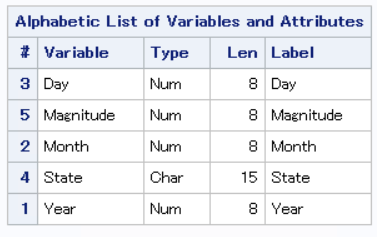
**Predictive analysis using SAS**

Question1:

**Dataset - Earthquake**

The given dataset has 310 observations with 5 variables

The variables fields are Magnitude of the earthquake, State in which it occurred, Day , Month, Year in which the earthquake occurred obtained using proc contents. Of which State is char variable and remaining are numbers.



The Dataset print using proc print

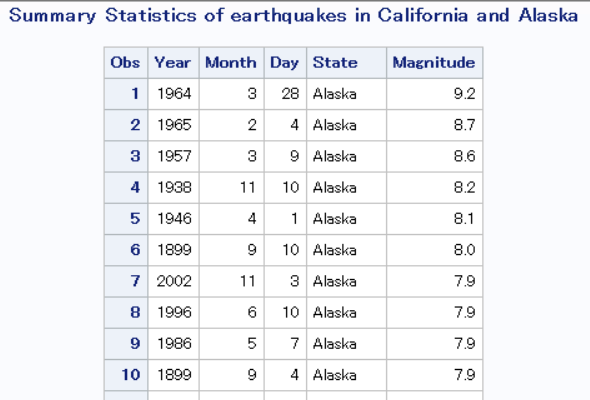


Question 1.a

Filter dataset to take only Alaska and California as these have the high observations of earthquakes.

This filtering can be done by using IF statement and output only rows with state=Alaska or State =California when the data is loaded.

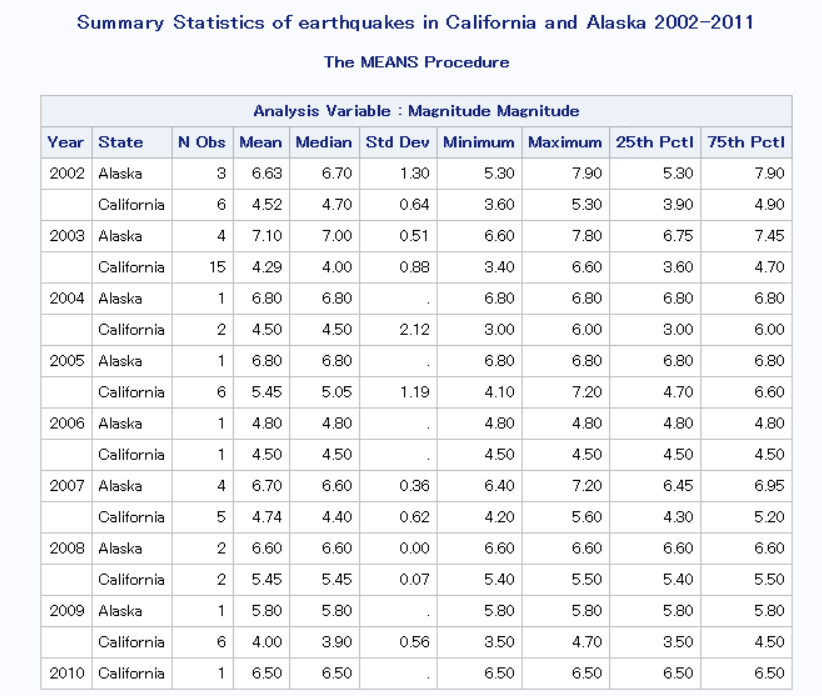
The proc contents show that the filtered dataset has 169 records.



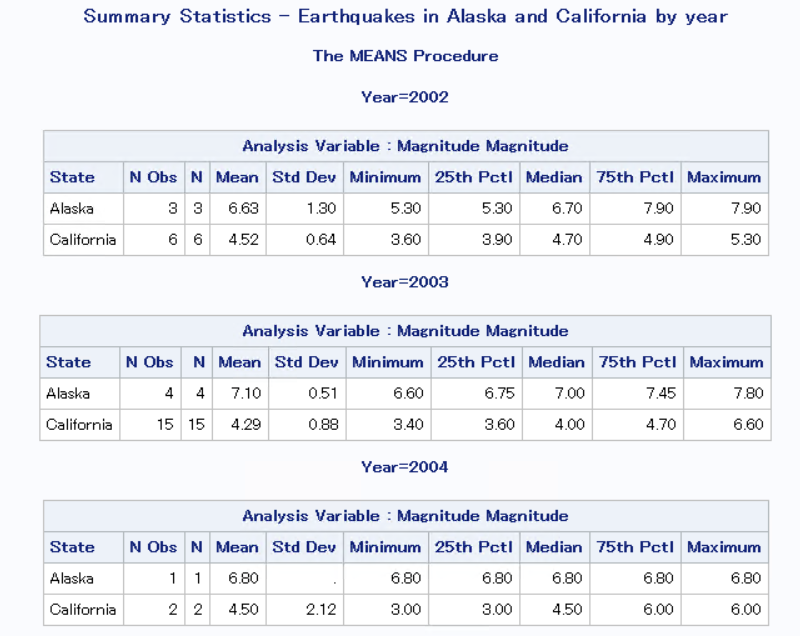
Question 1b. Summary statistics of earthquake in California and Alaska with year 2002-2011 only considered.

Year 2002-2011 can be filtered using WHERE clause with SET when the data is loaded.

Using PROC MEANS and specified what statistical parameters are required we will be able to obtain the below summary statistics

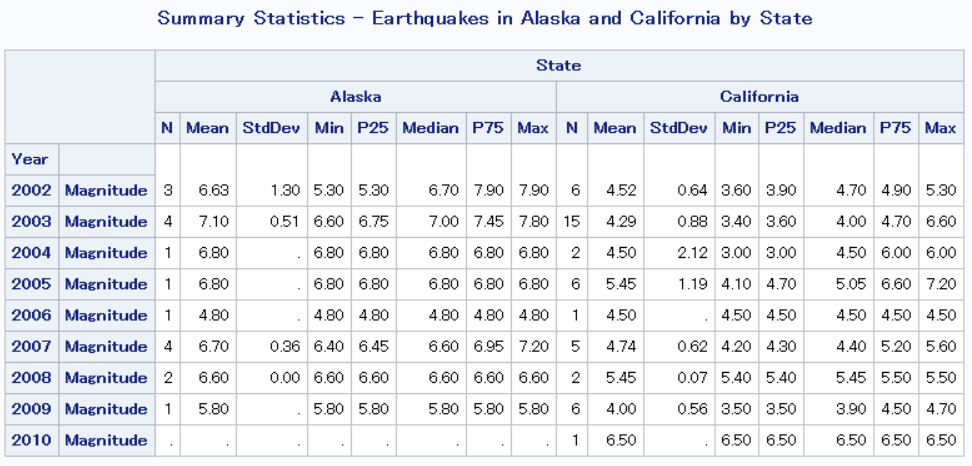


Question 1 c. To modify the tabular display to show split by Year of the earthquakes in the two States , we use BY clause as Year and CLASS clause as State with PROC MEANS

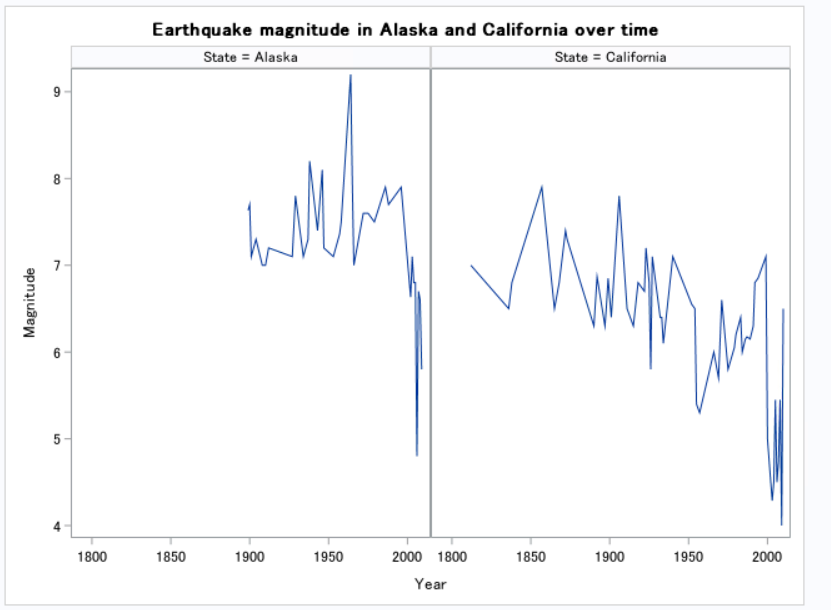


Question 1d. Display the earthquakes in the two states with year as 1st column and State as the header row.

This can be done by using the CLASS clause and variable with a \* showing that Year is column and State is header Row.



Question 1e. Display the time series plot for Alaska and California earthquakes side by side. This can be done by using the SGPANEL which displays the plot side by side. The time series plot is calculated using SERIES in PROC with the avg magnitude of earthquake over the years. Before doing this the dataset has to be sorted by year for SAS to arrange the data in times series. The times series has x and y axis with Year and Avg Magnitude.



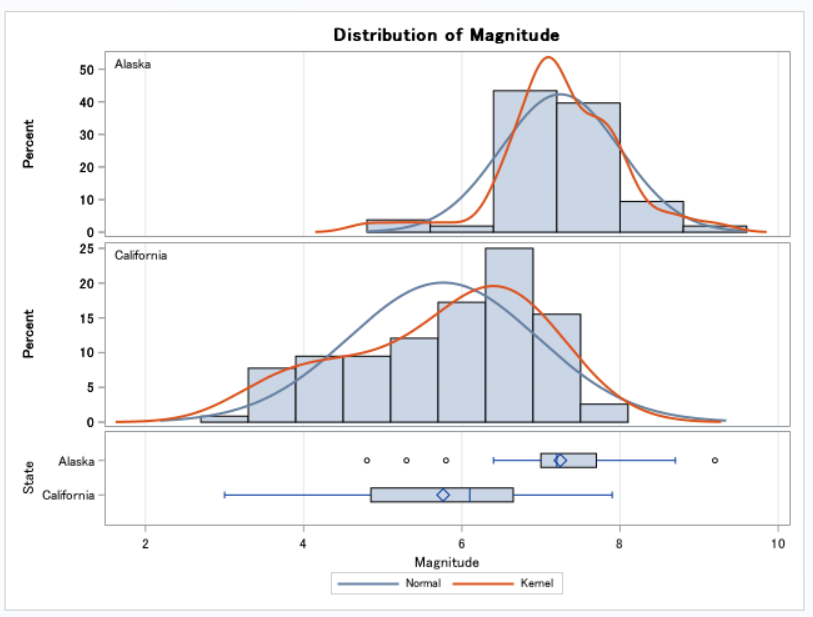
Question 1f: Test the following null hypothesis: “the average magnitude of earthquakes in California is equal to that of Alaska”.

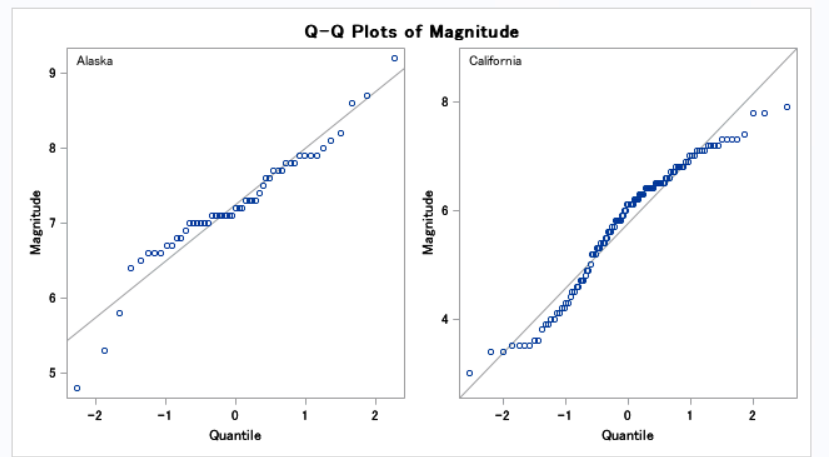
**Null Hypothesis:** The average magnitude of earthquakes in California is equal to the average magnitude of earthquake in Alaska.

**Action Hypothesis:** The average magnitude of earthquakes in California is not equal to the average magnitude of earthquake in Alaska.

To test this hypothesis, the t test is conducted with Magnitude as variable and State as the dependent variable PROC TTEST is used to calculate the t test scores. For t test the dataset has to be sorted first with by Year and State and then t test has to be conducted.







**Interpretation:**

The P value is seen to be less than 0.05 thus we can reject H0 and accept Ha. Thus, the conclusion is that the magnitude of earthquakes in California and Alaska are not equal. Looking at the t-value we can conclude that and F value we can conclude that we can reject the null hyptothesis.

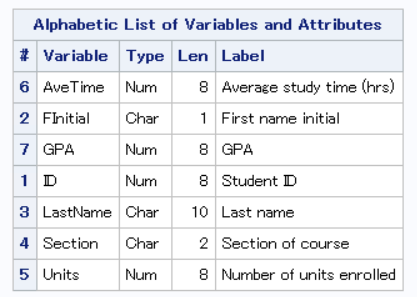
**Result: Earthquake magnitude in California is not equal to earthquake magnitude in Alaska**

**Question 2:**

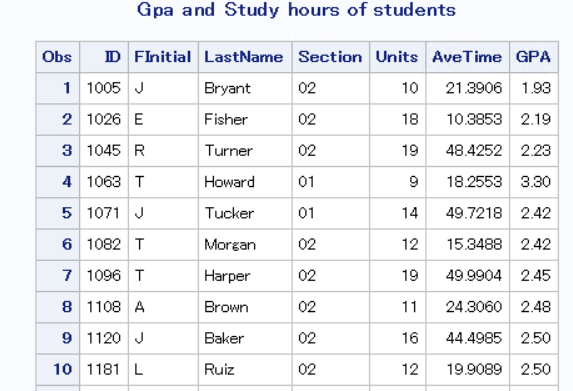
**Dataset – GPA**

The dataset has 122 observation with 7 variables

The variables as Section of course, average study time, name initial, GPA secured, Student IS, last name of student and number of units enrolled. The first initial, last initial and section are character fields



The dataset is displayed below.

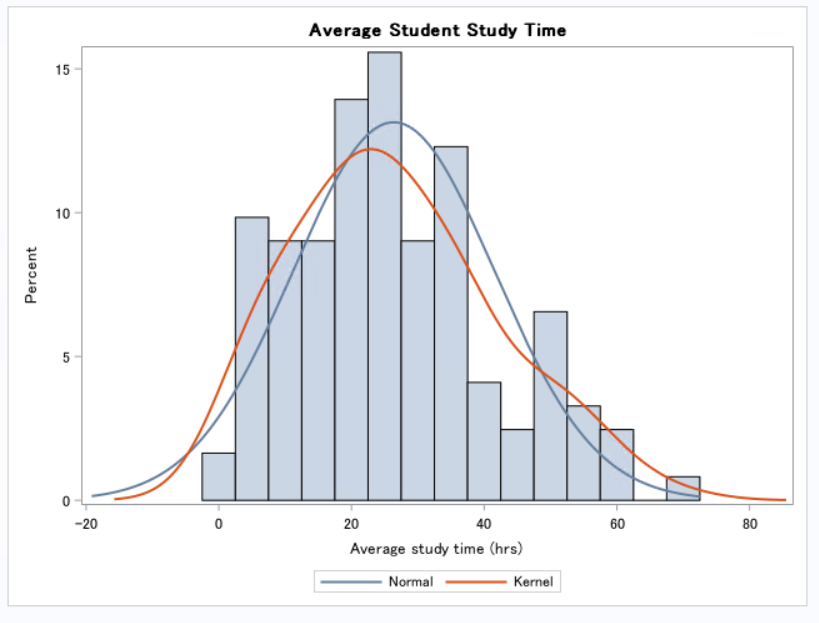


Question 2a: Display the histogram for hours of study. Use the start point=0 and bandwidth=5.

This can be done with the histogram function SGPLOT with PROC and histogram function with Average Time.

The Normal distribution is represented by the blue line and the kernel distribution is represented with orange line. The kernel line is a more accurate line showing the curves with the population of the data while normal distribution line provides a smoothened curve.

The graph shows that average study time is a normal distribution with more observations in 20-30 hours. The dataset has observations ranging 0-75 hours of average time spent in studying.



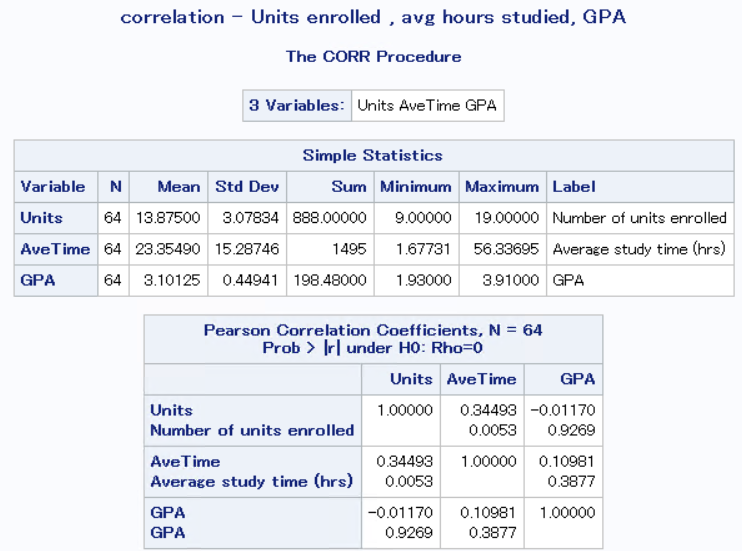
Question 2b: Check the correlation between Units enrolled and average hours studied with GPA. This checks if there is any direct impact of these two variables with GPA.

**Null Hypothesis H0:** There is no significant correlation between the variables (The variables are independent)

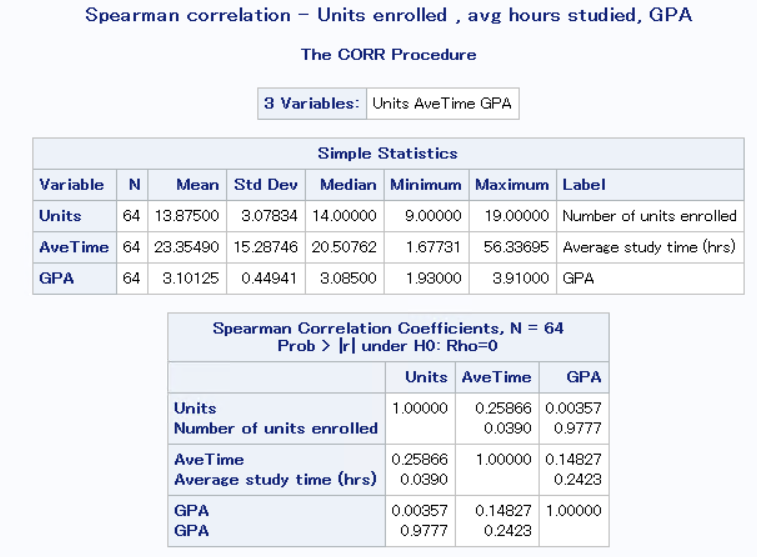
**Action Hypothesis Ha**: There is significant correlation between the variables (The variables are dependent)

The Pearson and Spearman tests are done for correlation to be certain of the impact and its surety.

The PROC CORR clause is used to perform Pearson correlation test on the variables.



The PROC CORR with SPEARMAN is used to perform spearman correlation on the variables to visualise its impact.

 `

With the Pearson test we can see conclude these interpretations:

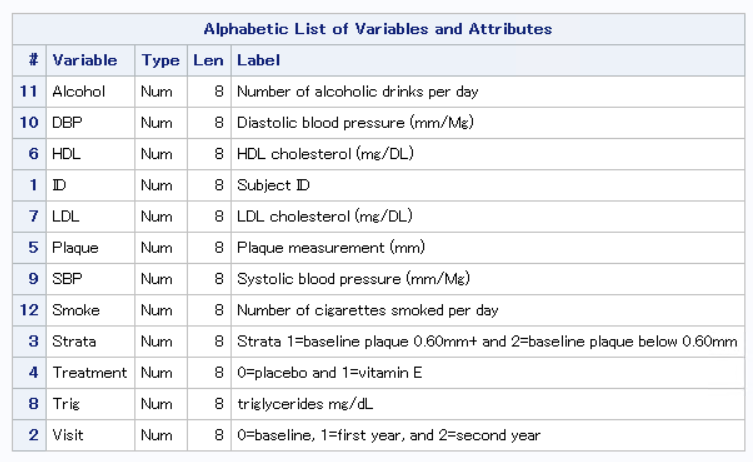
1. Units are positively trending with ave time, as units increase the number of hours studied is increased. Thus, there is a significant correlation between these two variables.
2. Units is negatively correlate with GPA. The more units the less the GPA gets to. Thus, there is no significance between these.
3. The relationship between ave time and GPA is negative, the more the hours the student studies the more the GPA the student secures. This is not true and is changed. Thus it can be concluded that there is no significance between these and that there may be other variables affecting these.

**Question 3:**

**Datatset : VITE**

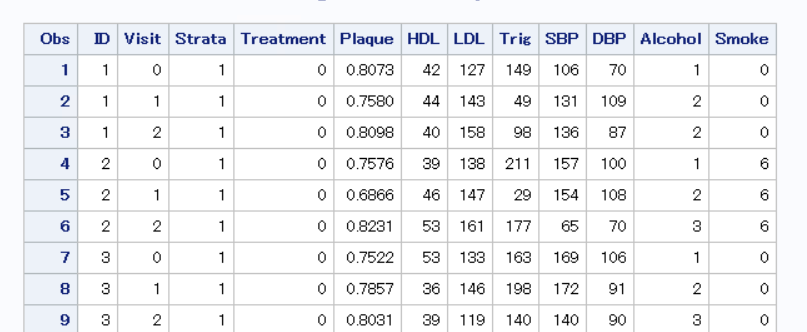
The dataset has 1500 observations with 12 variables

The variables are displayed below, and all are in numeric format.

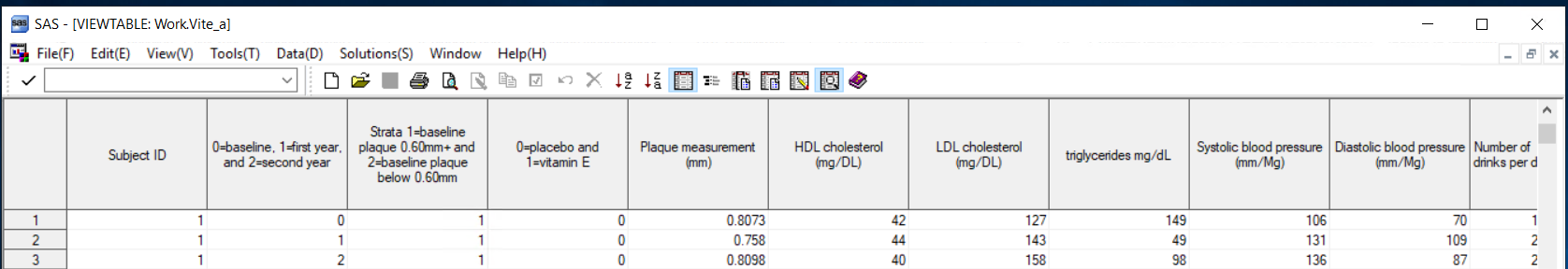


**Question 1a. Change the data display to show only headers as variable**

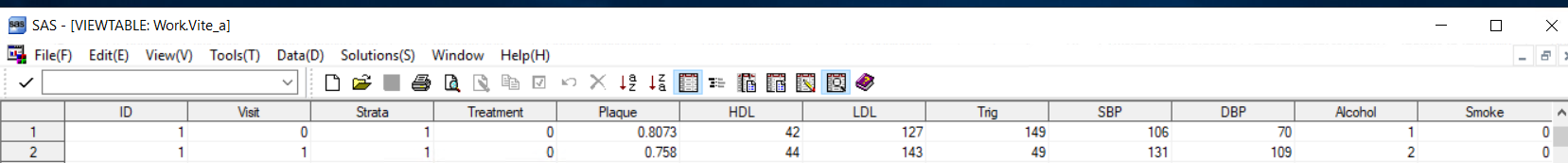
When the dataset is printed using PROC PRINT is as displayed below in the result viewer.

****

But when the file is opened to view the file we can see that the file is displayed this Label names.

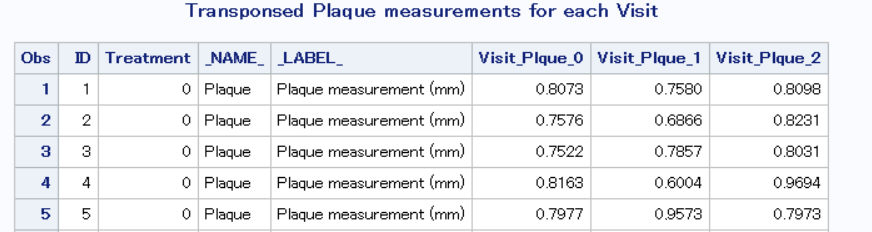


Then we need to change the view to display only column headers and not label names. This can be done by choosing View -> and select Column names.



Question 3b. Transpose the rows to represent the visits as columns rather than as rows. This compresses the dataset and the dataset is crisp to perform analysis.

This is done by using PROC TRASNPOSE. The dataset is sorted by ID and treatment. Visit is used as id to sort the plaque with prefix as visit\_plaque\_ .



Question 3c.

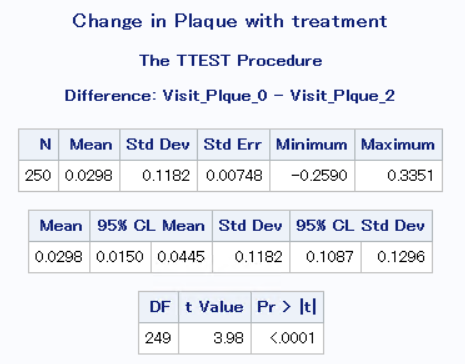
**Null Hypothesis H0:** There is no difference in plaque levels before and after treatment using vitamin E

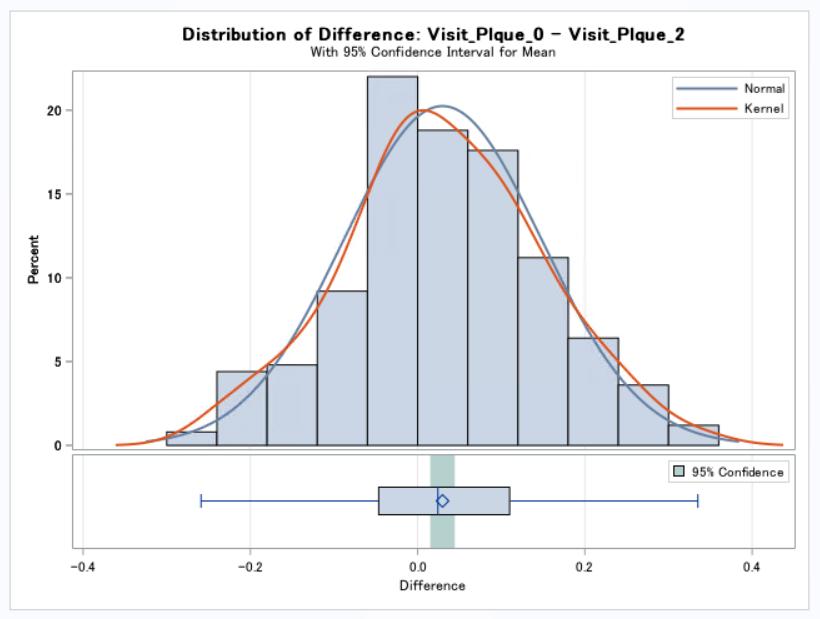
**Alternate Hypothesis H1:** There is a difference in plaque levels before and after treatment using vitamin E

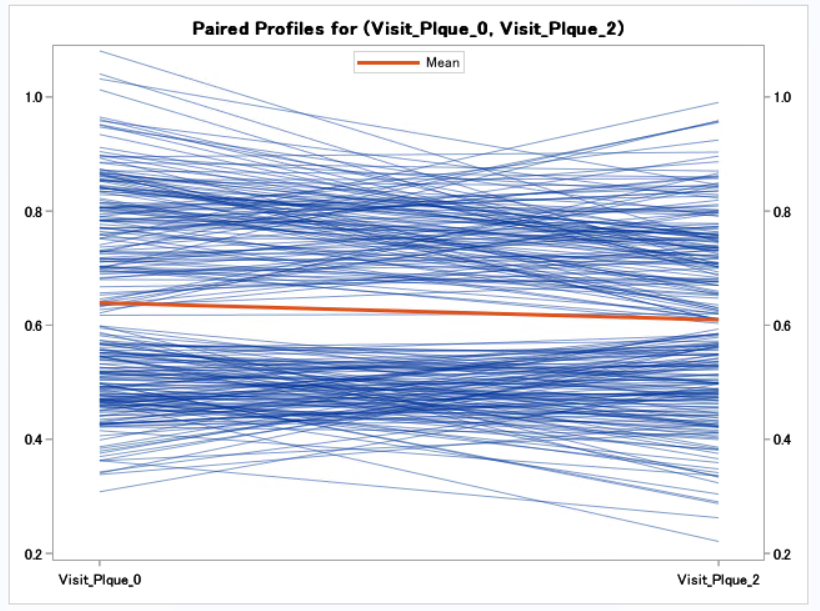
Assume there is no control group and check if there is a change in the plaque value in visit 0 and visit 2.

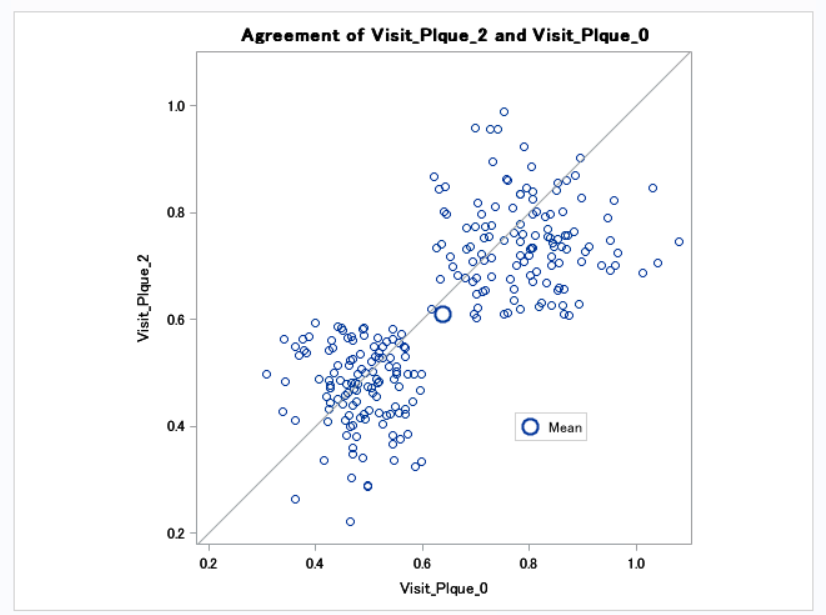
The dataset is first filtered using WHERE clause in SET function to remove the treatment=0. Thus, the resulting dataset only contains the observations when treatment is done.

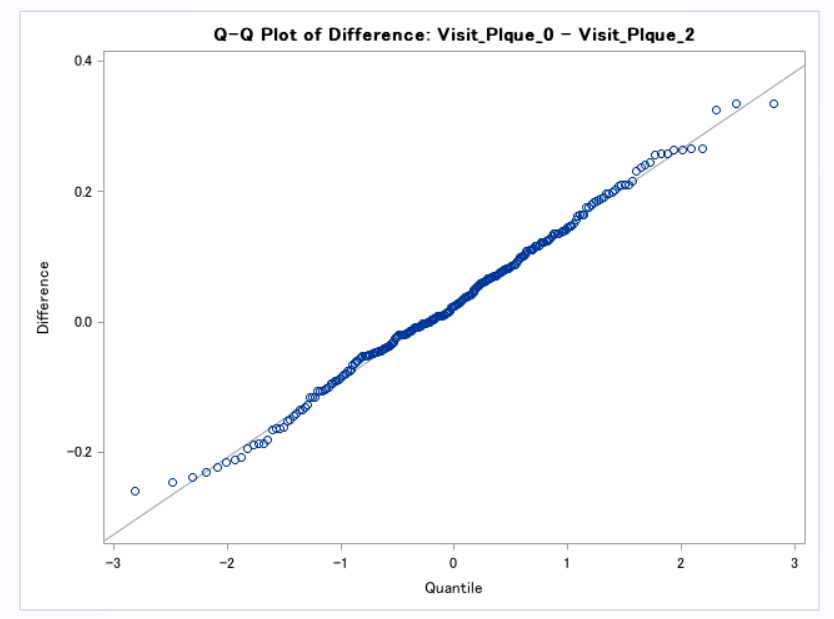
Ttest is now performed on the variables to check the null hypothesis.











**Interpretation:** We see that the P value is less for this test of paired variables and can thus reject the null hypothesis.

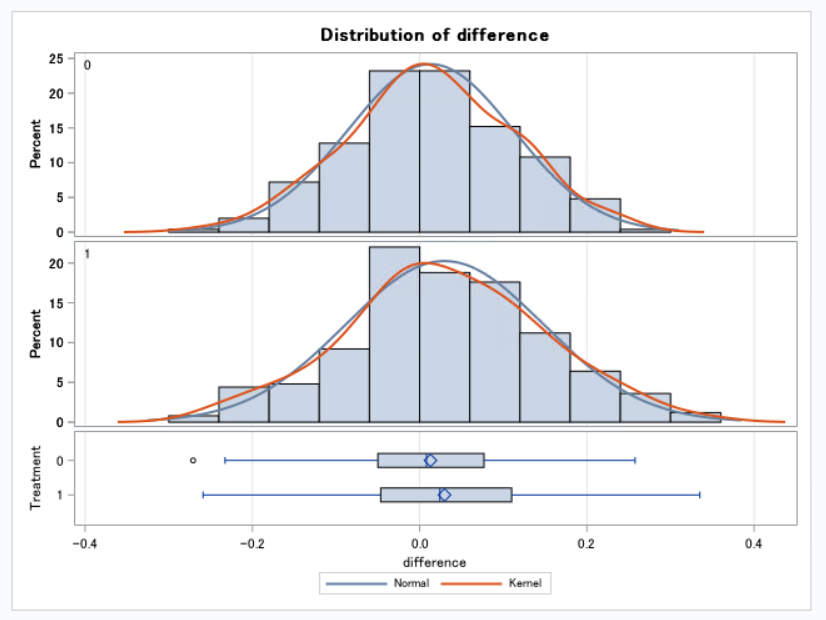
**Conclusion:** There is a difference in the plaque levels in the after the treatment. This shows that the treatment using Vitamin E helps the patients in the sample group of treatment =1.

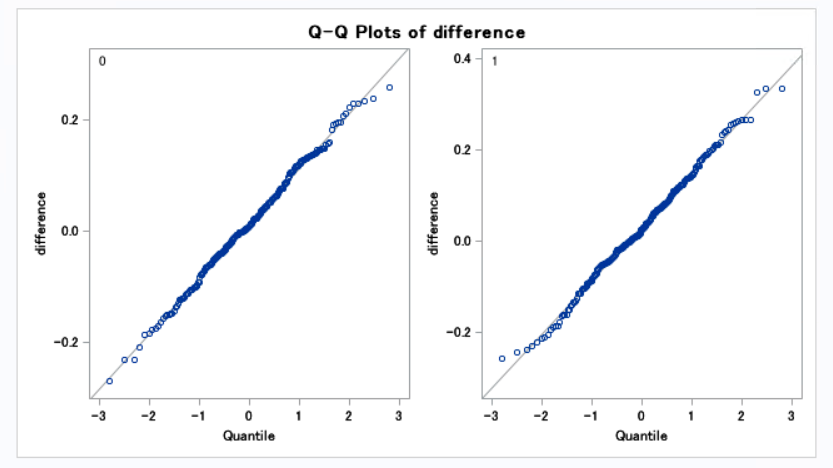
**Question 3d.** The dataset is now taken to include both the treatment 0 and 1. The control and the test group are taken into consideration for this. This helps to know the effect of the plague with the treatment and its effect without treatment. This helps decide if the treatment helped or not. Thus as the treatment was with vitamin E, treatment =1 helps decide is vitamin E helps to reduce the plague or not.

**Null Hypothesis H0:** There is no difference in plaque levels between the treatment and control groups

**Alternate Hypothesis H1:** There is a significant difference in plaque levels between the treatment and control groups







**Interpretation:**

Looking at the p value of 0.08 which is greater than 0.05 we accept the null hypothesis.

**Conclusion:** Stating that the there is no significant difference between the treatment group and the control group.

**Question 3e**. To state whether test 3c or 3d is better we need to check if the quality of the sample considered. When we take a sample inclusive of only the treatment =1 group which means only patients who received vitamin E, a betterment or degradation in the plague will not be conclusive if its because of just the treatment or was some other factor affecting the sample which the dataset didn’t consider or the right variables were not chosen for the test.

But in the case of test 3d, we have included both treatment 0 as well as treatment 1 to be inclusive of the dataset to show that when the patient who took the treatment and the patient who dint take treatment were both representative of a good sample and no other varying factor affects just one group. This helps to prove that vitamin consumption helped patients in treatment 1 and not consuming vitamin e by patients in treatment =0 shows that it did not help them.

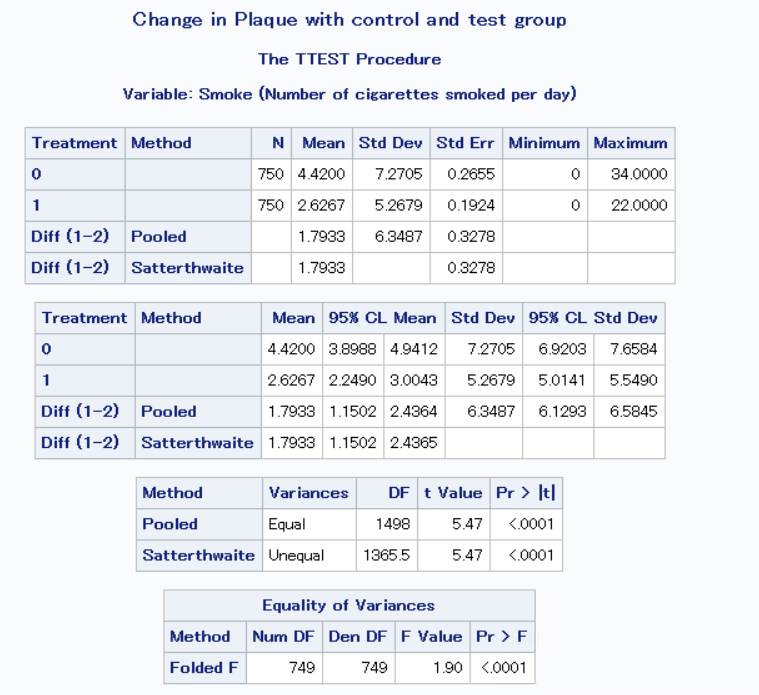
Thus, the test shows that there is difference in plaque levels in both the control as well as the treatment group with a significant difference in plaque levels in treatment group showing that it is because of treatment with vitamin E

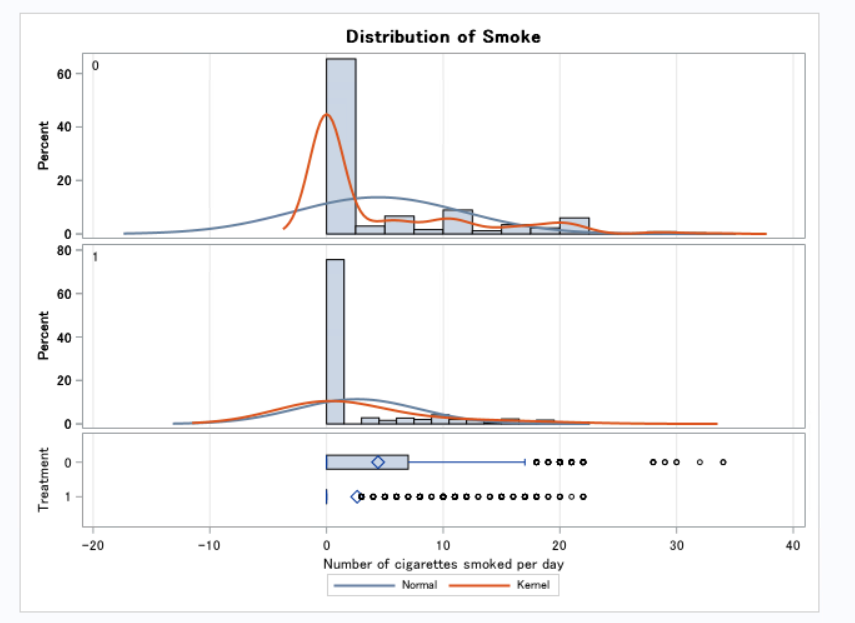
In conclusion, test done in 3D is good to get a proper analysis of the sample to show if vitamin E helped the patients to fight plague or not.

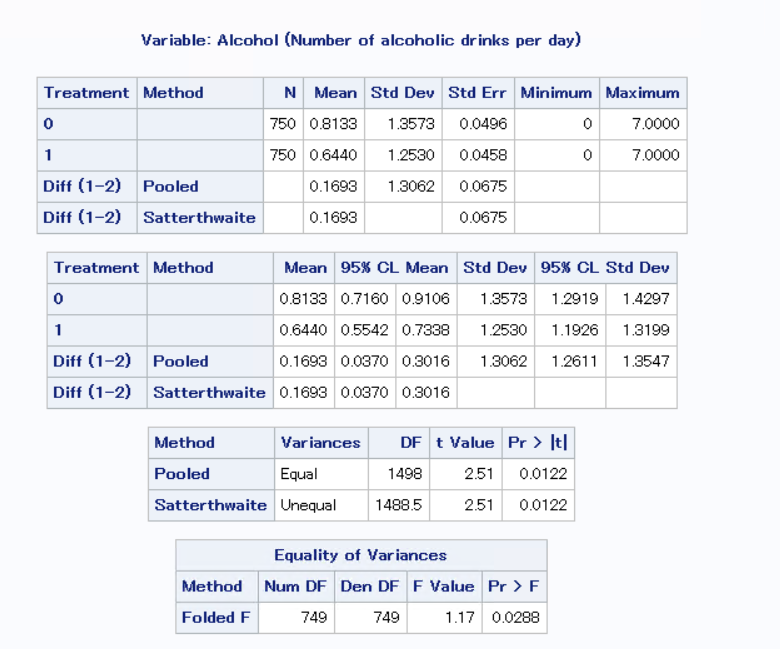
**Question 3F.** Randomising to show if the variable smoke and alcohol as represented in the dataset.

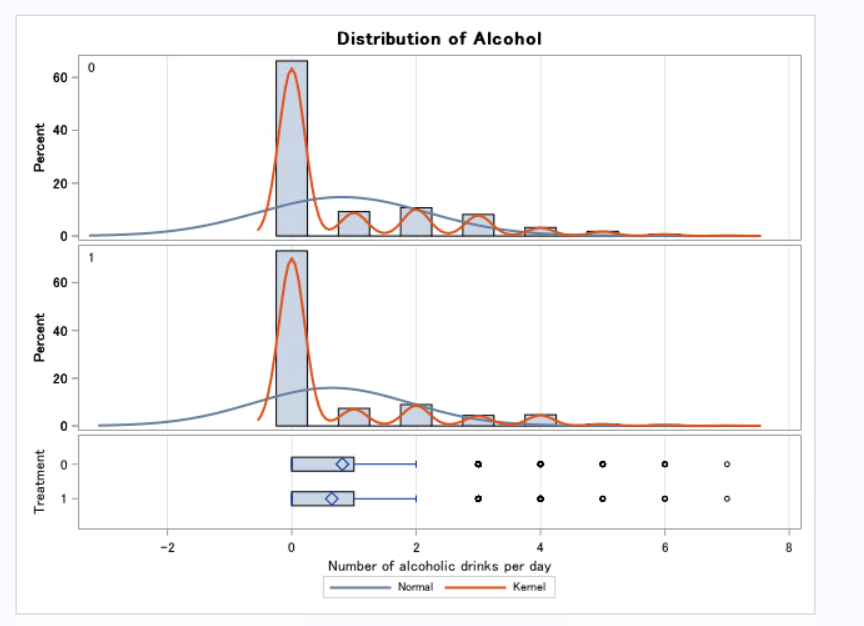
**Null Hypothesis H0**: There is no difference in averages with alcohol and smoke for both treatment and control groups

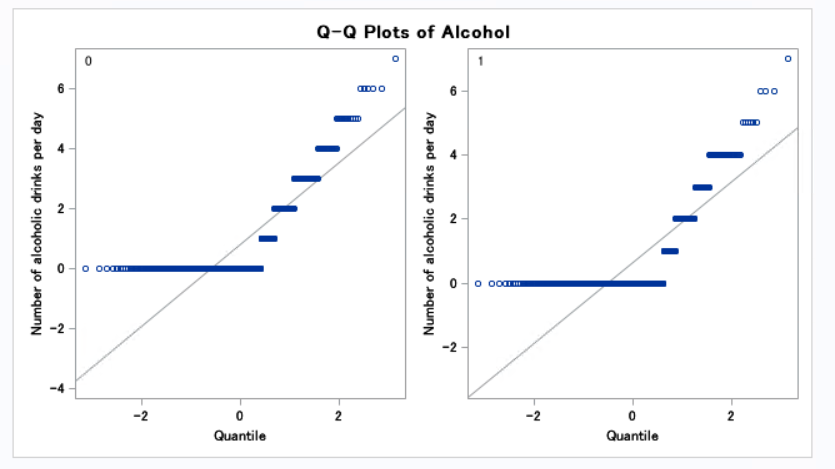
**Alternate Hypothesis H1**: There is a significant difference in averages with alcohol and smoke for both treatment and control groups





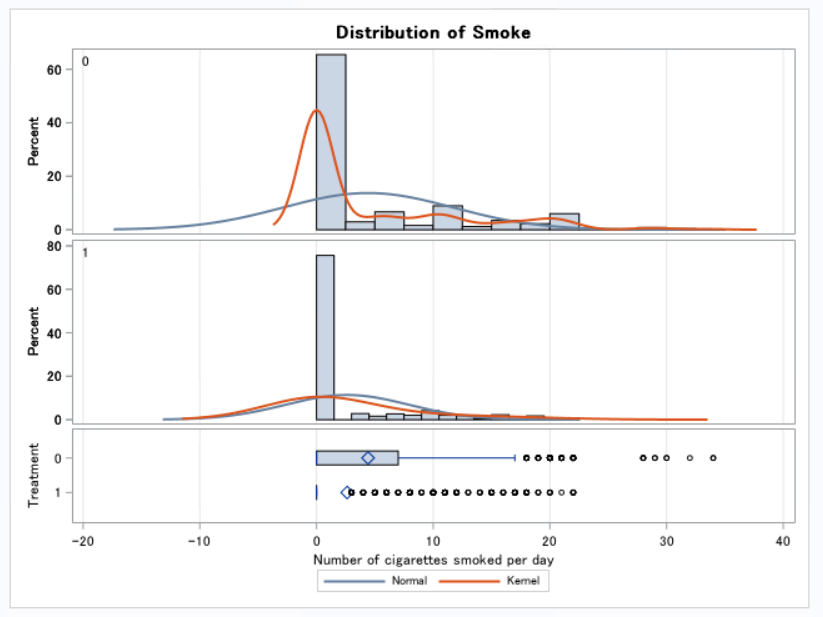


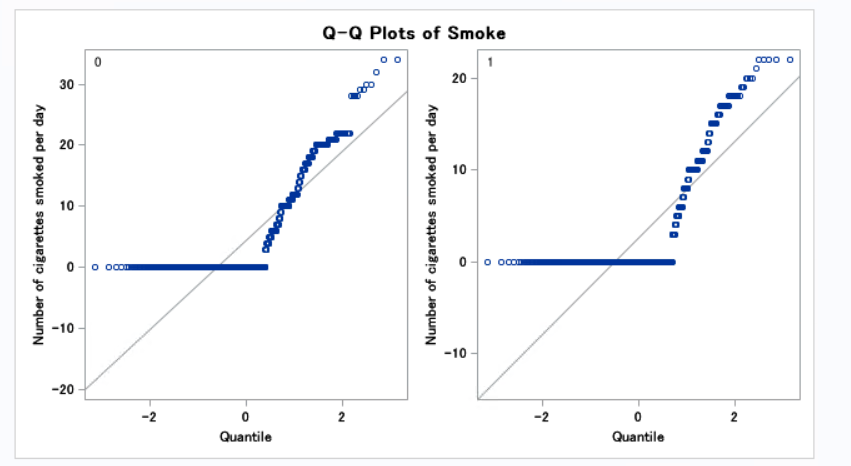




**Individual t test – I perform the individual t test to get a clear idea of the performance of that variable.**

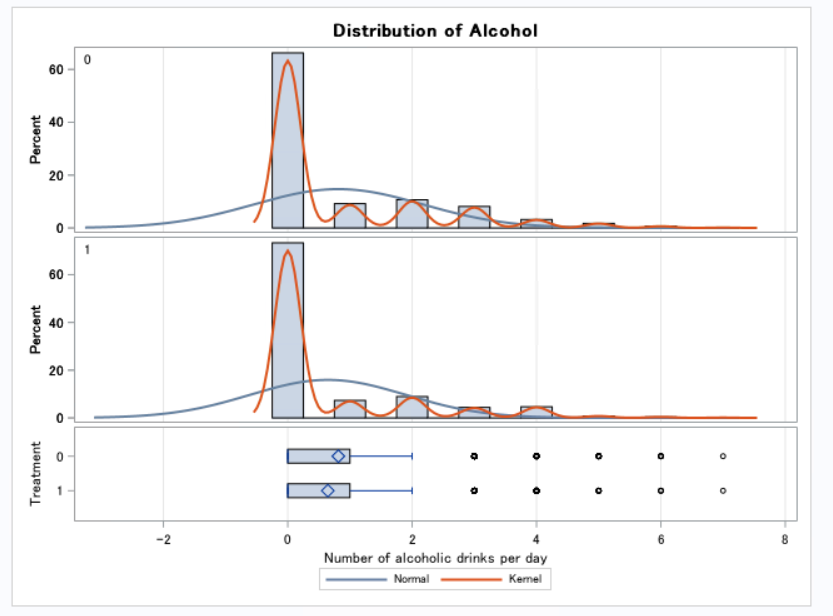


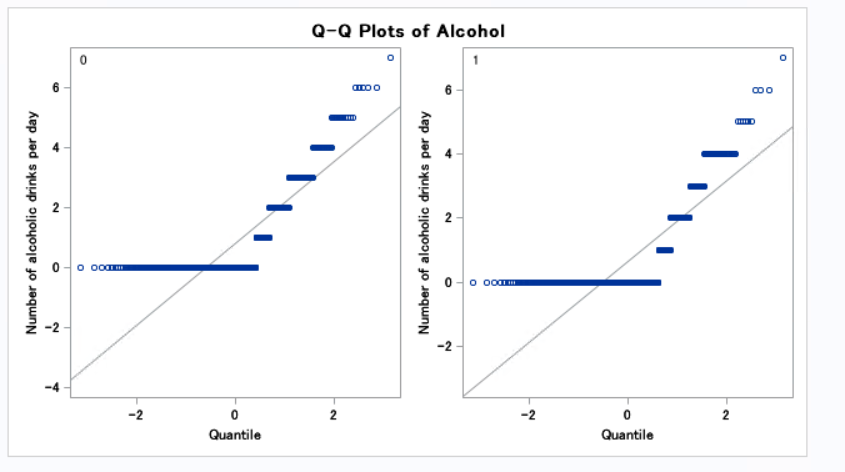




individual t test for alcohol







**Interpretation:**

We can see that the p value is less than alpha of 0.05 for both the methods in the graph thus we can reject the null hypothesis.

**Conclusion:**

There is a significant difference in averages with alcohol and smoke for both treatment and control groups