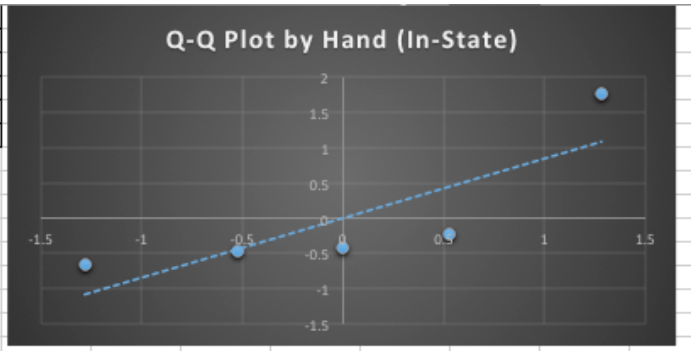
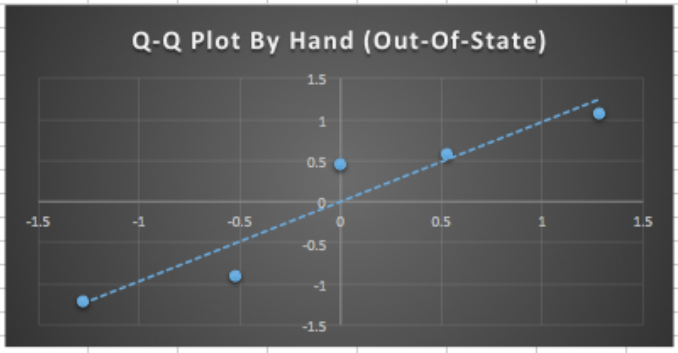
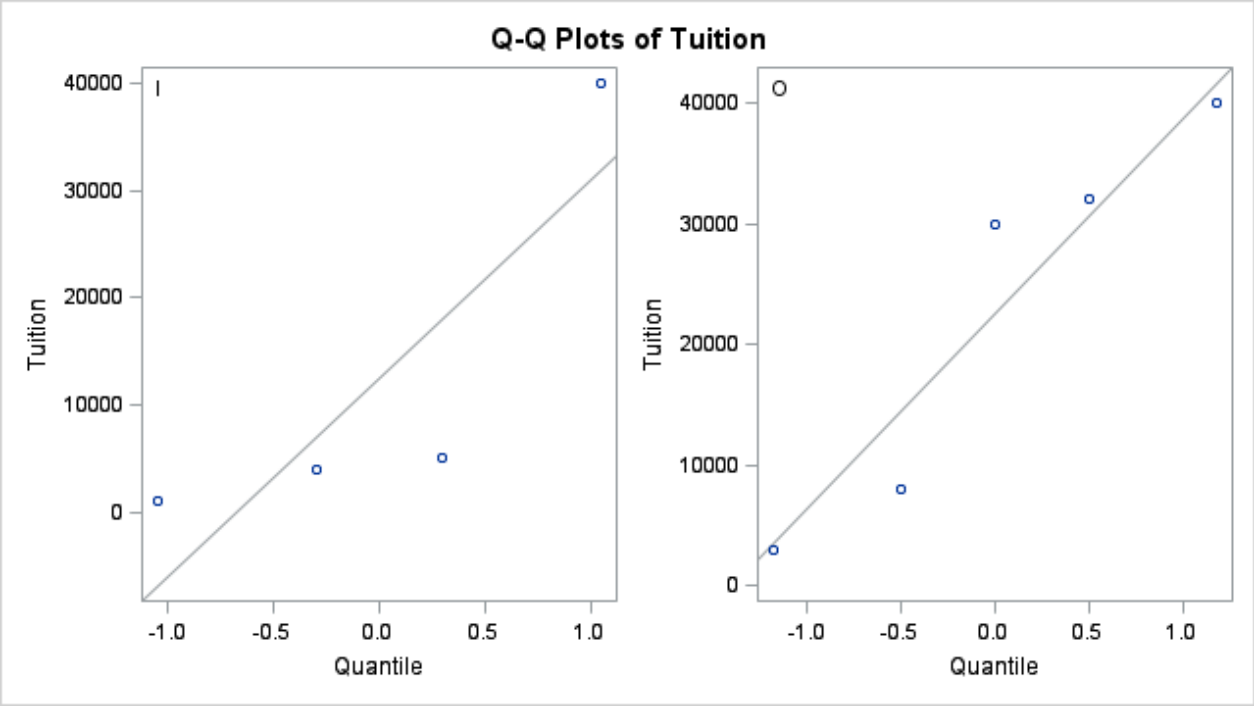
Question 1.

data	z-data	rank	middle	z-normal
1000	-0.65954383	1	0.1	-1.281551566
4000	-0.472880482	2	0.3	-0.524400513
5000	-0.410659366	3	0.5	0
8000	-0.223996018	4	0.7	0.524400513
40000	1.767079697	5	0.9	1.281551566
11600	average			
258300000	variance			
16071.71428	sd			



data	z-data	rank	middle	z-normal
3000	-1.213674641	1	0.1	-1.281551566
8000	-0.904063763	2	0.3	-0.524400513
30000	0.458224099	3	0.5	0
32000	0.58206845	4	0.7	0.524400513
40000	1.077445855	5	0.9	1.281551566
22600	average			
260800000	variance			
16149.30339	sd			





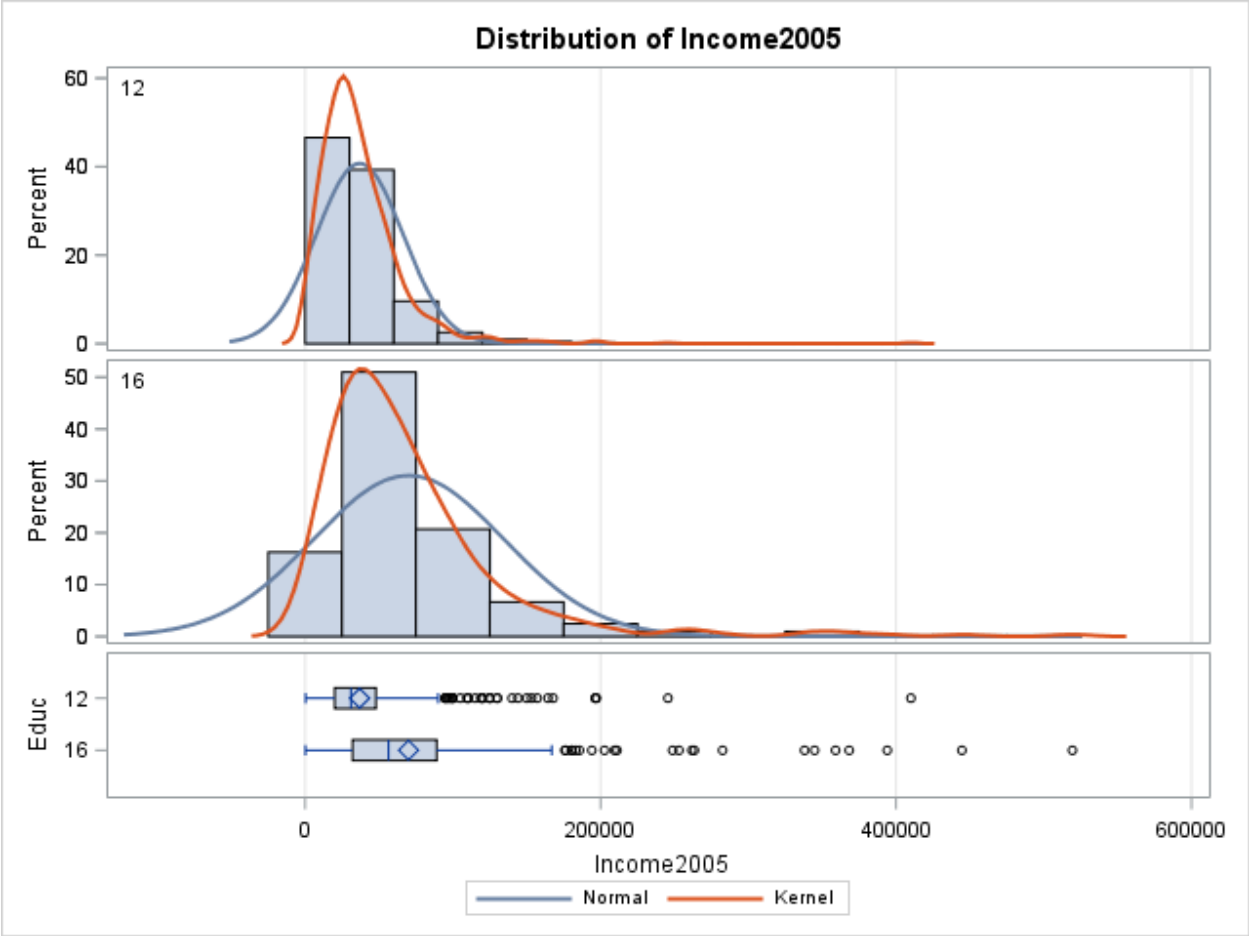
Our assumptions are H_0: population distribution is normal and H_a: population distribution is non-normal.

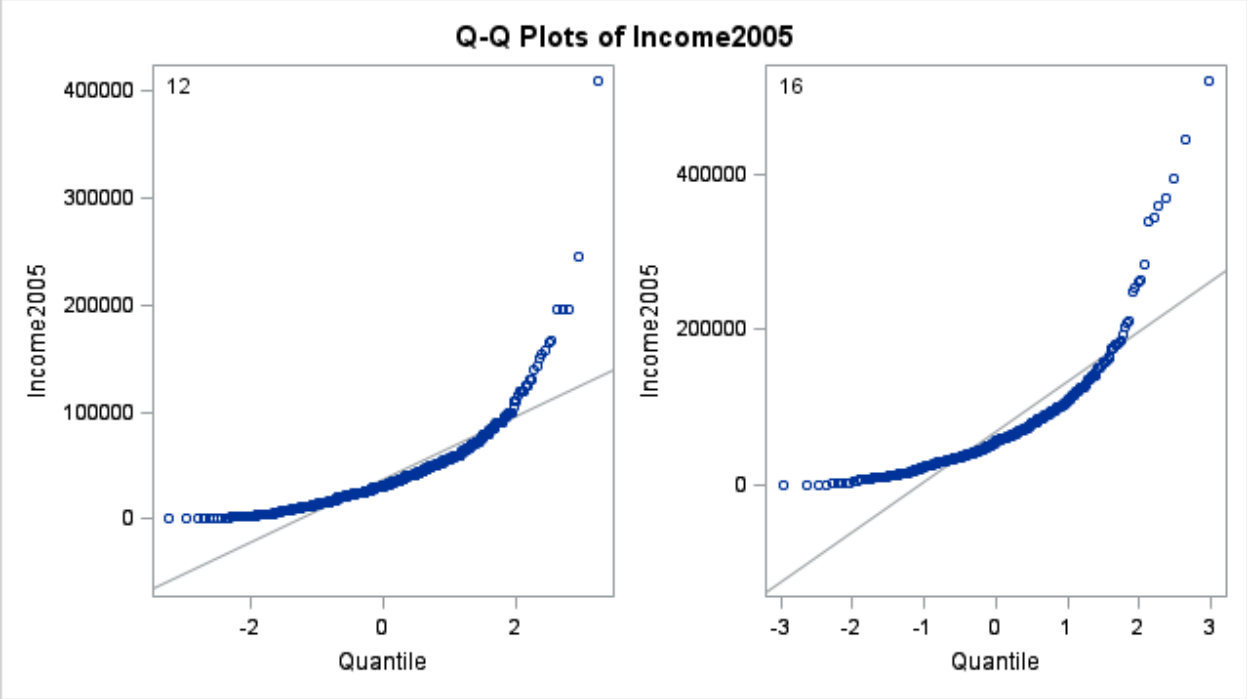
For the In-State data, there is a clear evidence against normality. There are 4 out of 5 values which are negative for the z-data. We would expect the first 2 values to be negative and the next 3 values to be positive for a distribution that is closer to a normal distribution. Furthermore, we need to investigate the outlier where the value is 40,000. Reject H_0.

For the Out-Of-State data, there is no evidence against normality. The first 2 values are negative, and the next 3 values are positive, which is a closer approximation to what we would expect in a normal distribution. Furthermore, there are no obvious outliers in the data. Fail to reject H_0 and continue to assume normality.

Question 2.

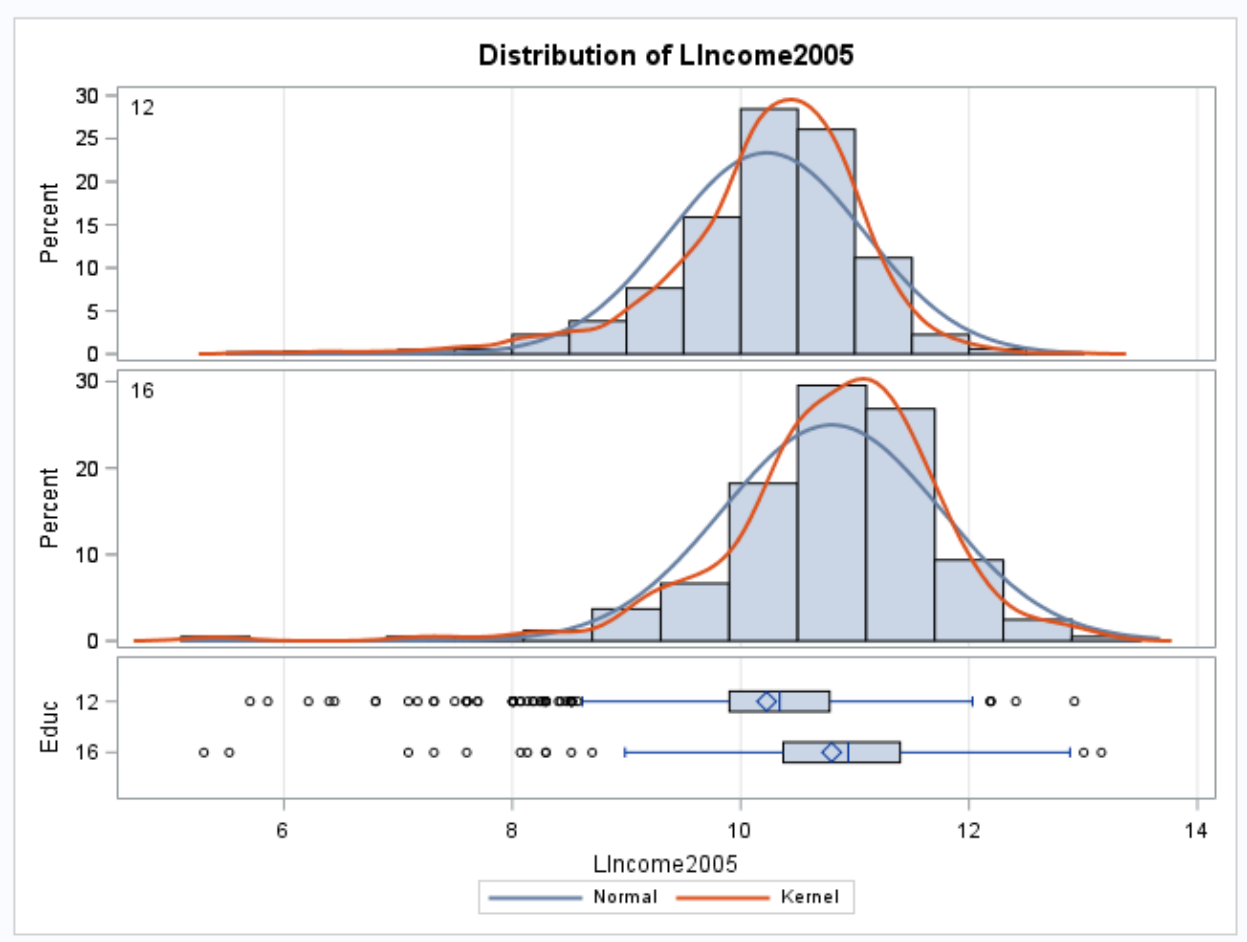
1. We are testing the claim that the distribution of incomes for those with 16 years of education exceeds the distribution for those with 12 years of education.
2. In order to perform a proper test, we must first address some fundamental assumptions of the data sets. Specifically, we need to test for normality, equal SD, and independence.

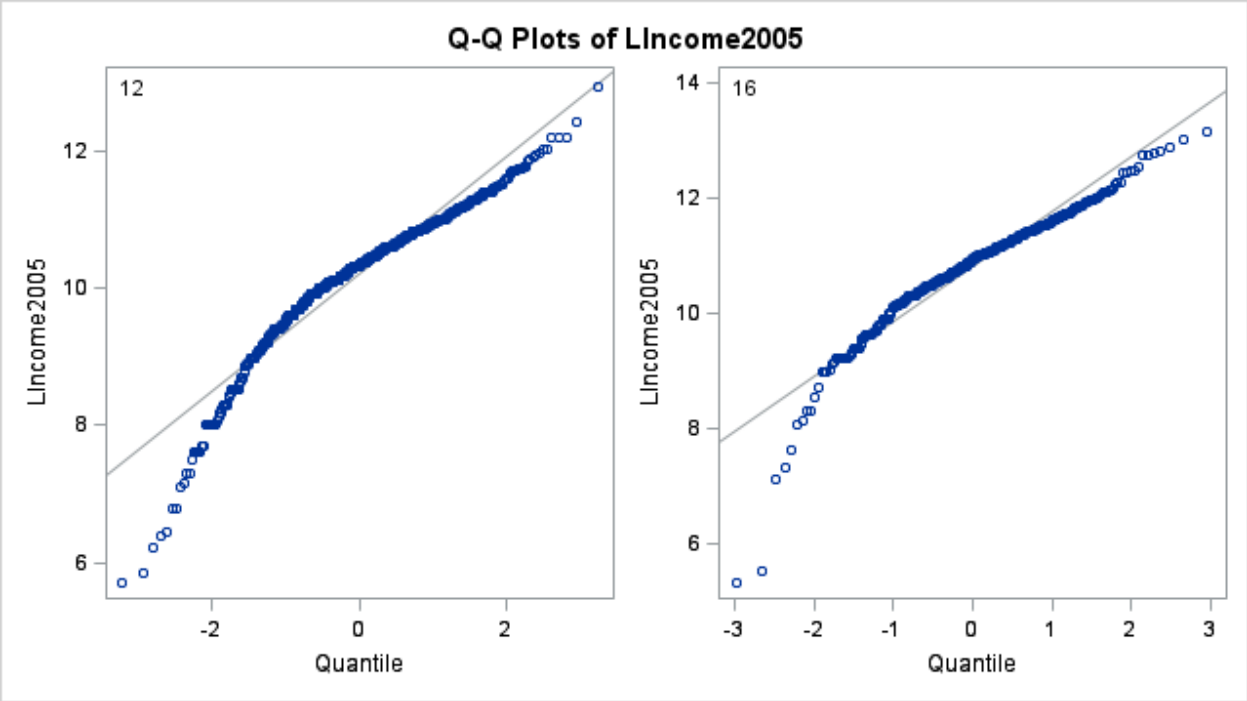

Distribution of Income2005
12
16
Percent
Educ
Income2005
Normal
Kernel


Q-Q Plots of Income2005
12
16
Income2005
Quantile

Original Data (see above histogram and Q-Q plot)

Normality – Each of the two samples in these groups should follow a normal distribution. Presence of outliers, for example, would increase the sample variance and lead to inaccurate t-values. Visually, the Q-Q plot of the original education data indicates that it does not follow a normal distribution. Furthermore, the ratio between the largest and smallest incomes in each group is extremely large – 410,008/300 = 1366.7 and 519,340/200 = 2596.7, for the 12 years of education and 16 years of education, respectively. The mean values in this case will not be meaningful as the data is skewed and the distribution has very long tails.





Log Transformed Data (see above histogram and Q-Q Plot)

Normality – After applying the natural log to the income values, we have a Q-Q plot that closely resembles the normal distribution. Even if there are some concerns of normality, the sample is large enough to ensure that the t-test is robust to this assumption. Also there is no evidence against the equality of the population standard deviations. The log transformed data is approximately normally distributed, and the anti-log of the mean on the logged scale will be approximately equal to the median of the untransformed data.

3. We will perform the 6-step hypothesis to gain inference on the median since we are dealing with the log of the data. For the 6-Step Hypothesis:

   A) Set up H_0 and H_1

   We are conducting a one-tail test to the claim that those with 16 years of education have a higher median income than those with 12 years of education. Therefore, we will construct our null hypothesis to assume u_{12} is greater than or equal to u_{16}.

   H0: $u_{12} >= u_{16}$

   HA: $u_{12} < u_{16}$

   B) Identify alpha and the critical value

   Our hypothesis test is with a significance level (alpha) level of .05.

DF = (1020 + 406) – 2 = 1424

Critical value = $t_{1424}(.05) = -1.655$

C) Identify the test Statistic

$x_{12} = 10.2272$, $x_{16} = 10.7971$

$s_{12} = 0.854$, $s_{16} = 0.9581$

$n_{12} = 1020$, $n_{16} = 406$

t = (10.2272-10.7971)/sqrt((0.854^2/1020) + (0.9581^2/406))

t = -10.45

D) Find P-Value

p =< .0001

E) Reject H_0 if the P-value is less than the significance level (alpha). Fail to reject if H_0 if it is not.

Reject H_0 since .0001 < .05

F) Conclusion

Since we have ran the t-test on the log of the data, we must take the anti-log of the mean difference to find out the degree in which the median differs on the original data. We can assume that the anti-log of the mean difference will give us the ratio of the median income of those with 16 years of education to the median income of those with 12 years of education. Taking the anti-log of -0.5699, we get the value 0.566. To get the upper limit of the median, we have to take the anti-log of -0.4844, which gives us a value of .616. The lower limit of the median, we have the anti-log of –infinity, which gives us a value of 0.

Educ	N	Mean	Std Dev	Std Err	Minimum	Maximum
12	1020	10.2272	0.8540	0.0267	5.7038	12.9239
16	406	10.7971	0.9581	0.0475	5.2983	13.1603
Diff (1-2)		-0.5699	0.8848	0.0519		

Educ	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
12		10.2272	10.1747	10.2797	0.8540	0.8185	0.8927
16		10.7971	10.7036	10.8906	0.9581	0.8964	1.0290
Diff (1-2)	Pooled	-0.5699	-Infty	-0.4844	0.8848	0.8535	0.9186
Diff (1-2)	Satterthwaite	-0.5699	-Infty	-0.4800			

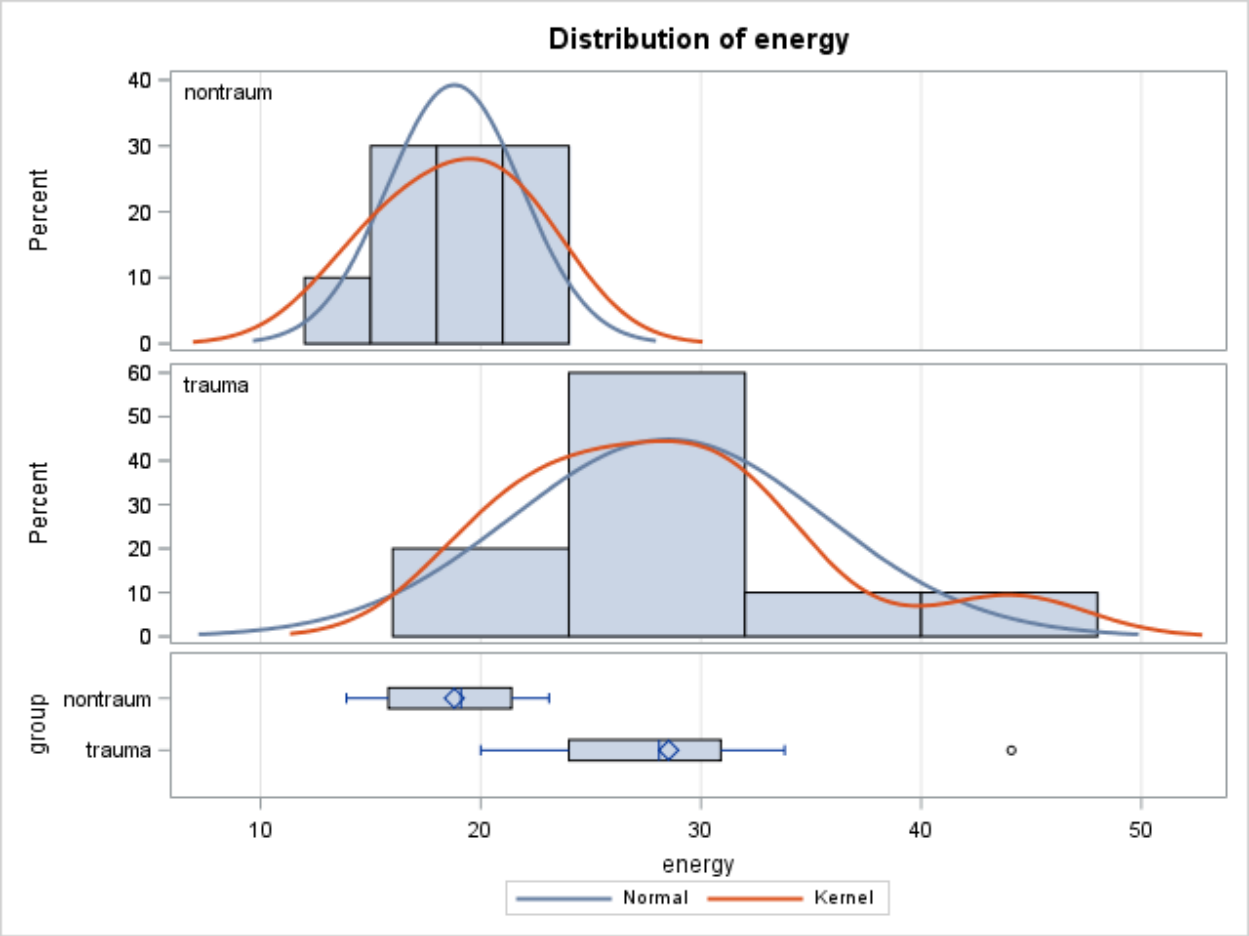
Method	Variances	DF	t Value	Pr < t
Pooled	Equal	1424	-10.98	<.0001
Satterthwaite	Unequal	674.82	-10.45	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	405	1019	1.26	0.0047

4. There is sufficient evidence at alpha = .05 level of significance that the median income of those with 16 years of education is 0.566 times that of those with 12 years of education. We have a 95% confidence level that the median income level of those with 16 years of education are up to 0.616 times higher than those with 12 years of education.

Bonus Question.

1. Based on the histogram and distribution, there is visual evidence that the variances differ in the non-trauma and trauma patients. As secondary evidence using the F-Test, we find the p-value of 0.0189 is smaller than the significance level of 0.05; hence we reject H_0 of equality and conclude there is evidence to suggest that the variances differ.



Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	9	9	5.44	0.0189

2. u_{NT} = 18.8, u_T = 28.53
   a. Mean Difference = 28.53 – 18.8 = 9.73
   b. Median Difference = 28.1 – 19.1 = 9

3. Random Sampling using Excel:

Group	Energy	Rand()
NT	20	0.58437926
NT	30.9	0.18131839
NT	19.9	0.80419437
NT	21.4	0.24855816
NT	44.1	0.63154023
NT	24	0.42948104

NT	13.9	0.88660599
NT	20.6	0.39645149
NT	17.9	0.61734361
NT	23.1	0.42016226
T	15.8	0.41377354
T	21.7	0.16085427
T	15.4	0.36142133
T	30	0.2518776
T	20.6	0.98021153
T	30.6	0.57991294
T	33.8	0.78386425
T	18.3	0.62663564
T	25.1	0.67114108
T	26.2	0.21320052
AvgNT	**23.58**	
AvgT	**23.75**	
MedianNT	**21**	
MedianT	**23.4**	

4. Mean and Median Differences
    a. Sample 1
        i. Mean: 0.17
        ii. Median: 2.4

5. Mean and Median Differences (Continued, please see appendix for Excel spreadsheet calculation)
    a. Sample 2
        i. Mean: -1.61
        ii. Median: -1.7
    b. Sample 3
        i. Mean: -0.82
        ii. Median: -0.35
    c. Sample 4
        i. Mean: 1.38
        ii. Median: -0.1
    d. Sample 5
        i. Mean: 3.85
        ii. Median: 4.8
    e. Sample 6
        i. Mean: -0.57

ii. Median: -2.4

6. The randomization test would not be affected much for the median difference since the median is a statistically robust parameter. Since the median is calculated as the middle value, not the average of all the values, an outlier will not affect the tests as much as the mean of all the values. However, the randomization test will be affected for the mean difference as an outlier will raise the mean significantly especially in a small data set. Although it would not stop us from running the randomization test, the outlier value will affect the data significantly and our mean differences would not be valid.

   The t-test would be affected by an outlier as it would mean that it would not satisfy the normality assumption. We would not be able to use the t-test and would have to investigate the outlier before proceeding.

Appendix: Calculation from Excel for sample mean and median differences

Group	Energy		Energy	Rand()		Energy	Rand()		Energy	Rand()		Energy	Rand()		Energy	Rand()		Energy	Rand()
NT	13.9		20	0.47571132		16.8	0.14866304		20.7	0.10410433		18.2	0.67355674		6.1	0.93377337		21.4	0.88673211
NT	15.4		30.9	0.67976771		21.2	0.09358439		19.3	0.91984925		22.1	0.50832644		12.9	0.00655116		30.6	0.03880809
NT	15.8		19.9	0.44893135		22.1	0.90305829		13.6	0.11801353		12.3	0.00249071		16.6	0.62934466		25.1	0.33491598
NT	17.9		21.4	0.21289278		20.6	0.81179309		20.6	0.18511001		15	0.20633937		20.6	0.20890345		26.2	0.67841257
NT	18.3		44.1	0.47135046		12	0.60683036		17.2	0.04919021		19.2	0.32100888		19.1	0.5196478		30.9	0.83983042
NT	19.9		24	0.16920447		24	0.99149039		12.3	0.43872706		17.2	0.55723251		13.6	0.15804568		18.3	0.33910324
NT	20.6		13.9	0.86623237		18.2	0.9352845		17.2	0.75022576		19.3	0.89678083		12.3	0.64198917		30	0.21677285
NT	21.4		20.6	0.49783526		18.7	0.75273378		22.1	0.77059853		24	0.64178947		12	0.49673675		15.8	0.53513866
NT	21.7		17.9	0.47955098		19.1	0.35987778		19.5	0.69113386		19.5	0.82356445		26.7	0.03870988		20.6	0.2511589
NT	23.1		23.1	0.39526446		19.3	0.69764243		17.4	0.72029022		6.1	0.34919327		19.3	0.17510419		20.6	0.44519569
T	20		15.8	0.14296321		24.3	0.02857788		10.9	0.39030777		17.4	0.01233487		18.5	0.55711827		21.7	0.88351698
T	20.6		21.7	0.62545493		6.1	0.48137599		11.8	0.61916741		13.6	0.26497995		22.2	0.7425484		33.8	0.74091668
T	24		15.4	0.8414933		17.5	0.19201071		19.1	0.62803296		21.3	0.40797248		17.5	0.78296422		13.9	0.8642121
T	25.1		30	0.44428375		17.2	0.91014115		16.8	0.3727291		19.8	0.60747921		20.3	0.29920326		23.1	0.05759122
T	26.2		20.6	0.08799695		17.5	0.28107868		12	0.56830146		24.3	0.78848337		20.7	0.22583317		17.9	0.25197869
T	30		30.6	0.42171802		20.3	0.60782256		20.3	0.82848273		12	0.49315511		17.5	0.69896566		19.9	0.49620269
T	30.6		33.8	0.15597747		20.5	0.89734589		18.5	0.07496668		24	0.41941866		21.6	0.1427316		24	0.19672455
T	30.9		18.3	0.67583514		17.4	0.62451573		22.6	0.77655704		17.2	0.4031117		19.5	0.31030297		20	0.95319109
T	33.8		25.1	0.25501091		22.2	0.55566869		17.5	0.37314231		20.3	0.10598217		21.2	0.88515193		44.1	0.92390081
T	44.1		26.2	0.21130268		12.9	0.00410527		22.2	0.6119794		16.8	0.74144343		18.7	0.8073472		15.4	0.62245113
AvgNT	18.8	AvgNT	23.58		AvgNT	19.2		AvgNT	17.99		AvgNT	17.29		AvgNT	15.92		AvgNT	23.95	
AvgT	28.53	AvgT	23.75		AvgT	17.59		AvgT	17.17		AvgT	18.67		AvgT	19.77		AvgT	23.38	
AvgDiff	9.73	AvgDiff	0.17		AvgDiff	-1.61		AvgDiff	-0.82		AvgDiff	1.38		AvgDiff	3.85		AvgDiff	-0.57	
MedianNT	19.1	MedianNT	21		MedianNT	19.2		MedianNT	18.35		MedianNT	18.7		MedianNT	15.1		MedianNT	23.25	
MedianT	28.1	MedianT	23.4		MedianT	17.5		MedianT	18		MedianT	18.6		MedianT	19.9		MedianT	20.85	
MedianDiff	9	MedianDiff	2.4		MedianDiff	-1.7		MedianDiff	-0.35		MedianDiff	-0.1		MedianDiff	4.8		MedianDiff	-2.4	