesis (equal means). Since randomization was used to select the sample populations, it is safe to assume that amongst the population of individuals age 41-49 – there is sufficient evidence to suggest that the increased education led to higher income earned. Because this was an observational study, direct causal inference cannot be made. Though the sample size is very large – ideally large enough to account for confounding variables – there are other features that should be added to this set to enhance the analysis:

* Starting salary
* Race
* State
* Graduate vs non-Graduates,
* Years on job
* Currently unemployed are excluded

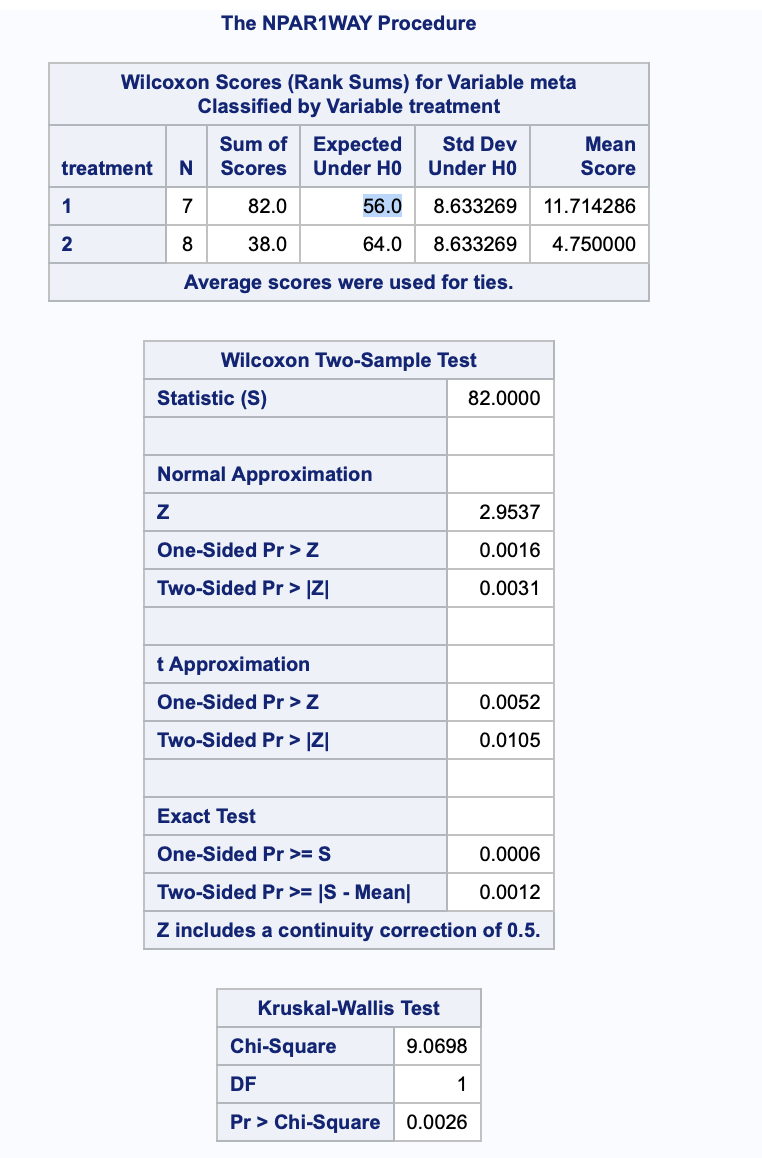
Of these – state, race and starting salary are the most important. Sample size should be sufficient to account for variations in gender.

1. **R-Code:**



1. Comparison to my methodology in HW 3 is somewhat difficult given that I made an inference about the median in HW 3 and am making an inference about the mean in this assignment but given the large amount of skew in both population examples and the fact that there are no zeroes in the dataset makes using the log transformed data appropriate. Using the log transformation to make inferences about the median also seems more appropriate given that the mean is sensitive to outliers whereas the median is not.

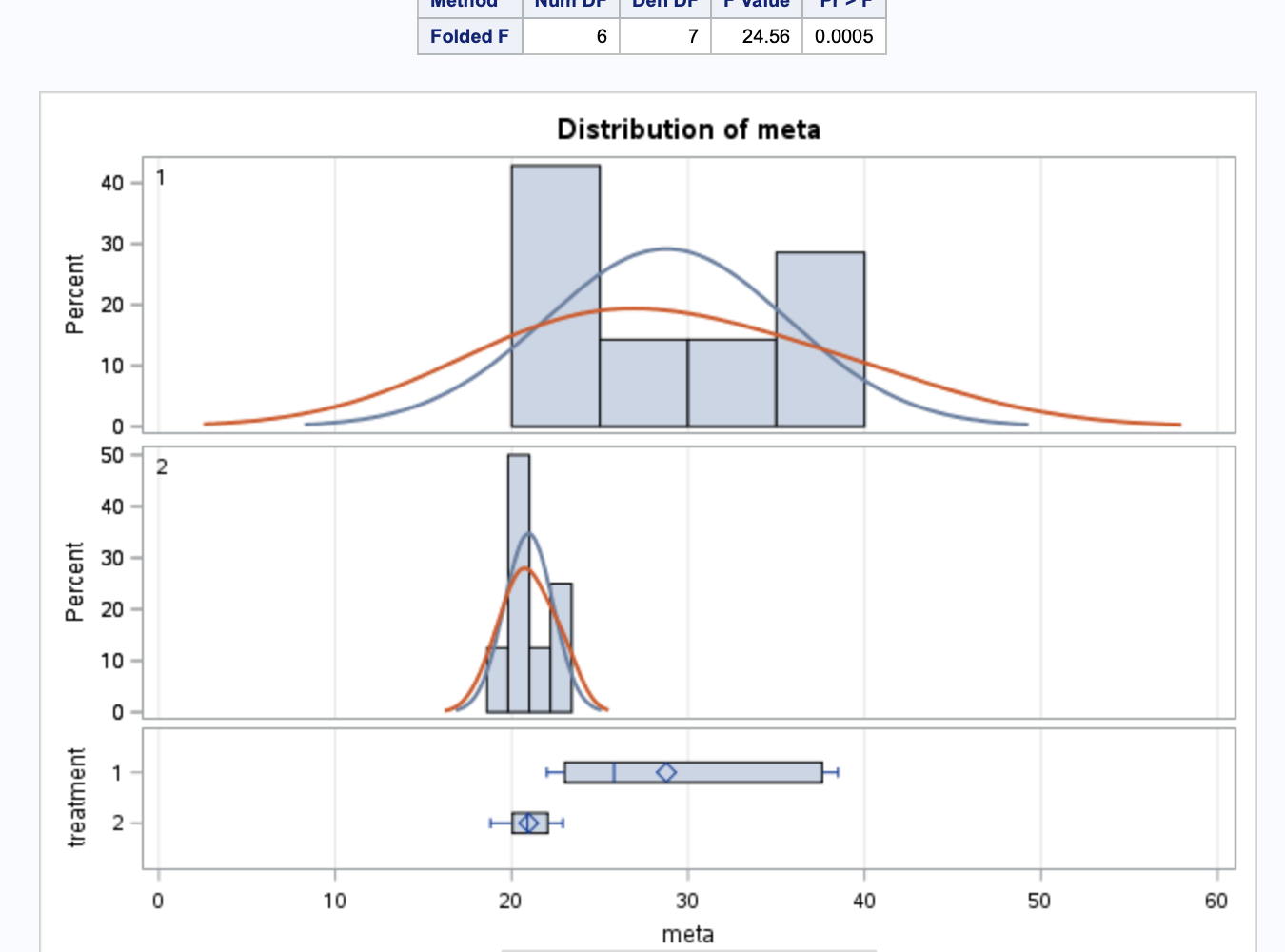
**Question 4:**

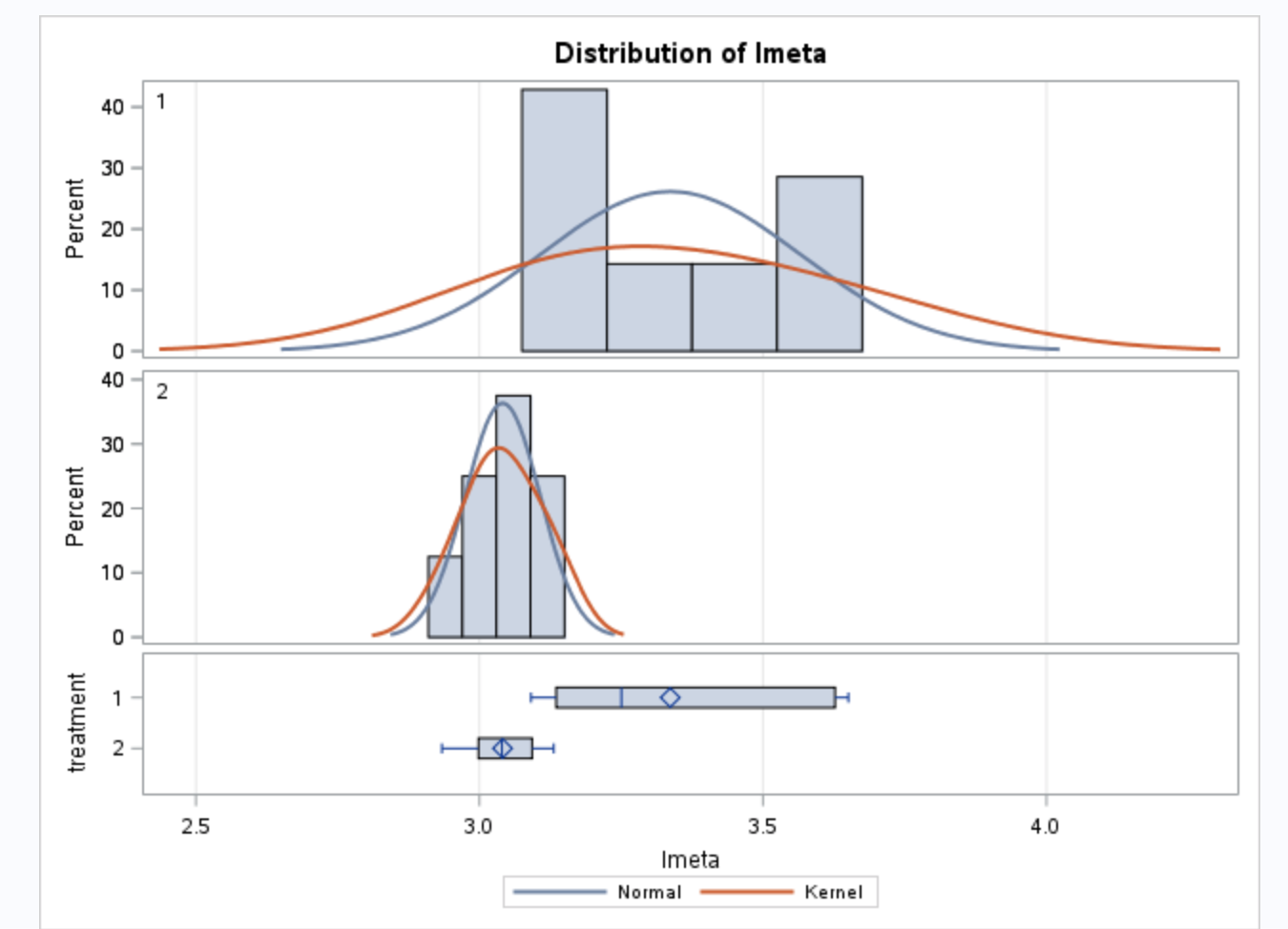
1. **Problem 20:**
2. ****
3. **State the Problem:**

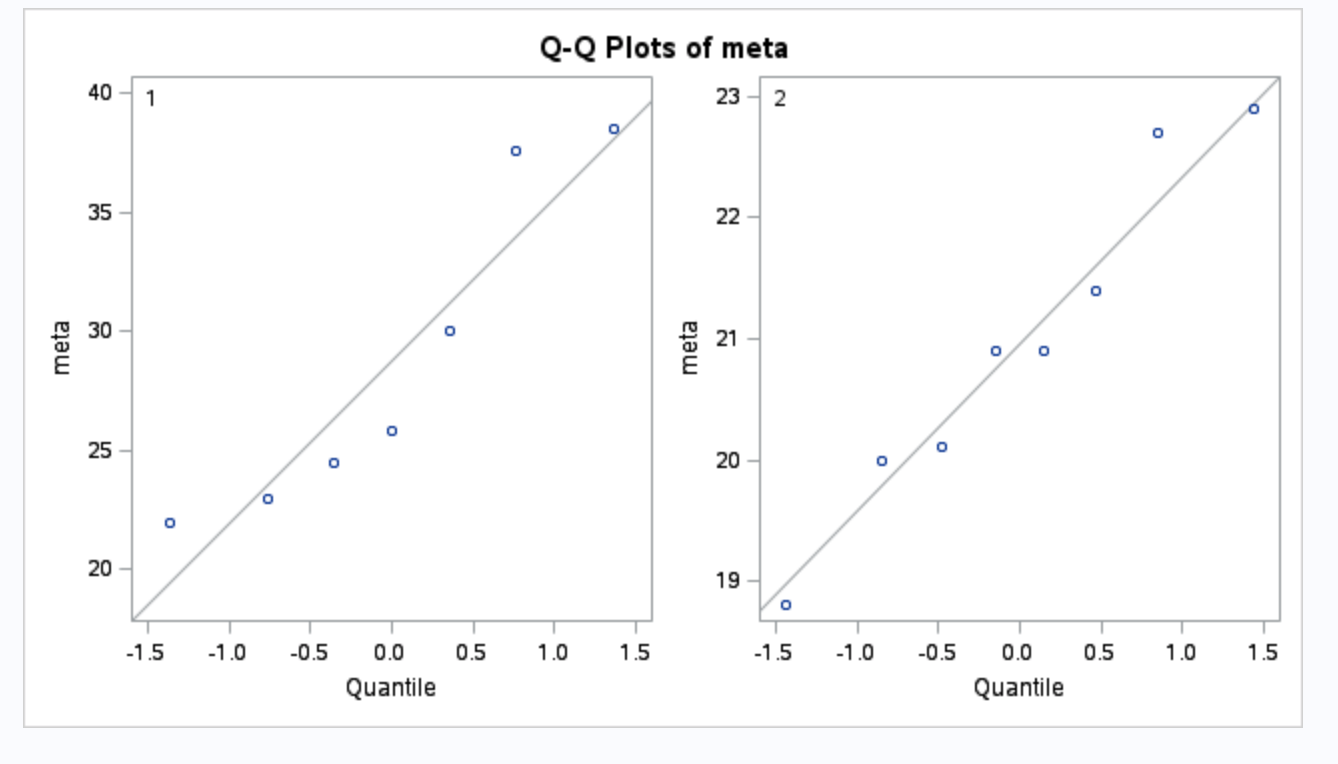
A hospital wishes to measure the effect of trauma (multiple broken bones) on the metabolic rate (measured in kcal burned per kg of body weight per day). The assumption is that the metabolic rate will be higher due to the body burning extra energy to recover from the additional trauma.

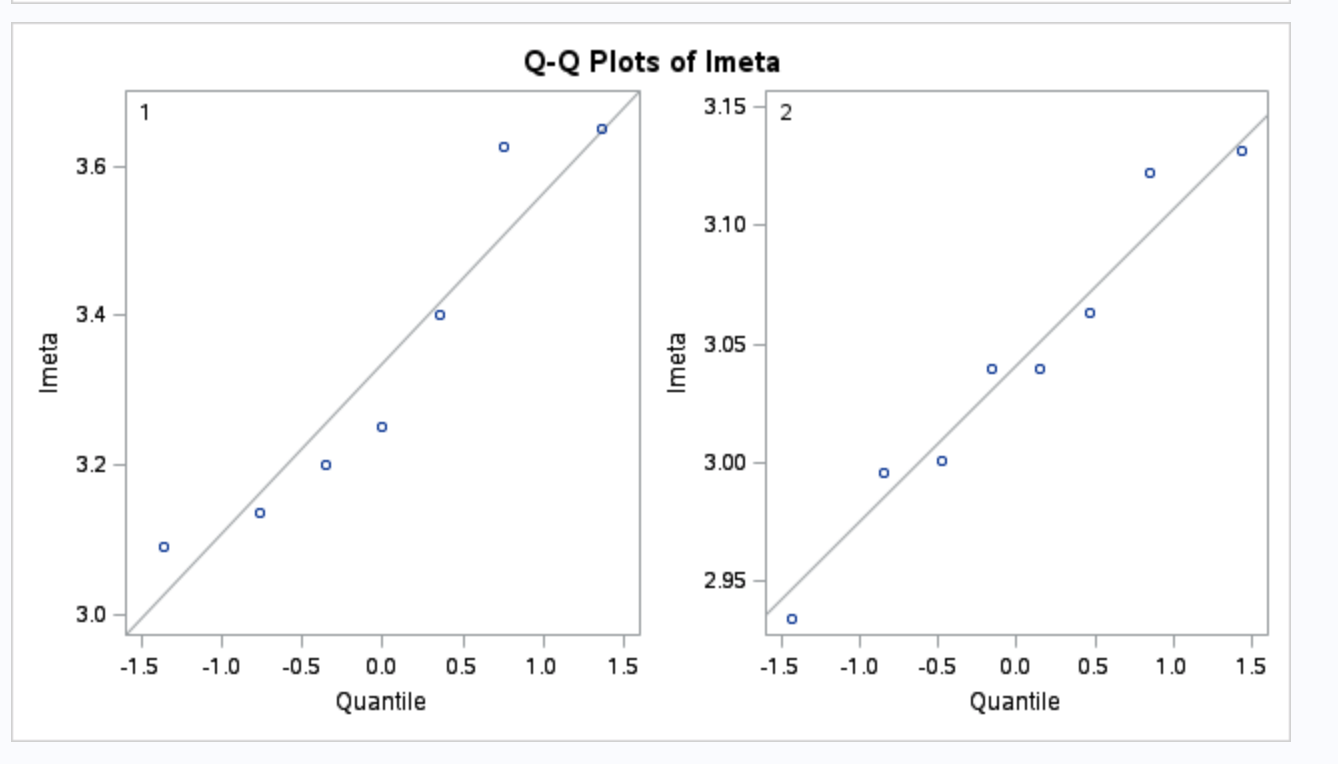
**Assess Validity of Standard T-Test:**

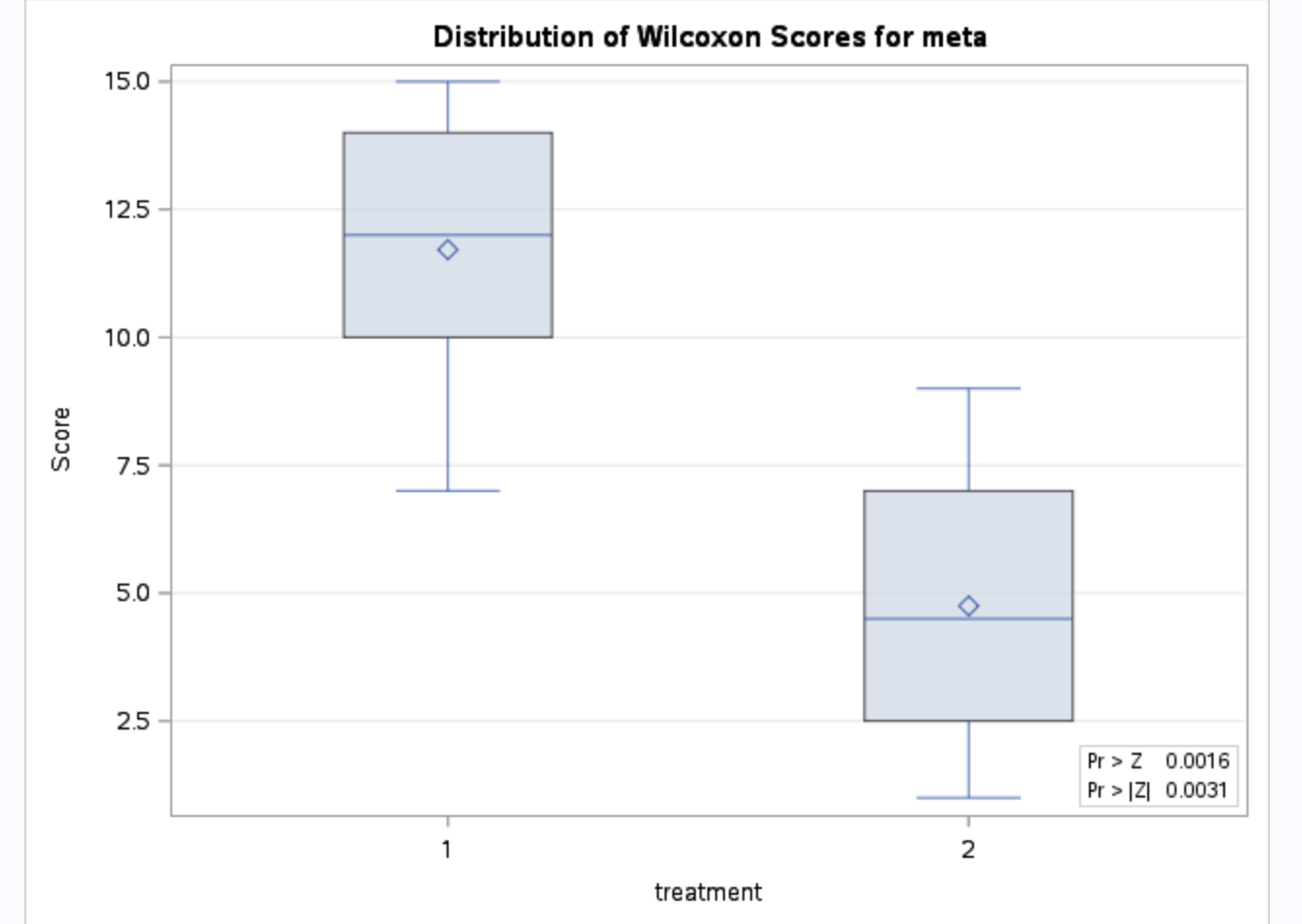
Visual inspection of histograms and box plots for both the raw and log transformed data indicate high levels of skew and unequal variances. In order to circumvent these limitations, we will employ a Wilcoxon Ranksum test as it is a non-parametric test that assumes no underlying distribution for the raw data but relies on the normality of ranks associated with the ordering of the raw data. The ranksum test is valid when sample sizes are greater than 5 and when there are relatively few ties – both criteria are met in this scenario.











6 Step Analysis on treatment data:

* **Identify H0 and Ha:**

H0: Distribution of trauma vs non-trauma patients are equal

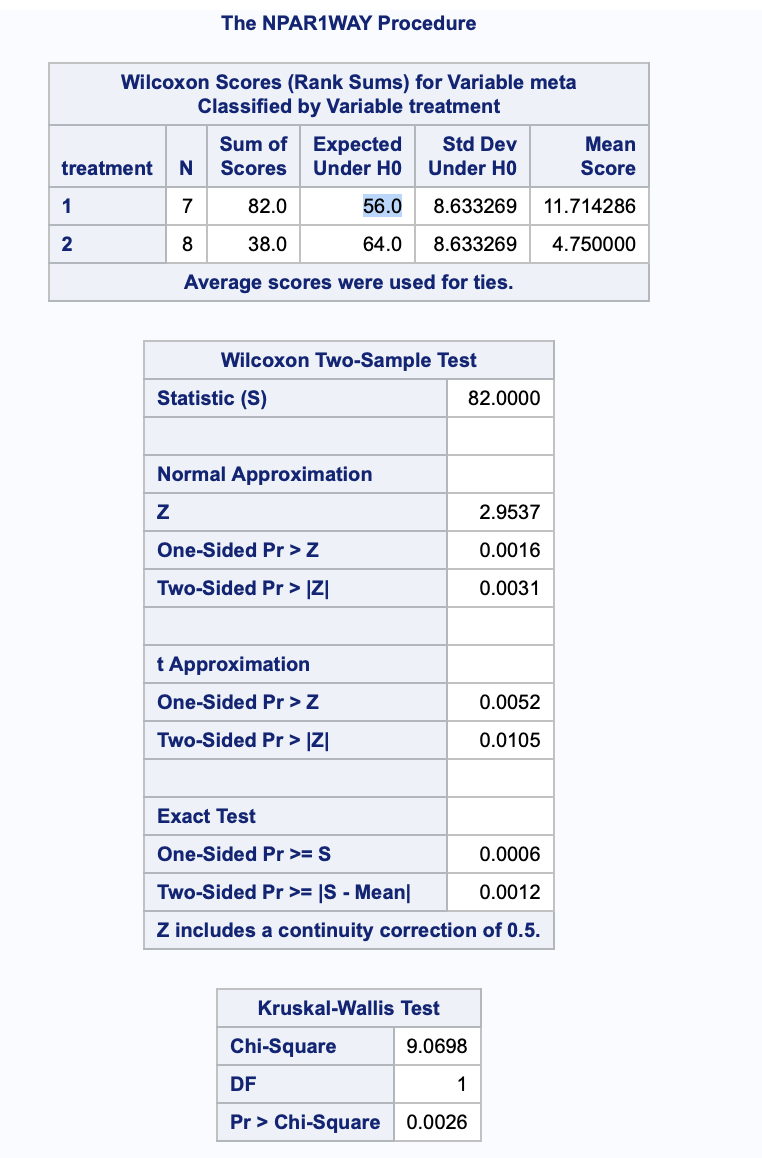
Ha: Distribution of trauma is greater than distribution of non-trauma

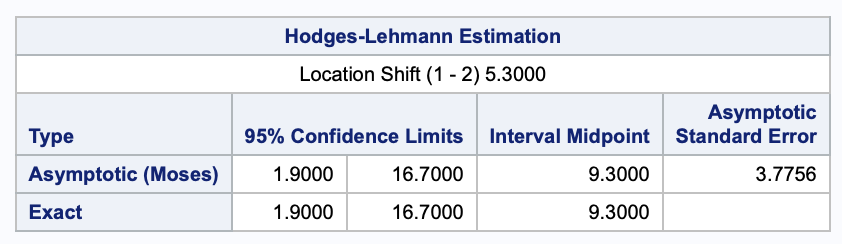
* **Critical Values:**

At alpha = .01 (95% CL) a one-sided t-statistic df = 15

Critical t = 1.753

T-statistic = 2.9537

****



* **P-Value:** 0.0016 (1 sided from normal) (0.0006 exact value)

Reject H0

* **Conclusions/Scope:**

In this observational study there is clear evidence to support the assumption that the distribution of metabolic rates is higher in the sample of trauma patients. The low p-value (0.0016) obtained in the Wilcoxon ranksum test causes us to reject the null hypothesis of equal distributions and strongly consider the alternative.

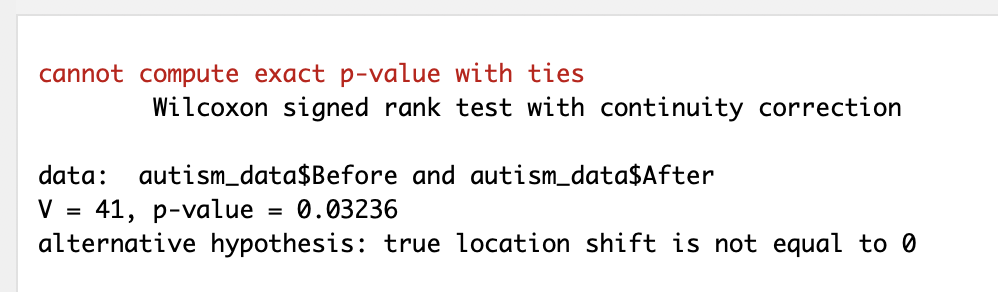
A point estimate of 9.3 kcal/kg/day was observed in the study with a 95% CL of [1.9, 16.7] kcal/kg/day.

In this study – it is assumed that the relationship between kcal expended per kg of body mass to be relatively constant since it is used as a means of comparison between individuals. This is an assumption that should be investigated further.

**Question 5:**

**a)**

**b)**

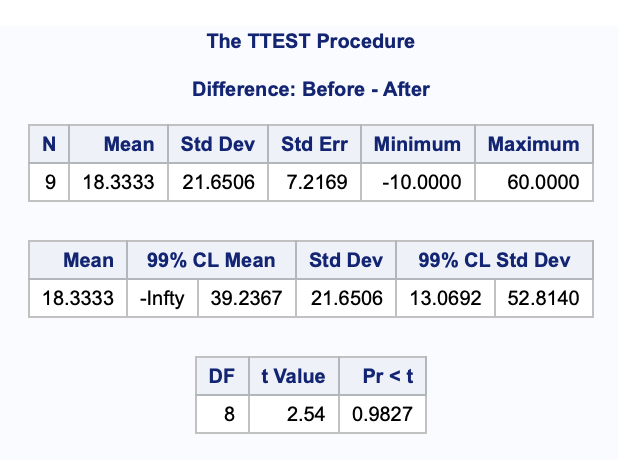
****

**c) 6 Step Hypothesis Test:**

* **Identify H0 and Ha:**

H0: Mean of differences is zero

Ha: Mean of before is larger than after (treatment has positive effect on outcome, less time to complete the task)



* **Critical Values:**

At alpha = .01 (99% CL) a one-sided t-statistic df = 9

Critical t = 1.833

T-statistic = 2.132

* **P-Value:**

Exact from SAS p= 0.0313

Reject H0 (mean difference is non-zero)