1. The United States Geological Survey provides data on earthquakes of historical interest. The SAS data set called EARTHQUAKES contains data about earthquakes with a magnitude greater than 2.5 in the United States and its territories. The variables are year, month, day, state, and magnitude.

(a) California and Alaska are the two states with the highest number of earthquakes in the country. Create a new data set that includes only these two states and use this data set to answer the following questions.

Code:

```
/*Part a*/
/*Import Data Earth Quake*/
LIBNAME HWDATA 'HWDATA';

DATA HWDATA.earthquakes;
INFILE 'earthquakes.sas7bdat';
INPUT YEAR MONTH DAY STATE MAGNITUDE;
RUN;

PROC SQL;
select *
from HWDATA.earthquakes
WHERE State = 'Alaska' or State = 'California'
ORDER BY State;
quit;
```

Output:

Time Series State/Magnitude

Year	Month	Day	State	Magnitude
1927	10	24	Alaska	7.1
2003	3	17	Alaska	7.1
1958	4	7	Alaska	7.3
1910	9	9	Alaska	7
1900	10	9	Alaska	7.7
1929	3	7	Alaska	7.8
1953	1	5	Alaska	7.1
2003	6	23	Alaska	6.9
1996	6	10	Alaska	7.9
1958	7	10	Alaska	7.7
2007	8	15	Alaska	6.5
1988	3	6	Alaska	7.7
1975	2	2	Alaska	7.6
1964	3	28	Alaska	9.2
1957	3	16	Alaska	7
1966	8	7	Alaska	7
1979	2	28	Alaska	7.5

(b) You are interested in the following statistics for the magnitude of earthquake:

- Mean
- Median
- Standard deviation
- Minimum and maximum
- 25th and 75th percentiles

Create a table that shows the above statistics across different states within each year. In particular, your table must have years at the first column and it must break down the results across different states in the second column. In order to make the table short, further assume you are interested only in recent years and want to create a table that shows the desired statistics from 2002 to 2011.

Code:

```
/*Part b*/
/*Import Data Earth Quake*/
LIBNAME HWDATA 'HWDATA';

DATA HWDATA.earthquakes;
INFILE 'earthquakes.sas7bdat';
INPUT YEAR MONTH DAY STATE MAGNITUDE;
RUN;

ODS CSV FILE = 'E:\Users\nsm190002\Desktop\HWDATA\questionb.csv';

/*Write SQL query to filter selected entries for Year(2002-2011)*/

PROC SQL;
select *
from HWDATA.earthquakes
WHERE Year BETWEEN 2002 and 2011
ORDER BY Year;
quit;
ODS CSV CLOSE;


/*Write SQL query to filter selected entries for Year(2002-2011)*/

PROC SQL;
select *
from HWDATA.earthquakes
WHERE Year BETWEEN 2002 and 2011
ORDER BY Year;
quit;
ODS CSV CLOSE;

/*Use Import Wizard to create a program for importing saved questionsb.csv file*/
PROC IMPORT OUT= HWDATA.QUESTIONB
            DATAFILE= "E:\Users\nsm190002\Desktop\HWDATA\questionb.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/*Print Statistics*/

PROC MEANS DATA = HWDATA.questionb n mean stddev min p25 median p75 max maxdec=2;
CLASS Year State;
VAR Magnitude;

title'Summary Statistics';
RUN;
```

Output:

Summary Statistics

The MEANS Procedure

Analysis Variable : Magnitude										
Year	State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
2002	Alaska	3	3	6.63	1.30	5.30	5.30	6.70	7.90	7.90
	California	6	6	4.52	0.64	3.60	3.90	4.70	4.90	5.30
	Indiana	1	1	4.60	.	4.60	4.60	4.60	4.60	4.60
	New York	2	2	4.20	1.27	3.30	3.30	4.20	5.10	5.10
	Oregon	1	1	4.50	.	4.50	4.50	4.50	4.50	4.50
	South Carolina	1	1	4.40	.	4.40	4.40	4.40	4.40	4.40
	Washington	2	2	3.90	0.28	3.70	3.70	3.90	4.10	4.10
	Wyoming	1	1	4.20	.	4.20	4.20	4.20	4.20	4.20
2003	Alabama	1	1	4.60	.	4.60	4.60	4.60	4.60	4.60
	Alaska	4	4	7.10	0.51	6.60	6.75	7.00	7.45	7.80
	Arkansas	1	1	4.00	.	4.00	4.00	4.00	4.00	4.00
	California	15	15	4.29	0.88	3.40	3.60	4.00	4.70	6.60
	Hawaii	1	1	4.70	.	4.70	4.70	4.70	4.70	4.70
	Idaho	1	1	3.30	.	3.30	3.30	3.30	3.30	3.30

(c) Modify you SAS code in (b) such that the results for each year is shown in a separate table.

Code:

```
PROC MEANS DATA = HWDATA.questionb n mean stddev min p25 median p75 max maxdec=2;
CLASS Year State;
VAR Magnitude;
by Year;

title'Summary Statistics';
RUN;
```

Output:

Year=2007

Analysis Variable : Magnitude										
Year	State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
2007	Alaska	4	4	6.70	0.36	6.40	6.45	6.60	6.95	7.20
	California	5	5	4.74	0.62	4.20	4.30	4.40	5.20	5.60
	Hawaii	1	1	5.40	.	5.40	5.40	5.40	5.40	5.40
	Montana	1	1	4.50	.	4.50	4.50	4.50	4.50	4.50

Year=2008

Analysis Variable : Magnitude										
Year	State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
2008	Alaska	2	2	6.60	0.00	6.60	6.60	6.60	6.60	6.60
	California	2	2	5.45	0.07	5.40	5.40	5.45	5.50	5.50
	Illinois	1	1	5.40	.	5.40	5.40	5.40	5.40	5.40
	Nevada	2	2	5.50	0.71	5.00	5.00	5.50	6.00	6.00

Year=2009

Analysis Variable : Magnitude

(d) Now, assume you want to show the same results in part (b) but with the difference that years are shown is the first column and the states are shown in the top row.

Code:

```
/*Part D*/

PROC TABULATE DATA = HWDATA.questionb;
CLASS Year State;
VAR Magnitude;
/*TABLE  Year*Magnitude,(n mean stddev min p25 median p75 max)*(State);*/
TABLE  (Year)*Magnitude,State*(n mean stddev min p25 median p75 max);
By Year;
title'Summary Statistics';
RUN;
```

Output:

Summary Statistics

Year=2009

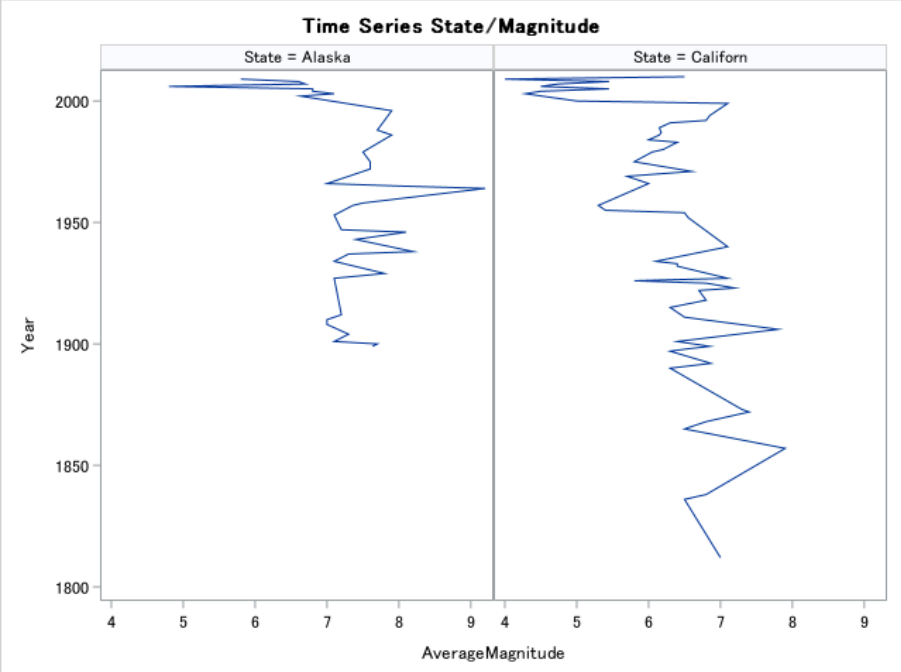
		State																																	
		Alaska								California								Colorado								Hawaii											
		N	Mean	StdDev	Min	P25	Median	P75	Max	N	Mean	StdDev	Min	P25	Median	P75	Max	N	Mean	StdDev	Min	P25	Median	P75	Max	N	Mean	StdDev	Min	P25	Median	P75	Max	N	Mean	Std
Year																																			
2009	Magnitude	1	5.80	.	5.80	5.80	5.80	5.80	5.80	6	4.00	0.56	3.50	3.50	3.90	4.50	4.70	1	3.70	.	3.70	3.70	3.70	3.70	3.70	1	5.20	.	5.20	5.20	5.20	5.20	5.20	1	3.00	

(e) You are interested in how the magnitude of earthquakes is trending over time for each state. In one graph, plot two time series plots, side by side, which shows the trend of average magnitude of earthquakes over time for the two states.

Code:

```
/*Part E*/

PROC IMPORT OUT= HWDATA.QUESTION1A
            DATAFILE= "E:\Users\nsm190002\Desktop\HWDATA\question1a.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

proc sort data = HWDATA.question1a;
by Year State;
run;

proc means data = HWDATA.question1a mean maxdec=2 noprint;
by Year State;
var Magnitude;
output out = means
mean = AverageMagnitude;
RUN;

PROC SGPANEL data = means;
PANELBY State;
SERIES Y=Year X=AverageMagnitude;
Title'Time Series State/Magnitude';

RUN;
```

Output:



(f) Test the following null hypothesis: "the average magnitude of earthquakes in California is equal to that of Alaska".

Code:

```
PROC IMPORT OUT= HWDATA.questione
            DATAFILE= "E:\Users\nsm190002\Desktop\HWDATA\questione.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
/*step 2 - sort data */
proc sort data = HWDATA.questione;
by Year State;
run;
/*step 3 - create Average Magnitude column*/

proc means data = HWDATA.questione mean maxdec=2 noprint;
by State;
var Magnitude;
output out = meane
mean = AverageMagnitude;
RUN;

/*Step 4 - Perform T-Test*/
proc ttest data = meane H0=0 SIDES=2;
class State;
var AverageMagnitude;
run;
```

Output:

Answer: Since P-value here is 0.0009 which is less than 0.05, so we reject the Null Hypothesis

Time Series State/Magnitude

The TTEST Procedure

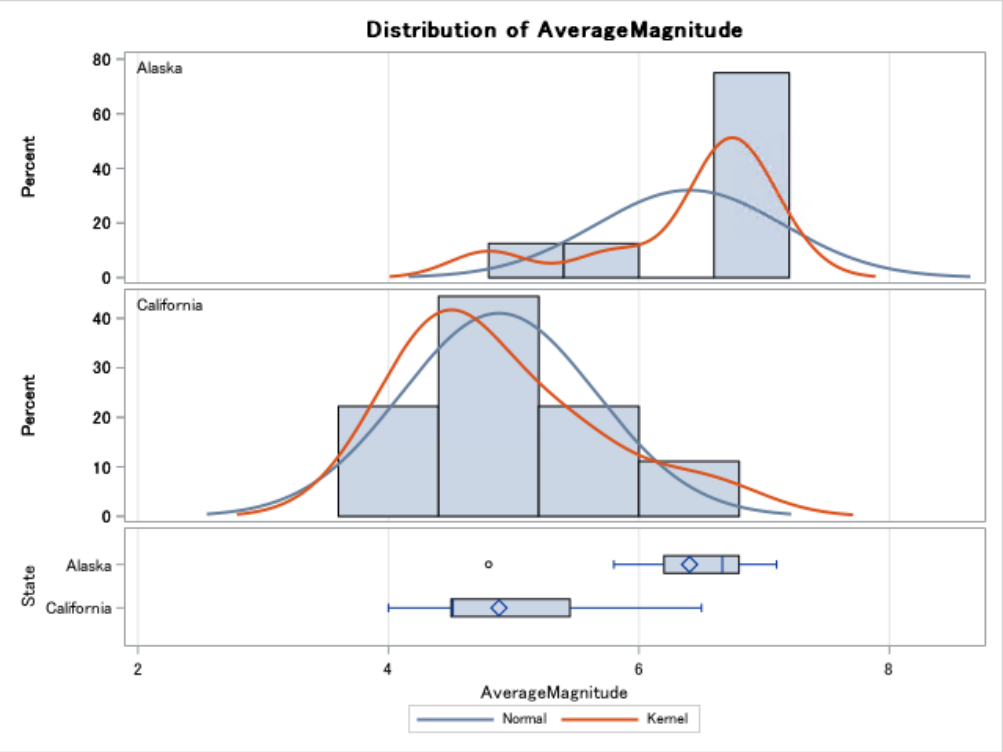
Variable: AverageMagnitude

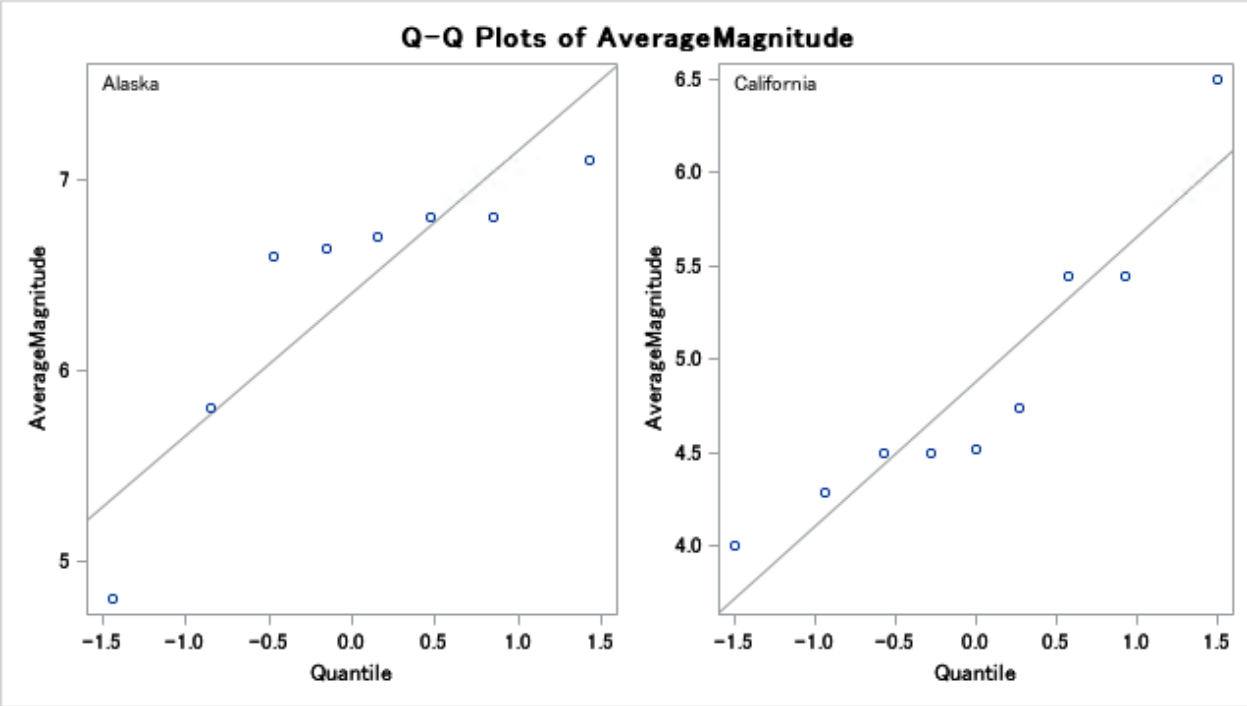
State	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Alaska		8	6.4042	0.7478	0.2644	4.8000	7.1000
California		9	4.8826	0.7779	0.2593	4.0000	6.5000
Diff (1–2)	Pooled		1.5216	0.7640	0.3712		
Diff (1–2)	Satterthwaite		1.5216		0.3703		

State	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Alaska		6.4042	5.7790	7.0293	0.7478	0.4944	1.5220
California		4.8826	4.2846	5.4805	0.7779	0.5254	1.4903
Diff (1–2)	Pooled	1.5216	0.7303	2.3128	0.7640	0.5644	1.1824
Diff (1–2)	Satterthwaite	1.5216	0.7318	2.3114			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	15	4.10	0.0009
Satterthwaite	Unequal	14.889	4.11	0.0009

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	8	7	1.08	0.9301


Distribution of AverageMagnitude
Alaska
California
Percent
State
Alaska
California
AverageMagnitude
Normal
Kernel


Q-Q Plots of AverageMagnitude
Alaska
California
AverageMagnitude
Quantile

2. Suppose that at a local university the study guidelines for the College of Science and Math are to study two to three hours per unit per week. The instructor of the class, Orientation to the Statistics Major, takes these guidelines very seriously. He asks students to record their study time each week, and at the end of the term he compares their average study time per week to their term GPA. The SAS data set called STUDY_GPA contains student identification information, orientation course-section number, number of units enrolled, average time studied, and term GPA.

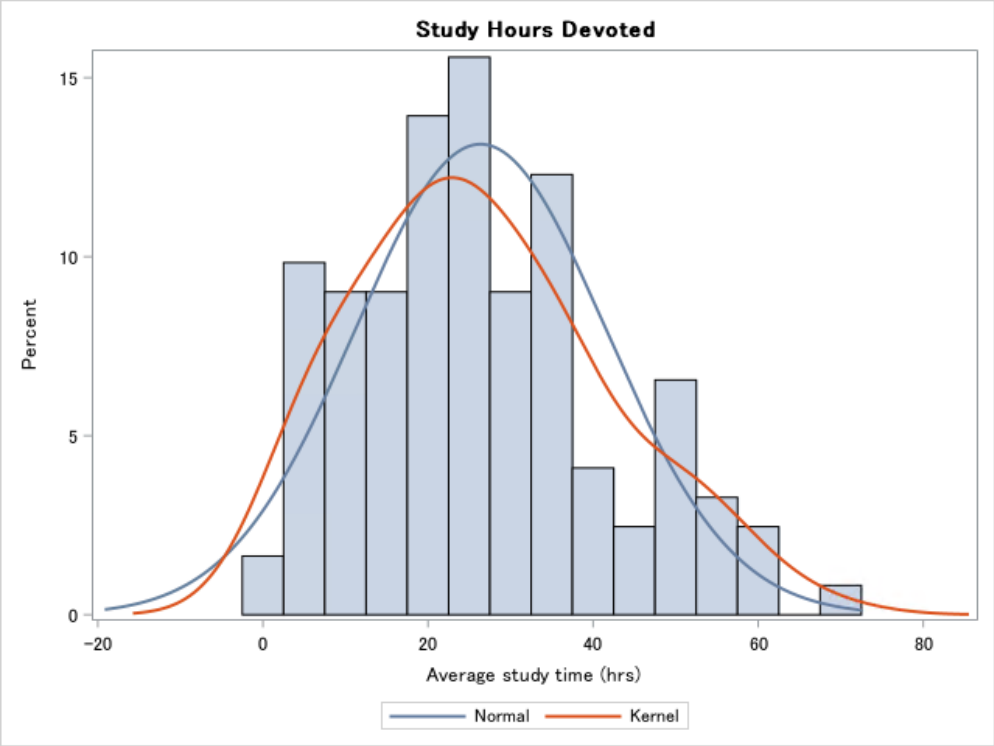
(a) Graph the histogram for hours of study. Use the start point=0 and bandwidth=5. Also, overlaid to this graph, display the plots for the kernel density and the best fitting normal curve. Using an eyeballing approach, can we say the hours of study follows a normal distribution?

Code:

```
/*part a*/
proc sgplot data = HWDATA.study_gpa;
histogram AveTime / binstart = 0 binwidth = 5;
density AveTime;
density AveTime / type = kernel;
title'Study Hours Devoted';
run;
```

Output:

Kernel density line shows that graph is almost near normal.



(b) Conduct a hypothesis test to check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 2. What is your conclusion? What variable you think may cause the other?

Code:

```
/*part b*/

ODS CSV FILE = 'E:\Users\nsm190002\Desktop\HWDATA\question2b.csv';

/*Write SQL query to filter selected entries for section '02'*/

PROC SQL;
select *
from HWDATA.study_gpa
WHERE Section eq '02';
quit;
ODS CSV CLOSE;

/*Use Import Wizard to create a program for importing saved questionsb.csv file*/
PROC IMPORT OUT= HWDATA.QUESTION2B1
            DATAFILE= "E:\Users\nsm190002\Desktop\HWDATA\question2b1.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

proc corr data=HWDATA.QUESTION2B1 plots=matrix(histogram);
var Units AveTime GPA;
title 'Correlation Units/AveTime/GPA';
run;
```

Output:

As we can see from the summary here:

1. For one unit increase in Units, there is 0.34 unit increase in Average Time
2. For One unit increase in Average Time, there is 0.10 units increase in GPA
3. For one unit increase in Units, there is 0.01 units decrease in GPA

The CORR Procedure

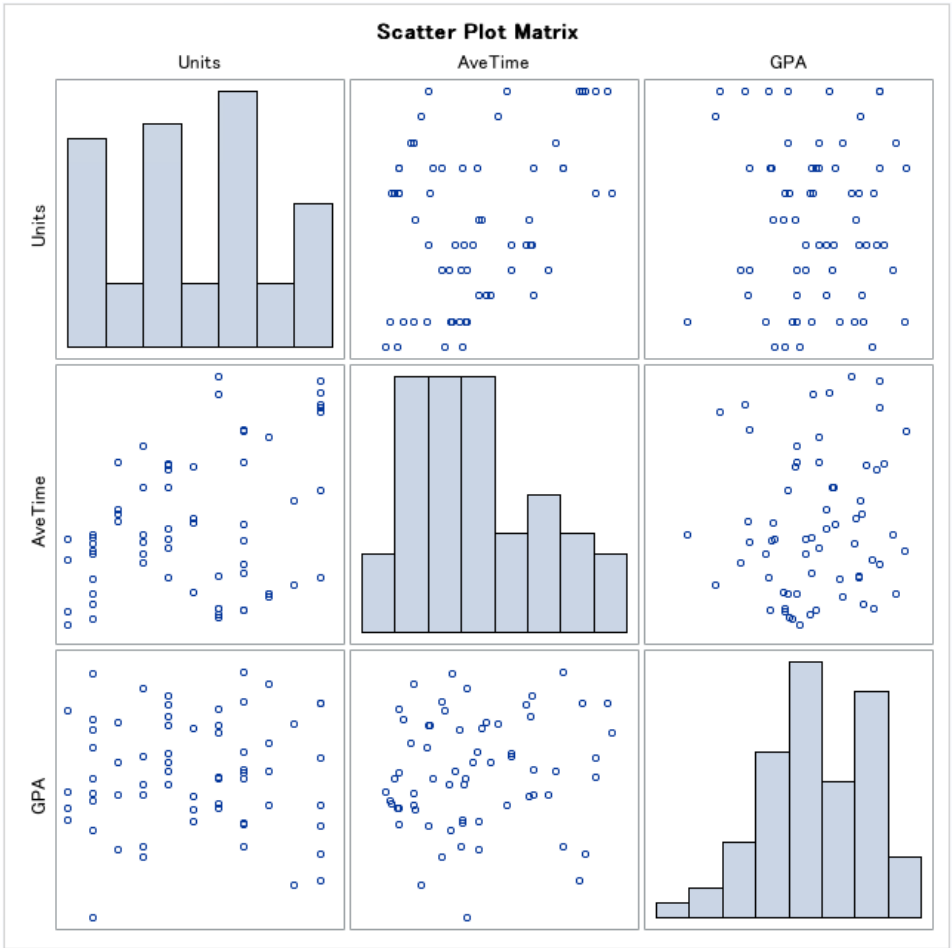
3 Variables:	Units AveTime GPA

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Units	64	13.87500	3.07834	888.00000	9.00000	19.00000
AveTime	64	23.35490	15.28746	1495	1.67731	56.33695
GPA	64	3.10125	0.44941	198.48000	1.93000	3.91000

Pearson Correlation Coefficients, N = 64
Prob > |r| under H0: Rho=0

	Units	AveTime	GPA
Units	1.00000	0.34493 0.0053	−0.01170 0.9269
AveTime	0.34493 0.0053	1.00000	0.10981 0.3877
GPA	−0.01170 0.9269	0.10981 0.3877	1.00000



(c) Assume there were no placebo group (i.e., treatment = 0) in your data set. Conduct a test to see whether there is a difference in plaque level before treatment and after the second visit? Interpret your results.

Code:

```
/*part c*/
/*Import file from part b*/
PROC IMPORT OUT= HWDATA.QUESTION3c1
            DATAFILE= "E:\Users\nsm190002\Desktop\HWDATA\question3c1.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

proc ttest data = HWDATA.Question3c1;
paired Visit0*Visit2;
title'Treatment Analysis 3c';
run;
```

Output:

For this experiment we only considered dataset with no control group (Placebo)

Hypothesis:

H0: V0(No treatment) – V2(After 2nd Visit in Treatment) = 0

H1: V0 – V2 NE 0

As we can see that P-value is less than 0.0001, which makes it significant enough to reject the Null.

We can safely say that there has been improved in patient's health after taking Vitamin E drug.

Treatment Analysis 3c

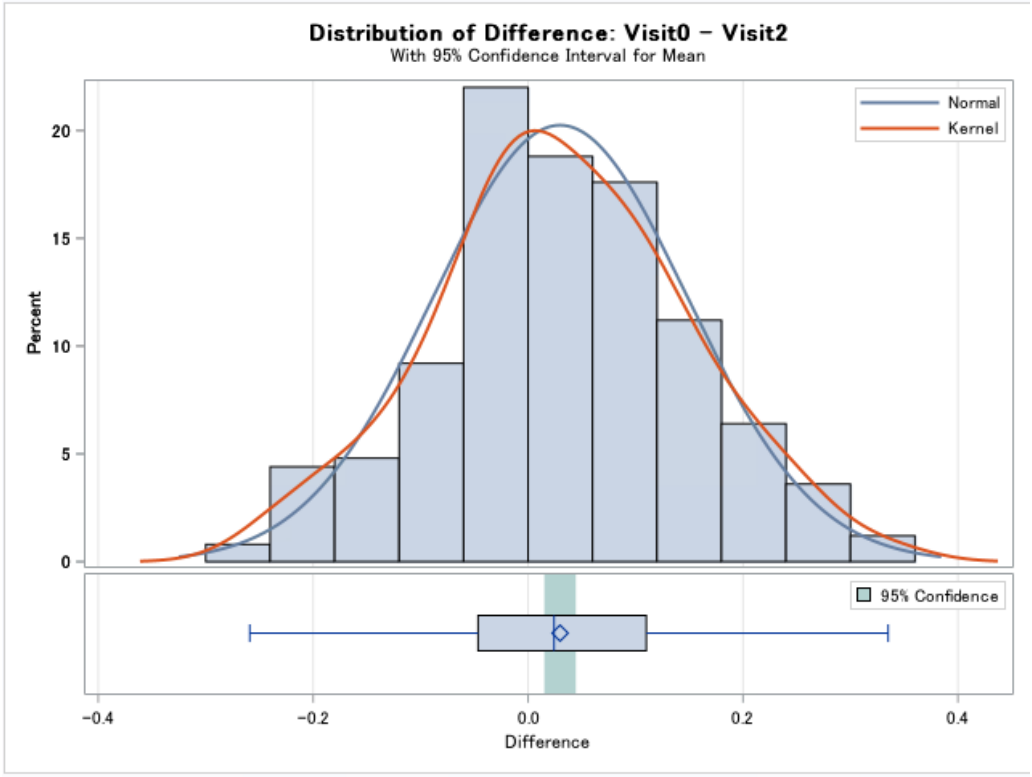
The TTEST Procedure

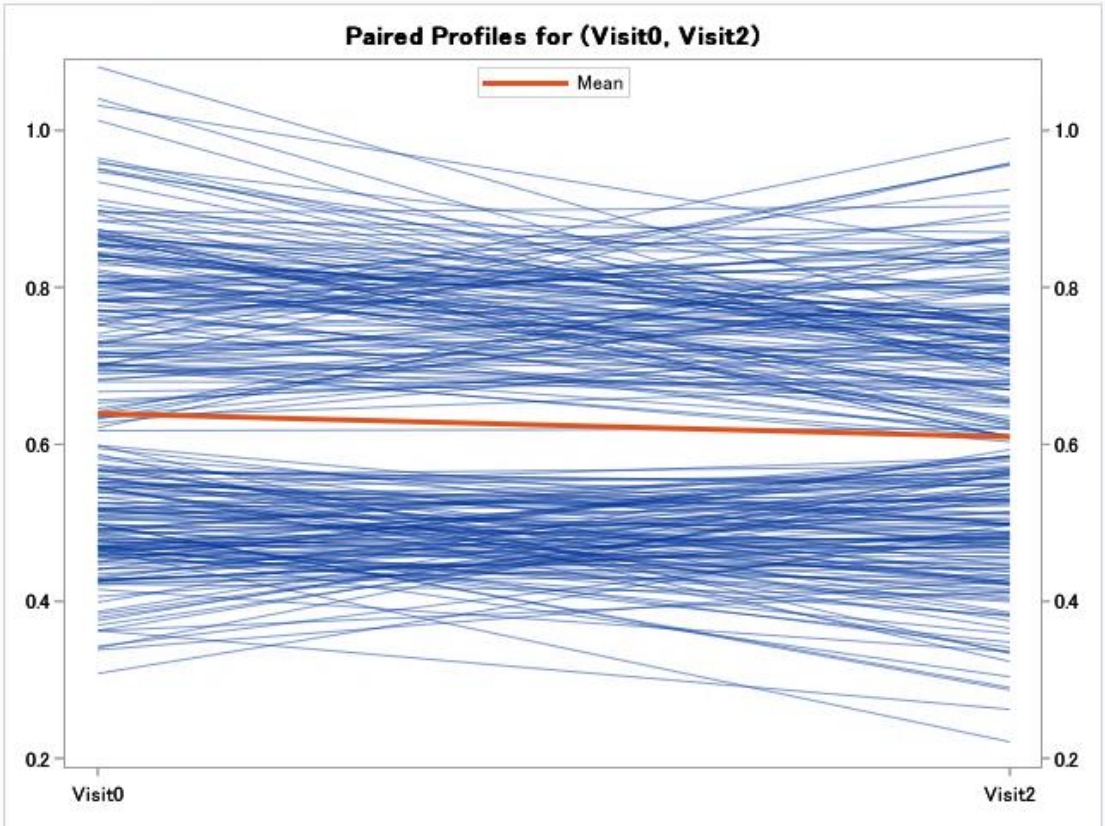
Difference: Visit0 – Visit2

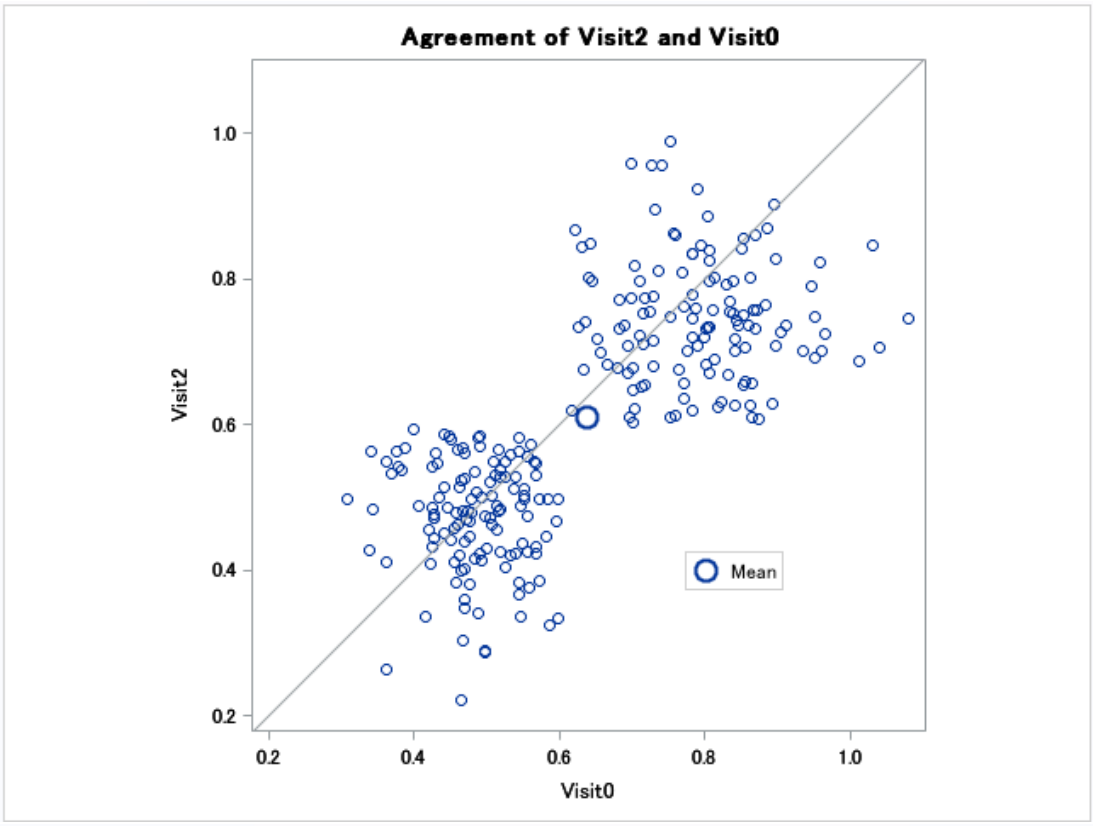
N	Mean	Std Dev	Std Err	Minimum	Maximum
250	0.0298	0.1182	0.00748	-0.2590	0.3351

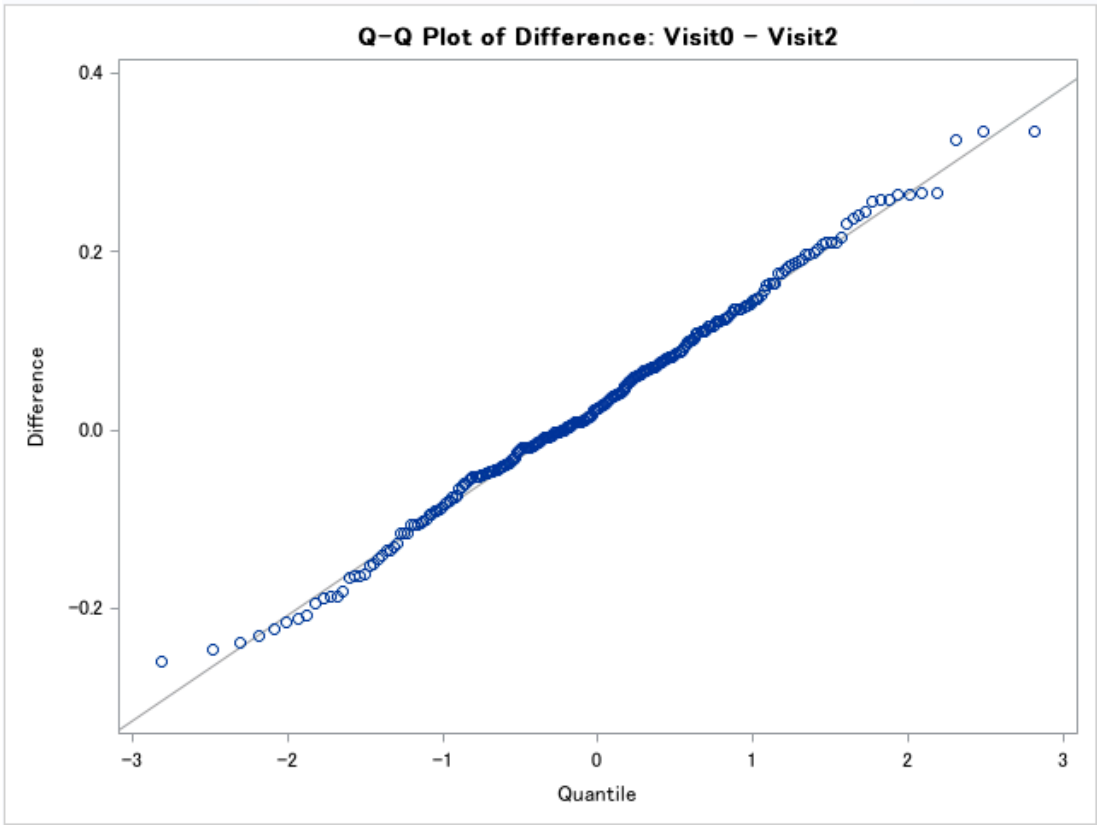
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.0298	0.0150	0.0445	0.1182	0.1087	0.1296

DF	t Value	Pr > |t|
249	3.98	<.0001


Distribution of Difference: Visit0 – Visit2
With 95% Confidence Interval for Mean
Normal
Kernel
Percent
20
15
10
5
0
95% Confidence
−0.4
−0.2
0.0
0.2
0.4
Difference


Paired Profiles for (Visit0, Visit2)
Mean
1.0
0.8
0.6
0.4
0.2
Visit0
Visit2


Agreement of Visit2 and Visit0
Visit2
Visit0
Mean
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0


Q-Q Plot of Difference: Visit0 - Visit2
Difference
Quantile
0.4
0.2
0.0
-0.2
-3
-2
-1
0
1
2
3

(d) Now, considering the fact that there is indeed a control group in your dataset, conduct a new test to check whether there is a difference in plaque level before treatment and after the second visit. Interpret your results.
(Hint: You need to make sure that the reduction in plaque level (if any) is indeed due to taking vitamin E. This is how having the control group in the study helps. In particular, you need to test whether the reduction in plaque level for treatment group is significantly less than that for control group)

Code:

```
/*part d*/
PROC IMPORT OUT= HWDATA.QUESTION3d
            DATAFILE= "E:\Users\nsm190002\Desktop\HWDATA\question3d.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
```

```
proc ttest data = HWDATA.Question3d;
paired Visit0*Visit2;
title'Treatment Analysis 3d';
run;
```

Output:

Hypothesis

H0: V0 – V2 = 0

H1: V0 – V2 NE 0

As we can see that P-value is 0.03 which is less than alpha = 0.05. So, we do not reject the Null and it means that there is no subsequent effect in change in health of people by taking Placebo Drug. This makes sense to check the effectiveness of the drug.

Treatment Analysis 3d

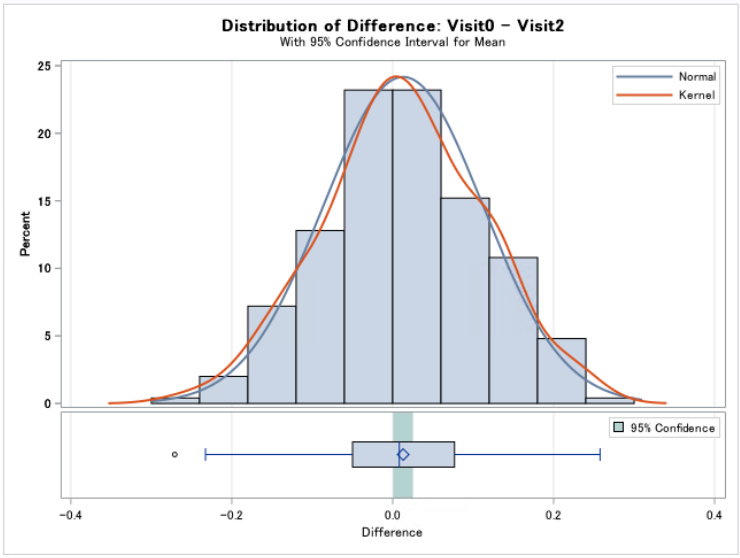
The TTEST Procedure

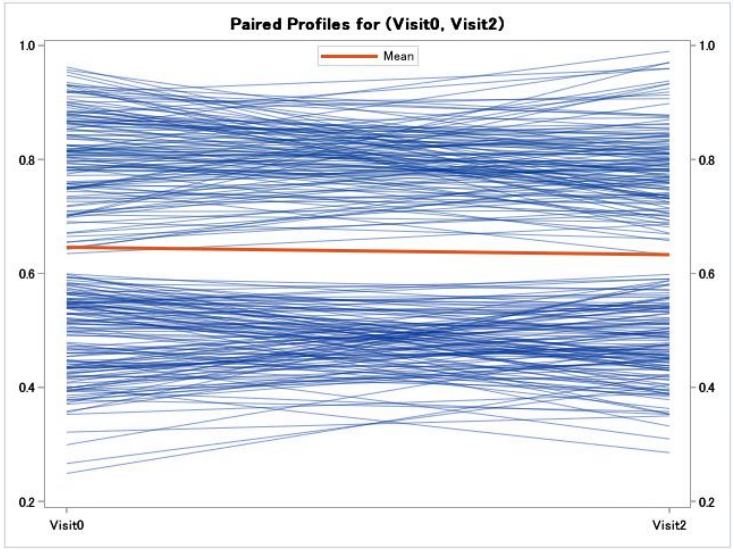
Difference: Visit0 – Visit2

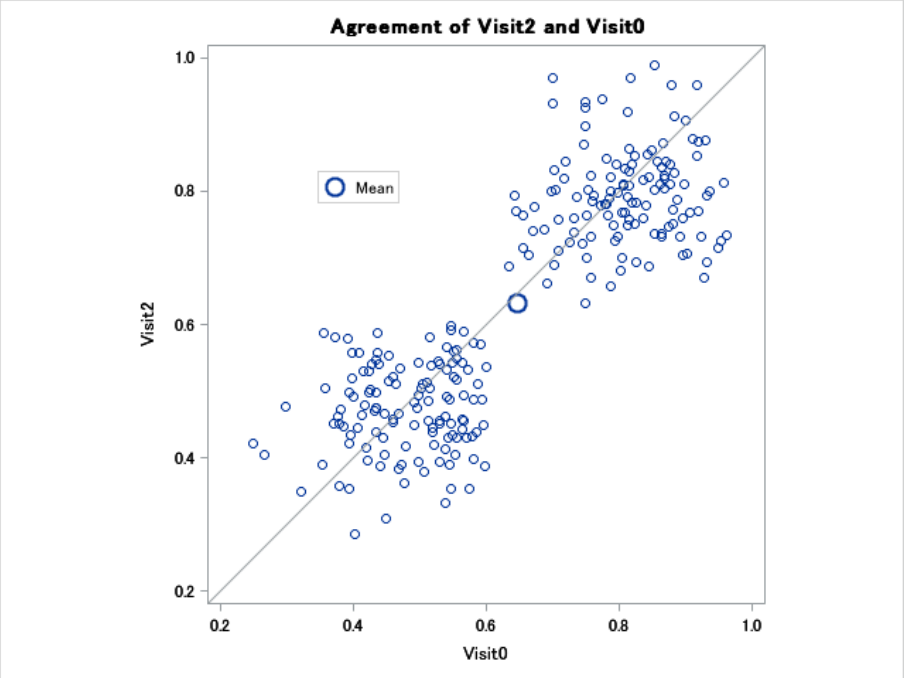
N	Mean	Std Dev	Std Err	Minimum	Maximum
250	0.0130	0.0990	0.00626	-0.2709	0.2577

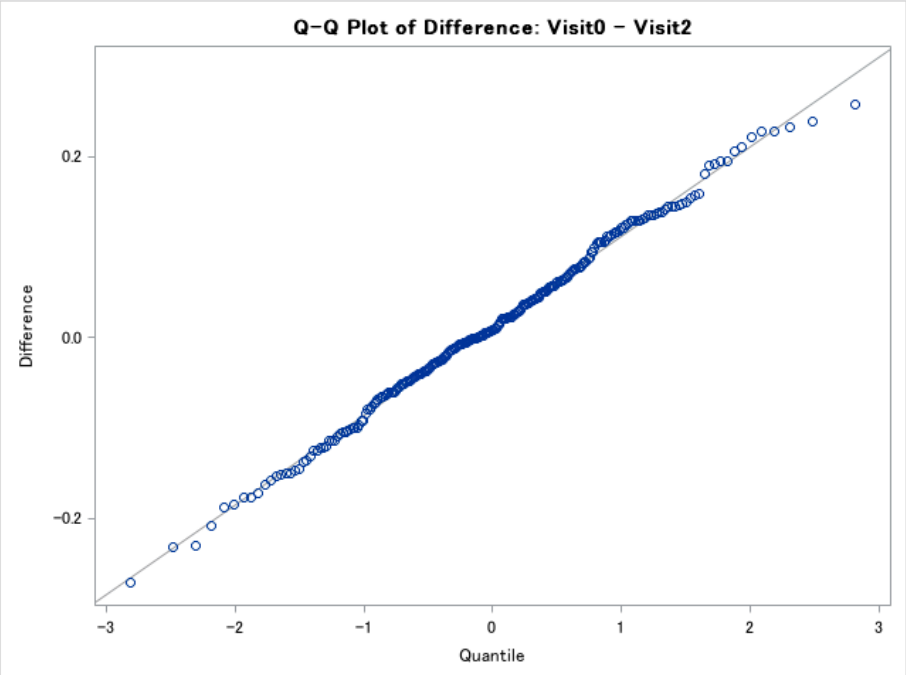
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0.0130	0.000628	0.0253	0.0990	0.0911	0.1086

DF	t Value	Pr > |t|
249	2.07	0.0395






Agreement of Visit2 and Visit0
Mean
Visit2
Visit0
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0


Q−Q Plot of Difference: Visit0 − Visit2
Difference
Quantile
0.2
0.0
−0.2
−3
−2
−1
0
1
2
3

(e) Which of the tests in part (c) and (d) is more reliable? Explain.
(Hint: This should be clear from the hint given above)

Answer: I think testing for T-Test in both the cases made sense to me since we wanted to check the effectiveness of the drug and indeed it did prove that by taking Vitamin E drug, the wellness of patient is improved.

(f) One of the critical factors in randomizing the subjects in control and treatment groups is to make sure that the subject are perfectly randomized in all aspects. Using the last two columns (i.e., alcohol and cigarette usage) of the original (long format) data, conduct two tests to check whether subjects are randomized perfectly. If they are perfectly randomized, then we should not expect much difference in alcohol (or cigarette) consumption for control vs. treatment groups.

Code:

```
/*print f*/
/*test 1*/
proc ttest data = HWDATA.Vite;
paired Alcohol*Smoke;
title'Paired T-Test Alcohol/Smoke';
run;
/*test 2*/
proc sgplot data = HWDATA.Vite;
hbox Alcohol /Category = Visit group=Treatment;
title'Alcohol Consumption';
RUN;

proc sgplot data = HWDATA.Vite;
hbox Smoke /Category = Visit group=Treatment;
title'Cigerrete Consumption';
RUN;
```

Output:

Paired T-Test Alcohol/Smoke

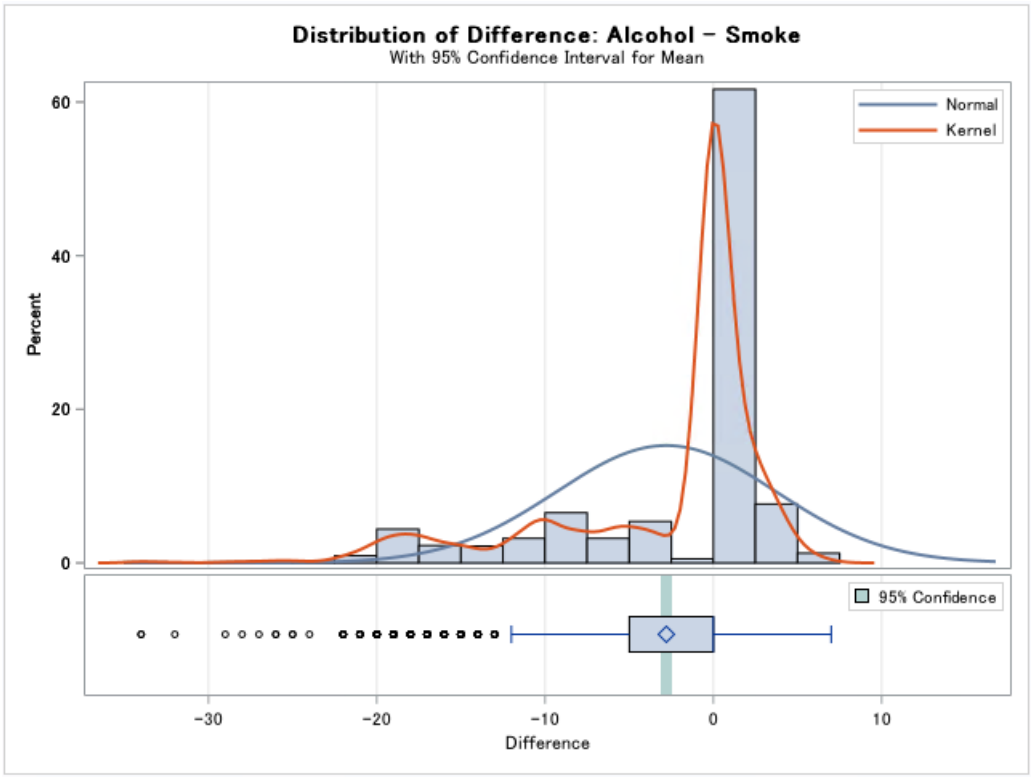
The TTEST Procedure

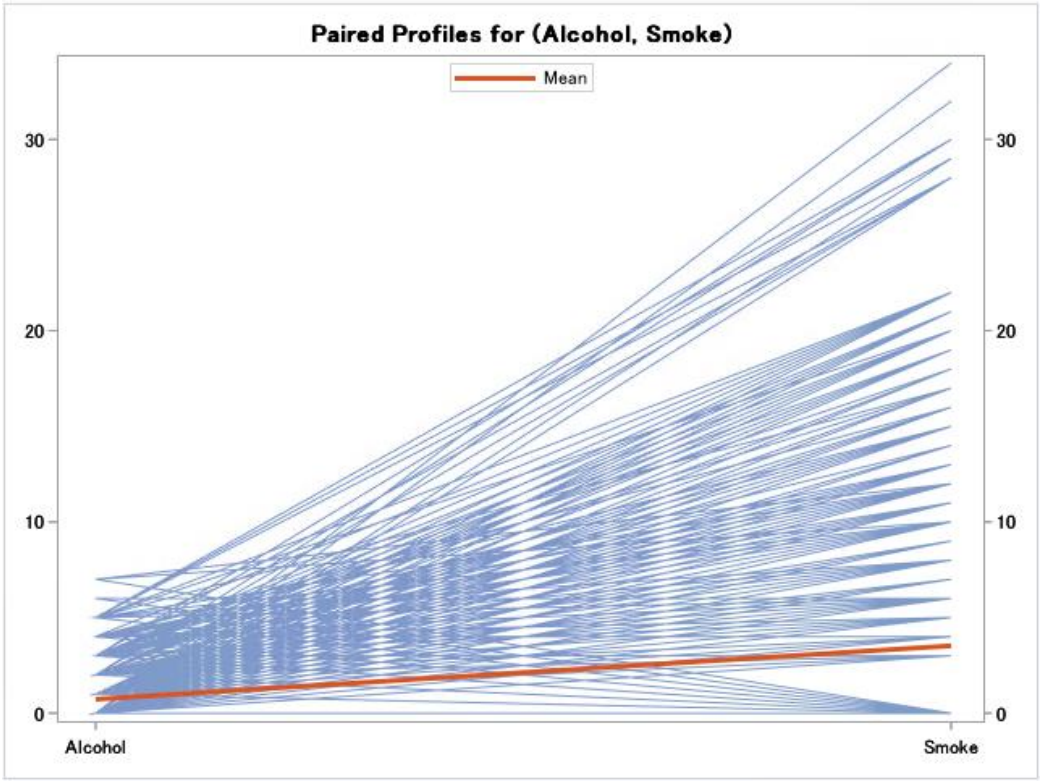
Difference: Alcohol - Smoke

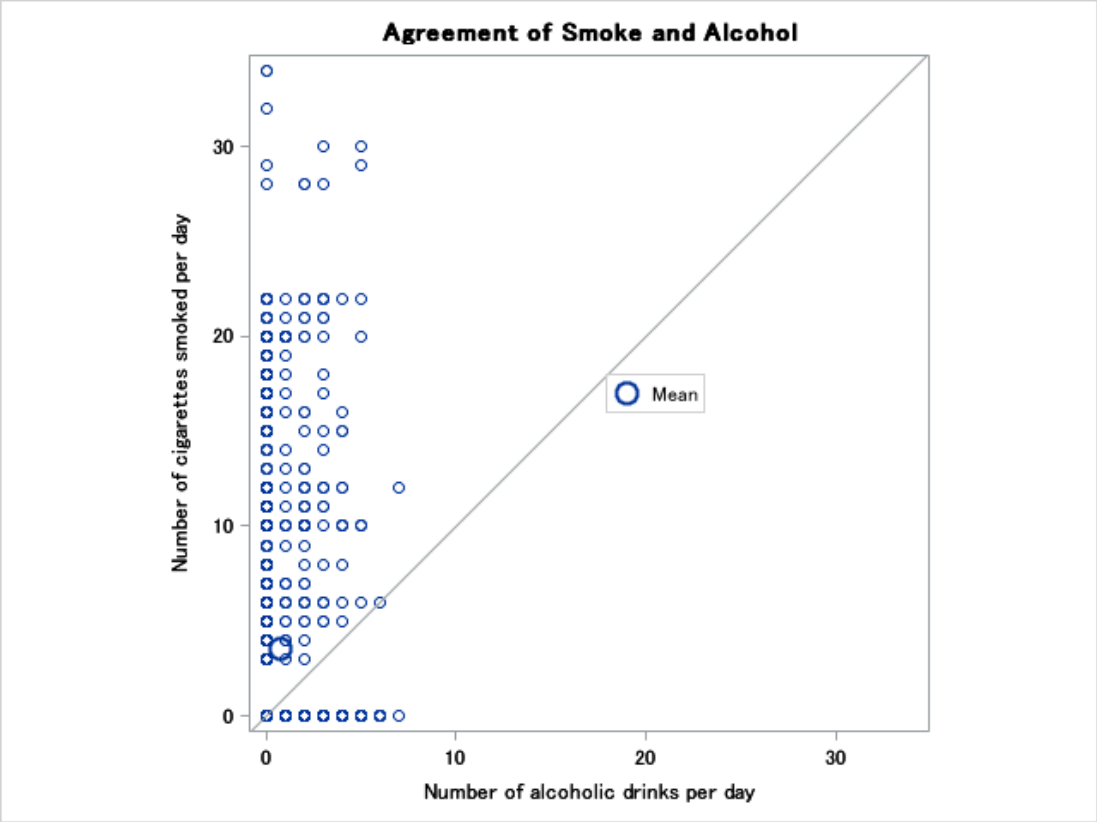
N	Mean	Std Dev	Std Err	Minimum	Maximum
1500	-2.7947	6.5249	0.1685	-34.0000	7.0000

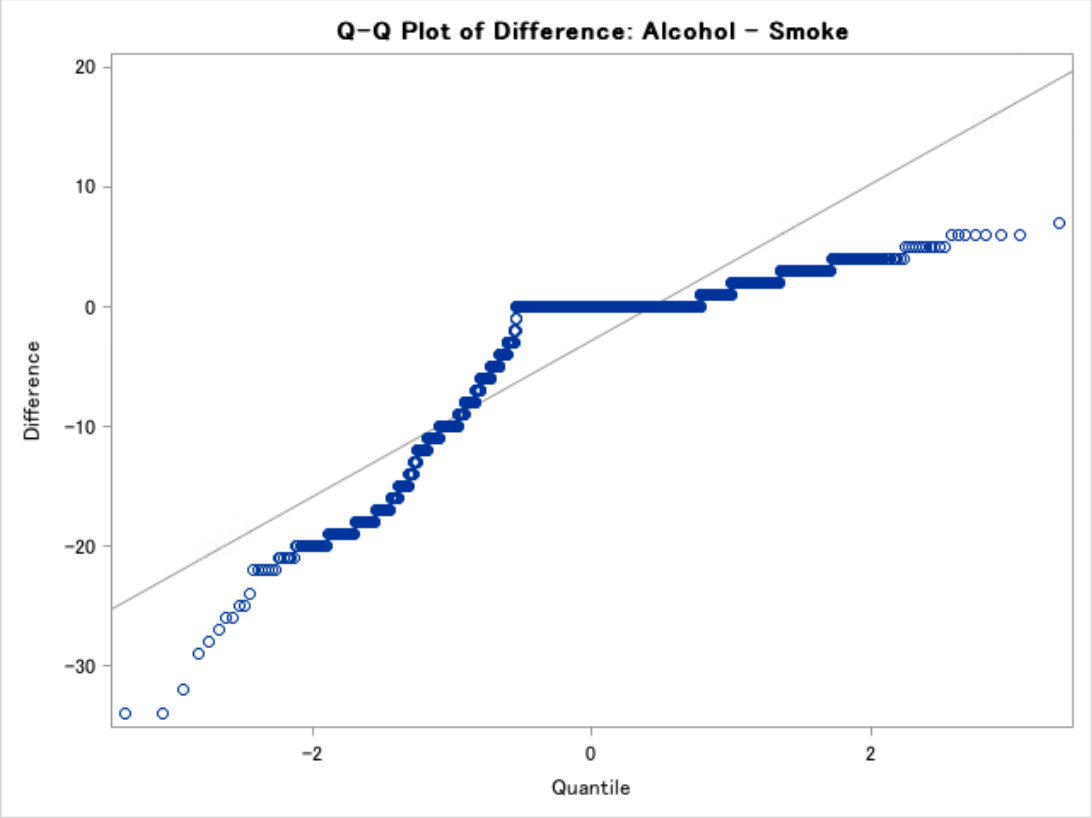
Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-2.7947	-3.1251	-2.4642	6.5249	6.2995	6.7672

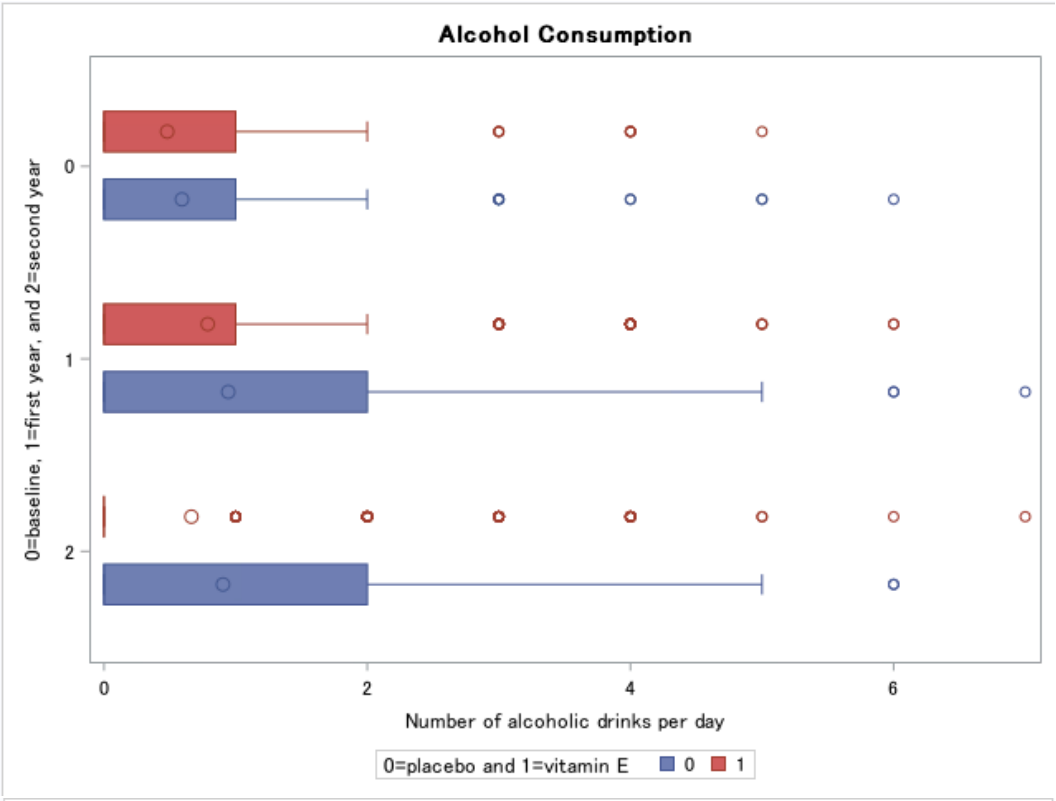
DF	t Value	Pr > |t|
1499	-16.59	<.0001

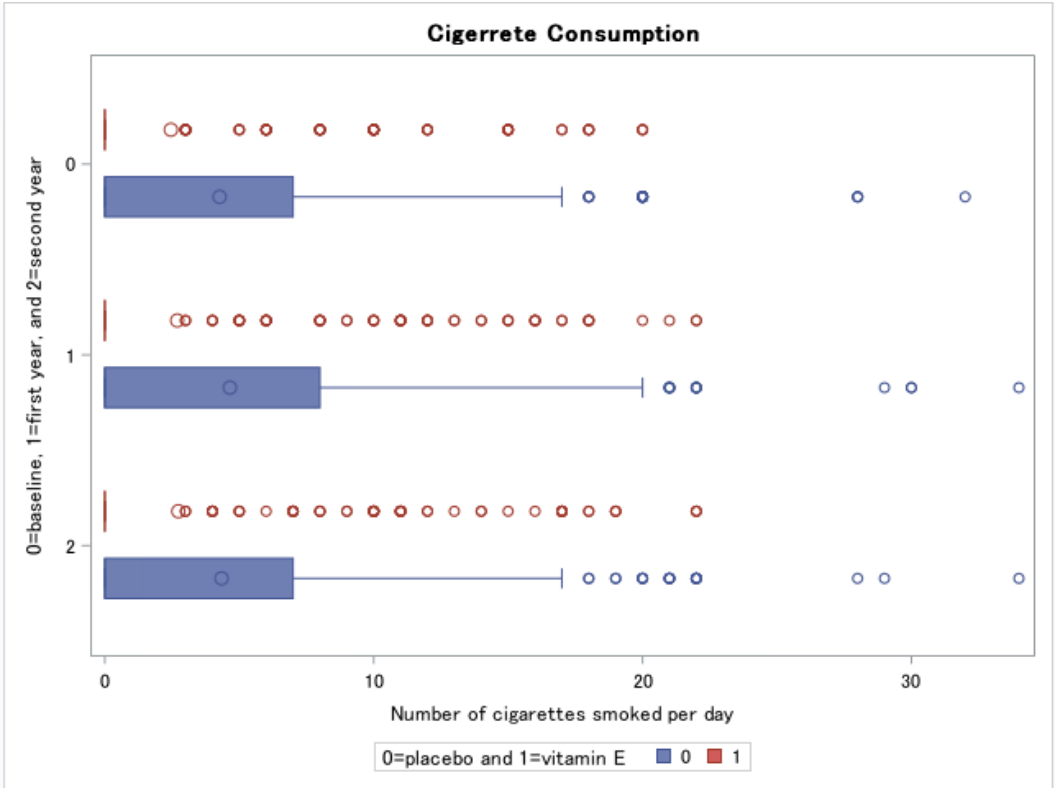

Distribution of Difference: Alcohol − Smoke
With 95% Confidence Interval for Mean
Normal
Kernel
95% Confidence
Percent
60
40
20
0
−30
−20
−10
0
10
Difference


Paired Profiles for (Alcohol, Smoke)
Mean
30
20
10
0
Alcohol
Smoke


Agreement of Smoke and Alcohol
Number of cigarettes smoked per day
Number of alcoholic drinks per day
Mean
0
10
20
30


Q-Q Plot of Difference: Alcohol - Smoke
Difference
Quantile
20
10
0
-10
-20
-30
-2
0
2


Alcohol Consumption
0=baseline, 1=first year, and 2=second year
0
1
2
0
2
4
6
Number of alcoholic drinks per day
0=placebo and 1=vitamin E
0
1


Cigerrete Consumption
0=baseline, 1=first year, and 2=second year
0
1
2
0
10
20
30
Number of cigarettes smoked per day
0=placebo and 1=vitamin E
0
1

Acknowledgements:

1. https://stats.idre.ucla.edu/sas/modules/how-to-reshape-data-long-to-wide-using-proc-transpose/
2. https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/257-30.pdf
3. https://documentation.sas.com/?docsetId=sqlproc&docsetTarget=n1e51p1t33ruyyn1iyzqonf3r0cm.htm&docsetVersion=9.4&locale=en#n149ioasupdfb0n1vs31jz24j38f
4. https://www.sas.com/content/dam/SAS/sv_se/doc/other1/Exporting-SAS-Data-Sets-and-Creating-ODS-reports-140219.pdf
5. https://documentation.sas.com/?docsetId=odsug&docsetTarget=p0oxrbinw6fjuwn1x23qam6dntyd.htm&docsetVersion=9.4&locale=en
6. https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Calgary-User-Group/Yankovsky-ExploringProcTtest-Apr2015.pdf
7. https://examples.yourdictionary.com/examples-of-control-groups.html
8. https://www.youtube.com/watch?v=FLQ_ittU1gc
9. https://www.quora.com/How-can-you-test-that-your-random-assignment-was-truly-random