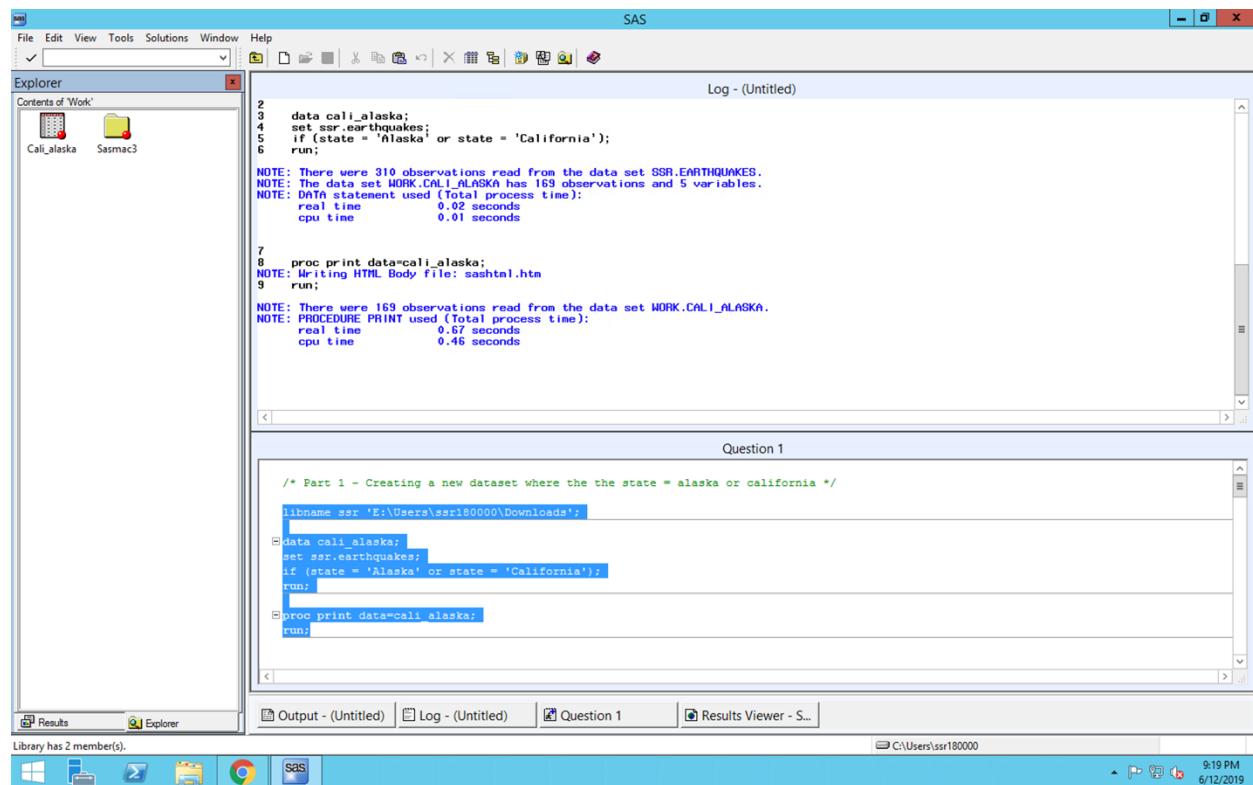


REPORT

TOPIC – HYPOTHESIS TESTING

Problem 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.



The screenshot shows the SAS software interface with the following details:

- Explorer:** Shows two datasets: Cali_alaska and Sasmac3.
- Log - (Untitled):** Displays the executed SAS code and its output.

```
2 data cali_alaska;
3 set ssr.earthquakes;
4 if (state = 'Alaska' or state = 'California');
5 run;

NOTE: There were 310 observations read from the data set SSR.EARTHQUAKES.
NOTE: The data set WORK.CALI_ALASKA has 169 observations and 5 variables.
NOTE: DATA statement used (Total process time):
      real time          0.02 seconds
      cpu time           0.01 seconds

7 proc print data=cali_alaska;
8 NOTE: Writing HTML Body file: sashml.htm
9 run;

NOTE: There were 169 observations read from the data set WORK.CALI_ALASKA.
NOTE: PROCEDURE PRINT used (Total process time):
      real time          0.67 seconds
      cpu time           0.46 seconds
```
- Question 1:** A code editor window containing the following SAS code:

```
/* Part 1 - Creating a new dataset where the state = alaska or california */
libname ssr 'E:\Users\ssr18000\Downloads';

data cali_alaska;
set ssr.earthquakes;
if (state = 'Alaska' or state = 'California');
run;

proc print data=cali_alaska;
run;
```
- Bottom Navigation:** Shows tabs for Output - (Untitled), Log - (Untitled), Question 1, and Results Viewer - S...; also displays the current working directory as C:\Users\ssr18000.
- System Tray:** Shows the date and time as 9:19 PM 6/12/2019.

SAS

File Edit View Tools Data Solutions Window Help

Explorer

Contents of 'Work'

Cal.alaska Sasmac3

VIEWTABLE: Work.Cali_alaska

	Year	Month	Day	State	Magnitude
1	1964	3	28	Alaska	9.2
2	1965	2	4	Alaska	8.7
3	1957	3	9	Alaska	8.6
4	1938	11	10	Alaska	8.2
5	1946	4	1	Alaska	8.1
6	1899	9	10	Alaska	8
7	2002	11	3	Alaska	7.9
8	1996	6	10	Alaska	7.9
9	1896	5	7	Alaska	7.9
10	1899	9	4	Alaska	7.9
11	1857	1	9	California	7.9
12	2003	11	17	Alaska	7.8
13	1987	11	30	Alaska	7.8
14	1929	3	7	Alaska	7.8
15	1906	4	18	California	7.8
16	1892	2	24	California	7.8
17	1988	3	6	Alaska	7.7
18	1958	7	10	Alaska	7.7
19	1900	10	9	Alaska	7.7
20	1975	2	2	Alaska	7.6
21	1972	7	30	Alaska	7.6
22	1979	2	28	Alaska	7.5
23	1943	11	3	Alaska	7.4
24	1872	3	26	California	7.4
25	1992	6	28	California	7.3
26	1965	3	30	Alaska	7.3
27	1958	4	7	Alaska	7.3
28	1952	7	21	California	7.3
29	1937	7	22	Alaska	7.3
30	1922	1	31	California	7.3
31	1904	8	27	Alaska	7.3
32	1873	11	23	California	7.3
33	2007	12	19	Alaska	7.2
34	2005	6	15	California	7.2
35	1952	4	25	California	7.2
36	1980	11	8	California	7.2
37	1947	10	16	Alaska	7.2

Results Log - (Untitled) Question 1 Results Viewer - S... VIEWTABLE: Wor...

NOTE: Table has been opened in browse mode.

C:\Users\ssr180000 9:20 PM 6/12/2019

SAS

File Edit View Tools Run Solutions Window Help

Results

Log - (Untitled)

NOTE: There were 169 observations read from the data set WORK.CALI_ALASKA.
 NOTE: PROCEDURE PRINT used (Total process time):
 real time 0.67 seconds
 cpu time 0.46 seconds

```
10 proc means data=cali_alaska mean stddev min p25 median p75 max maxdec= 2;
11 where year>=2002 and year<=2011;
12 class year state;
13 var magnitude;
14 title 'Summary Statistics of Magnitude';
15 run;
```

NOTE: There were 61 observations read from the data set WORK.CALI_ALASKA.
 WHERE (year>=2002 and year<=2011);
 NOTE: PROCEDURE MEANS used (Total process time):
 real time 0.07 seconds
 cpu time 0.07 seconds

Question 1 *

```
/* Part 2 - tablulating the summary statistics for magnitude of the earthquakes across different states within each year from
2002 to 2011*/
proc means data=cali_alaska mean stddev min p25 median p75 max maxdec= 2;
where year>=2002 and year<=2011;
class year state;
var magnitude;
title 'Summary Statistics of Magnitude';
run;
```

Output - (Untitled) Log - (Untitled) Question 1 * Results Viewer - S...

NOTE: 6 Lines Submitted.

C:\Users\ssr180000 Ln 19, Col 34 9:21 PM 6/12/2019

SAS

File Edit View Go Tools Solutions Window Help

Results

Print: The SAS System
Means: Summary Statistics of Magnitude

Results Viewer - SAS Output

Summary Statistics of Magnitude

The MEANS Procedure

Analysis Variable : Magnitude Magnitude

Year	State	N Obs	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
2002	Alaska	3	6.63	1.30	5.30	5.30	6.70	7.90	7.90
	California	6	4.52	0.64	3.60	3.90	4.70	4.90	5.30
2003	Alaska	4	7.10	0.51	6.60	6.75	7.00	7.45	7.80
	California	15	4.29	0.88	3.40	3.60	4.00	4.70	6.60
2004	Alaska	1	6.80	-	6.80	6.80	6.80	6.80	6.80
	California	2	4.50	2.12	3.00	3.00	4.50	6.00	6.00
2005	Alaska	1	6.80	-	6.80	6.80	6.80	6.80	6.80
	California	6	5.45	1.19	4.10	4.70	5.05	6.60	7.20
2006	Alaska	1	4.80	-	4.80	4.80	4.80	4.80	4.80
	California	1	4.50	-	4.50	4.50	4.50	4.50	4.50
2007	Alaska	4	6.70	0.36	6.40	6.45	6.60	6.95	7.20
	California	5	4.74	0.62	4.20	4.30	4.40	5.20	5.60
2008	Alaska	2	6.60	0.00	6.60	6.60	6.60	6.60	6.60
	California	2	5.45	0.07	5.40	5.40	5.45	5.50	5.50
2009	Alaska	1	5.80	-	5.80	5.80	5.80	5.80	5.80
	California	6	4.00	0.56	3.50	3.50	3.90	4.50	4.70
2010	California	1	6.50	-	6.50	6.50	6.50	6.50	6.50

Output - (Untitled) Log - (Untitled) Question 1 * Results Viewer - ...

C:\Users\ssr180000 9:22 PM 6/12/2019

SAS

File Edit View Tools Run Solutions Window Help

Results

Print: The SAS System
Means: Summary Statistics of Magnitude
Means: Summary Statistics

Log - (Untitled)

```

16 proc sort data=cali_alaska;
17 by year;
18 run;

NOTE: There were 169 observations read from the data set WORK.CALI_ALASKA.
NOTE: The data set WORK.CALI_ALASKA has 169 observations and 5 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time          0.00 seconds
      cpu time          0.00 seconds

19 proc means data=cali_alaska n mean stddev min p25 median p75 max maxdec= 2;
20 where year>=2002 and year<=2011;
21 by year;
22 class state;
23 var magnitude;
24 title 'Summary Statistics';
25 run;

NOTE: There were 61 observations read from the data set WORK.CALI_ALASKA.
      WHERE (year>=2002 and year<=2011);
NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.07 seconds
      cpu time          0.04 seconds

```

Question 1 *

```

/* Part 3 - Modify SAS code in (a) such that the results for each year are shown in a separate table */

proc sort data=cali_alaska;
by year;
run;

proc means data=cali_alaska n mean stddev min p25 median p75 max maxdec= 2;
where year>=2002 and year<=2011;
by year;
class state;
var magnitude;
title 'Summary Statistics';
run;

```

Output - (Untitled) Log - (Untitled) Question 1 * Results Viewer - sa... C:\Users\ssr180000 Ln 34, Col 14 9:23 PM 6/12/2019

NOTE: 7 Lines Submitted.

SAS

File Edit View Go Tools Solutions Window Help

Results

Print: The SAS System
Means: Summary Statistics of Magnitu
Means: Summary Statistics

Results Viewer - sashtml

California 1 1 4.50 . 4.50 4.50 4.50 4.50 4.50

Year=2007

Analysis Variable : Magnitude Magnitude

State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
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

Year=2008

Analysis Variable : Magnitude Magnitude

State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
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

Year=2009

Analysis Variable : Magnitude Magnitude

State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Alaska	1	1	5.80	. .	5.80	5.80	5.80	5.80	5.80
California	6	6	4.00	0.56	3.50	3.50	3.90	4.50	4.70

Year=2010

Analysis Variable : Magnitude Magnitude

State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
California	1	1	6.50	. .	6.50	6.50	6.50	6.50	6.50

Output - (Untitled) Log - (Untitled) Question 1 * Results Viewer - ...

C:\Users\ssr180000 9:23 PM 6/12/2019

SAS

File Edit View Tools Run Solutions Window Help

Results

Print: The SAS System
Means: Summary Statistics of Magnitu
Means: Summary Statistics
Tabulate: Summary Statistics of Magnit

Log - (Untitled)

```

24 title 'Summary Statistics';
25 run;
NOTE: There were 61 observations read from the data set WORK.CALI_ALASKA.
      WHERE (year>=2002 and year<=2011);
NOTE: EXECUTION MENTS used (Total process time):
      real time          0.07 seconds
      cpu time           0.04 seconds

26 proc tabulate data=cali_alaska;
27 where year>=2002 and year<=2011;
28 class year state;
29 var magnitude;
30 table year*(magnitude),state*(N Mean StdDev Min p25 Median p75 Max);
31 title 'Summary Statistics of Magnitude';
32 run;
NOTE: There were 61 observations read from the data set WORK.CALI_ALASKA.
      WHERE (year>=2002 and year<=2011);
NOTE: EXECUTION TABULATE used (Total process time):
      real time          0.10 seconds
      cpu time           0.09 seconds

```

Question 1 *

```

/* Part 4 - show the same results in part(a) but with the difference that years are shown in the first column and the states
   are shown in the top row. */

33 proc tabulate data=cali_alaska;
34 where year>=2002 and year<=2011;
35 class year state;
36 var magnitude;
37 table year*(magnitude),state*(N Mean StdDev Min p25 Median p75 Max);
38 title 'Summary Statistics of Magnitude';
39 run;

```

Output - (Untitled) Log - (Untitled) Question 1 * Results Viewer - sa...

C:\Users\ssr180000 Ln 44, Col 18 9:24 PM 6/12/2019

NOTE: 7 Lines Submitted.

SAS

File Edit View Go Tools Solutions Window Help

Results

Print: The SAS System
Means: Summary Statistics of Magnitude
Means: Summary Statistics
Tabulate: Summary Statistics of Magnitude

Results Viewer - sashtml

Analysis Variable : Magnitude Magnitude

State	N Obs	N	Mean	Std Dev	Minimum	25th Pctl	Median	75th Pctl	Maximum
Alaska	1	1	5.80	.	5.80	5.80	5.80	5.80	5.80
California	6	6	4.00	0.56	3.50	3.50	3.90	4.50	4.70

Year=2010

Summary Statistics of Magnitude

Year	Alaska						California										
	N	Mean	StdDev	Min	P25	Median	P75	Max	N	Mean	StdDev	Min	P25	Median	P75	Max	
	2002	Magnitude	3	6.63	1.30	5.30	5.30	6.70	7.90	7.90	6	4.52	0.64	3.60	3.90	4.70	4.90
2003	Magnitude	4	7.10	0.51	6.60	6.75	7.00	7.45	7.80	15	4.29	0.88	3.40	3.60	4.00	4.70	6.60
2004	Magnitude	1	6.80	.	6.80	6.80	6.80	6.80	6.80	2	4.50	2.12	3.00	3.00	4.50	6.00	6.00
2005	Magnitude	1	6.80	.	6.80	6.80	6.80	6.80	6.80	6	5.45	1.19	4.10	4.70	5.05	6.60	7.20
2006	Magnitude	1	4.80	.	4.80	4.80	4.80	4.80	4.80	1	4.50	.	4.50	4.50	4.50	4.50	4.50
2007	Magnitude	4	6.70	0.36	6.40	6.45	6.60	6.95	7.20	5	4.74	0.62	4.20	4.30	4.40	5.20	5.60
2008	Magnitude	2	6.60	0.00	6.60	6.60	6.60	6.60	6.60	2	5.45	0.07	5.40	5.40	5.45	5.50	5.50
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
2010	Magnitude	1	6.50	.	6.50	6.50	6.50	6.50	6.50

Output - (Untitled) Log - (Untitled) Question 1 * Results Viewer - ...

C:\Users\ssr180000 9:24 PM 6/12/2019

SAS

File Edit View Tools Solutions Window Help

Results

Print: The SAS System
Means: Summary Statistics of Magnitude
Means: Summary Statistics
Tabulate: Summary Statistics of Magnitude
SGPanel: Trend in magnitude over time

Log - (Untitled)

```

real time      0.01 seconds
cpu time      0.01 seconds

proc sgpanel data= means;
  panelby state;
  series x=year y=AvgMagnitude;
  title 'Trend in magnitude over time';
run;

NOTE: PROCEDURE SGpanel used (Total process time):
      real time      1.80 seconds
      cpu time      0.40 seconds

NOTE: There were 92 observations read from the data set WORK.MEANS.

```

Question 1 *

```

/* Part 5 - 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. */

proc sort data=cali_alaska;
  by year state;
run;

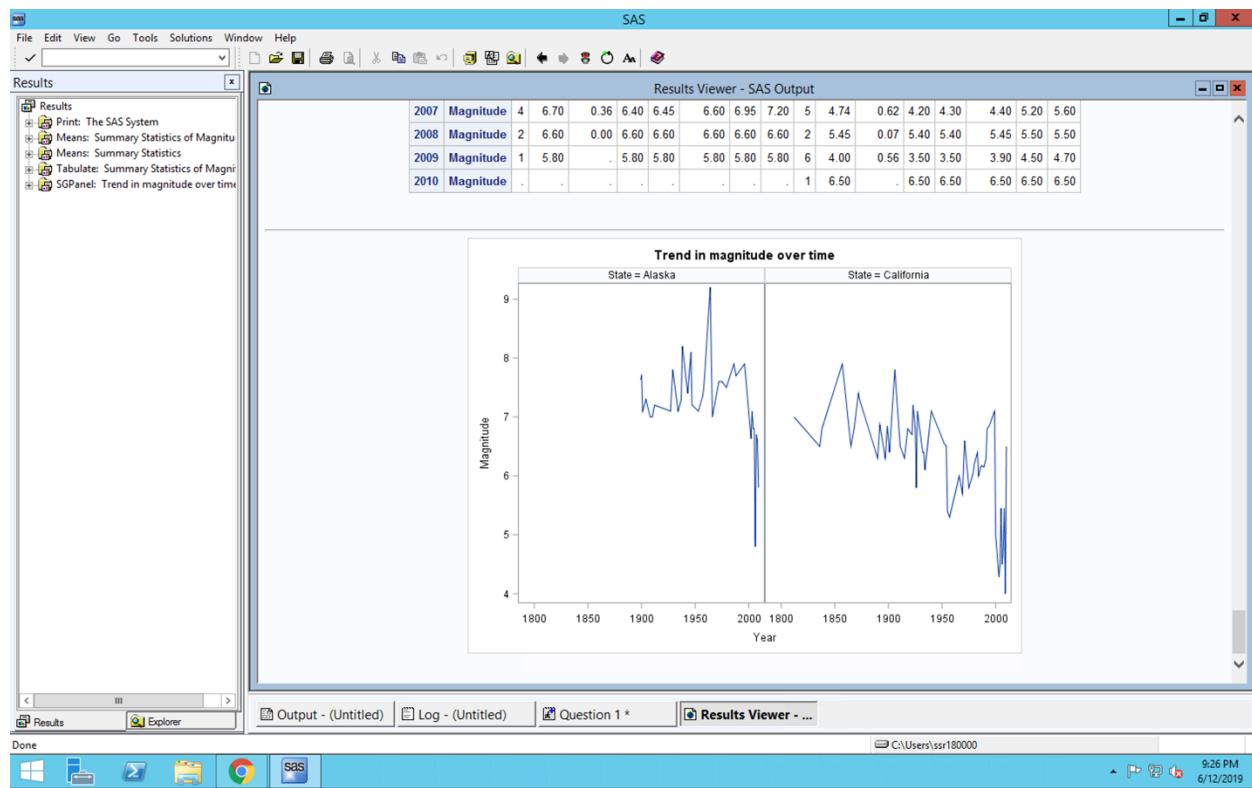
proc means data=cali_alaska mean maxdec= 2 noproctitle;
  by year state;
  var magnitude;
  output out=means
        mean= AvgMagnitude;
run;

proc sgpanel data= means;
  panelby state;
  series x=year y=AvgMagnitude;
  title 'Trend in magnitude over time';
run;

```

Output - (Untitled) Log - (Untitled) Question 1 * Results Viewer - S...

C:\Users\ssr180000 9:25 PM 6/12/2019



Is the average magnitude of earthquakes in California is significantly higher than that in Alaska?

The screenshot shows the SAS environment with multiple windows. The top window is the Log - (Untitled) window, which displays log messages and PROC TTEST and TTTEST statements. The bottom window is the Question 1 * window, which contains the following code and its execution results:

```

/* Part 6 - Test the hypothesis
H0: average magnitude of earthquakes in California <= average magnitude of earthquakes in Alaska
H1: average magnitude of earthquakes in California > average magnitude of earthquakes in Alaska
Test - 2 independent sample t-test
*/
proc ttest data=means;
  class state;
  var AvgMagnitude;
  title 'hypothesis testing';
run;

/*proc ttest data=means;
  class state;
  var AvgMagnitude;
  title 'hypothesis testing';
run;

/* Report the results:
When we look carefully at the hypothesis test of equality of variances for the population, it is clear that p-value>alpha ie 0.28>0.05
Hence we fail to reject the null implying presence of equal variances.
Now we check the pooled method for equal variances to observe that the p-value<alpha hence we reject the null.

Therefore we have found enough evidence to claim that average magnitude of earthquakes in California > average magnitude of
earthquakes in Alaska */

```

The null hypothesis is:

$$H_0: \mu_{California} \leq \mu_{Alaska}$$

Or equivalently

$$H_0: \mu_{Alaska} - \mu_{California} \geq 0$$

The screenshot shows the SAS Results Viewer window. The left pane displays a tree view of results, including 'Print: The SAS System', 'Means: Summary Statistics of Magnitude', 'Tabulate: Summary Statistics of Magnitude', 'SGPanel: Trend in magnitude over time', and 'Ttest: hypothesis testing'. The right pane shows the 'hypothesis testing' section of the 'The TTEST Procedure' for the variable 'AvgMagnitude (Magnitude)'. It includes three tables: 1) Descriptive statistics for Alaska and California, with a difference row. 2) Method selection table comparing Pooled and Satterthwaite methods. 3) Equality of Variances table for Folded F. Below these are two graphs: 'Distribution of AvgMagnitude' (a histogram) and 'Alaska' (a scatter plot). The bottom status bar shows the path 'Output - (Untitled) / Log - (Untitled) / Question 1 * / Results Viewer - ...' and the file path 'C:\Users\ssr180000'. The system tray at the bottom indicates the date and time as 6/12/2019 9:28 PM.

The p-value is very large, implying that we cannot reject the null. Hence, statistically, there is no significance evidence to believe that the average magnitude or earthquakes in California is larger.

Problem 2:

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

The screenshot shows the SAS software interface. The top menu bar includes File, Edit, View, Tools, Run, Solutions, Window, and Help. The left sidebar is titled 'Results' and lists various analysis types: Print, Means, Tabulate, SGPanel, Test, Contents, and SGPlot. The main area has two windows: 'Log - (Untitled)' and 'Question 2'. The 'Log' window displays the following SAS code and its execution results:

```

cpu time      0.04 seconds
61  proc sgplot data= study;
62    histogram AveTime / binstart = 0 binwidth = 5 ;
63    density AveTime / type = kernel;
64    density AveTime;
65    title 'Hours of study';
66  run;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time         0.52 seconds
      cpu time          0.07 seconds

NOTE: There were 122 observations read from the data set WORK.STUDY.

```

The 'Question 2' window contains the following SAS code:

```

libname ssr 'E:\Users\ssr180000\Downloads';

data study;
  set ssr.study_gpa;
run;

/*get to know the variable names used */
proc contents data=study;
run;

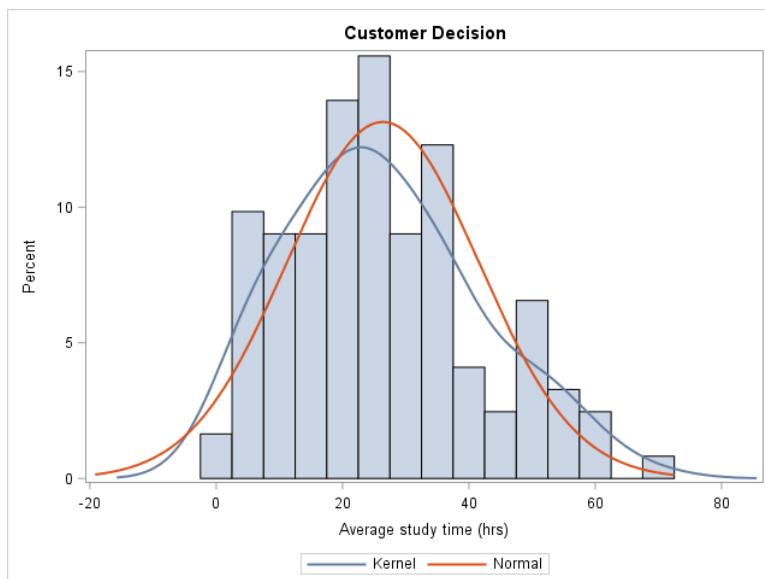
/* Part 1 - Graph the histogram for hours of study */

proc sgplot data= study;
  histogram AveTime / binstart = 0 binwidth = 5 ;
  density AveTime / type = kernel;
  density AveTime;
  title 'Hours of study';
run;

/* Yes it can be said that hours of study follows a normal distribution. */

```

At the bottom of the interface, there are tabs for 'Output - (Untitled)', 'Log - (Untitled)', 'Question 2', and 'Results Viewer - sa...'. The status bar at the bottom right shows 'Ln 12, Col 54', 'C:\Users\ssr180000', '9:40 PM', and '6/12/2019'.



Eyeballing: The kernel density is a bit skewed to the left, but not too much. We can't strongly argue that it looks like a normal distribution, however, normality assumption may not be a big assumption since the skewedness is not significant.

The screenshot shows the SAS software interface. The top menu bar includes File, Edit, View, Tools, Run, Solutions, Window, and Help. The main window has a 'Results' pane on the left containing a tree view of analysis results. The central area contains two panes: 'Log - (Untitled)' and 'Question 2 *'. The 'Log' pane displays SAS log output, including PROC SGPLOT and PROC UNIVARIATE statements. The 'Question 2 *' pane contains a SAS program snippet and a note about statistical tests for normality.

```

NOTE: PROCEDURE CONTENTS used (Total process time):
      real time         0.04 seconds
      cpu time          0.04 seconds

61  proc splot data= study;
62    histogram AveTime / binstart = 0 binwidth = 5 ;
63    density AveTime / type = kernel;
64    density AveTime;
65    title 'Hours of study';
66  run;

NOTE: PROCEDURE SGPLT used (Total process time):
      real time         0.52 seconds
      cpu time          0.07 seconds

NOTE: There were 122 observations read from the data set WORK.STUDY.

67  proc univariate data=study normal;
68  var AveTime;
69  run;

NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time         0.06 seconds
      cpu time          0.04 seconds

```

/* Part 2 - Conduct a statistical test to check whether the hours of study follows a normal distribution. */

proc univariate data=study normal;
var AveTime;
run;

/* We know that Skewness indicates how asymmetrical the distribution is (Whether it is more spread out on one side) while kurtosis indicates how flat or peaked the distribution is. The normal distribution has values of 0 for both skewness and kurtosis.

When we observe the results produced by the univariate statistical procedure, we can conclude that since the value of skewness and kurtosis is very close to zero, our estimation of the variable ' average hours of study ' is true to have an almost normal distribution.

The screenshot shows the SAS software interface with the 'Results' pane on the left. The central area is a 'Results Viewer - sashml' window titled 'Hours of study'. It displays the 'The UNIVARIATE Procedure' results for the variable 'AveTime' (Average study time (hrs)).

Moments

N	122	Sum Weights	122
Mean	26.3651016	Sum Observations	3216.54239
Std Deviation	15.1768129	Variance	230.335649
Skewness	0.48016013	Kurtosis	-0.307681
Uncorrected SS	112675.08	Corrected SS	27870.6135
Coeff Variation	57.564022	Std Error Mean	1.37404408

Basic Statistical Measures

Location	Variability		
Mean	26.36510	Std Deviation	15.17681
Median	24.82108	Variance	230.33565
Mode	.	Range	68.23397
		Interquartile Range	21.45978

Tests for Location: Mu0=0

Test	Statistic	p Value
Student's t	t 19.18796	Pr > t < .0001
Sign	M 61	Pr >= M < .0001
Signed Rank	S 3751.5	Pr >= S < .0001

Tests for Normality

Causality and Correlation.

Check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 1.

The screenshot shows the SAS software interface with three main windows:

- Log - (Untitled)**: Displays the log output of the SAS code run. It includes the following text:

```
real time      0.05 seconds
cpu time      0.04 seconds

130 proc corr data=study;
131   where Section='01';
132   var Units AveTime GPA;
133 run;

NOTE: PROCEDURE CORR used (Total process time):
      real time      0.04 seconds
      cpu time      0.04 seconds
```
- Question 2 ***: A text editor window containing the SAS code and a hypothesis test analysis.

```
/* Part 3 - (c) Conduct a hypothesis test to check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 1

H0: variables are not related
H1: variables are related
*/

proc corr data=study;
  where Section='01';
  var Units AveTime GPA;
run;

/* When we look closely at the correlation matrix generated, we can observe that the p-value for correlation between units and average time and that of average time and gpa is smaller than alpha. Hence statistically significant. Whereas the correlation between units and gpa has p-value>alpha hence not statistically significant.

According to me more the units taken by the student, more the number of hours he studies for them in the week. Thus the variable units is a good candidate for causing the variable AveTime.

*/
```
- Results**: A tree-view results browser window showing various analysis nodes like Print, Means, Tabulate, SGPanel, Test, Contents, SGPlot, Univariate, and Corr.

The taskbar at the bottom shows the Windows Start button, File Explorer, Google Chrome, and the SAS application icon. The system tray indicates the date as 6/12/2019 and the time as 10:22 PM.

The screenshot shows the SAS Results Viewer interface. On the left, the 'Results' pane lists various analysis nodes, including 'Print: The SAS System', 'Means: Summary Statistics of Magnitu...', and 'Corr: Hours of study'. The main pane displays the 'Results Viewer - SAS Output' window.

Hours of study
The CORR Procedure
3 Variables: Units AveTime GPA

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Units	58	13.79310	3.15538	800.00000	9.00000	19.00000	Number of units enrolled
AveTime	58	29.68670	14.46548	1722	0.77286	69.00683	Average study time (hrs)
GPA	58	3.30138	0.39409	191.48000	2.42000	3.94000	GPA

Pearson Correlation Coefficients, N = 58
Prob > |r| under H0: Rho=0

	Units	AveTime	GPA
Units	1.00000	0.42598	-0.15327
Number of units enrolled	0.0009	0.2507	
AveTime	0.42598	1.00000	-0.34324
Average study time (hrs)	0.0009	0.0083	
GPA	-0.15327	-0.34324	1.00000
GPA	0.2507	0.0083	

Below the viewer are tabs for 'Output - (Untitled)', 'Log - (Untitled)', 'Question 2 *', and 'Results Viewer - ...'. The taskbar at the bottom shows icons for File Explorer, Google Chrome, and the SAS application, along with system status information.

From this result, it can be seen that average study time and number of units registered are positively correlated. There is no other significant correlation. This correlation is very sensible: the more units you register, the more time you spend on studying.

Problem 3:

A study was conducted to see whether taking vitamin E daily would reduce the levels of atherosclerotic disease in a random sample of 500 individuals. Clinical measurements, including thickness of plaque of the carotid artery (taken via ultrasound), were recorded at baseline and at two subsequent visits in a SAS data set called VITE. Patients were divided into two strata according to their baseline plaque measurement.

Difference in plaque level before treatment and after the second visit?

H_0

: For treatment group, plaque level after the second visit is greater than plaque level at the beginning

For this part, we need to run a paired t test for treatment group to check whether taking vitamin E changes the plaque level. If we reject the null, we can conclude that taking vitamin E is effective in reducing the plaque level for treatment group.

Note that, in general, rejecting a one sided test is easier. However, when using a one sided test we should have a reasonable prior assumption. In this case, for example, it is believed that taking vitamin E should reduce the plaque level, hence, a one sided test is reasonable. If there is no such reasoning, we have to use a two sided test.

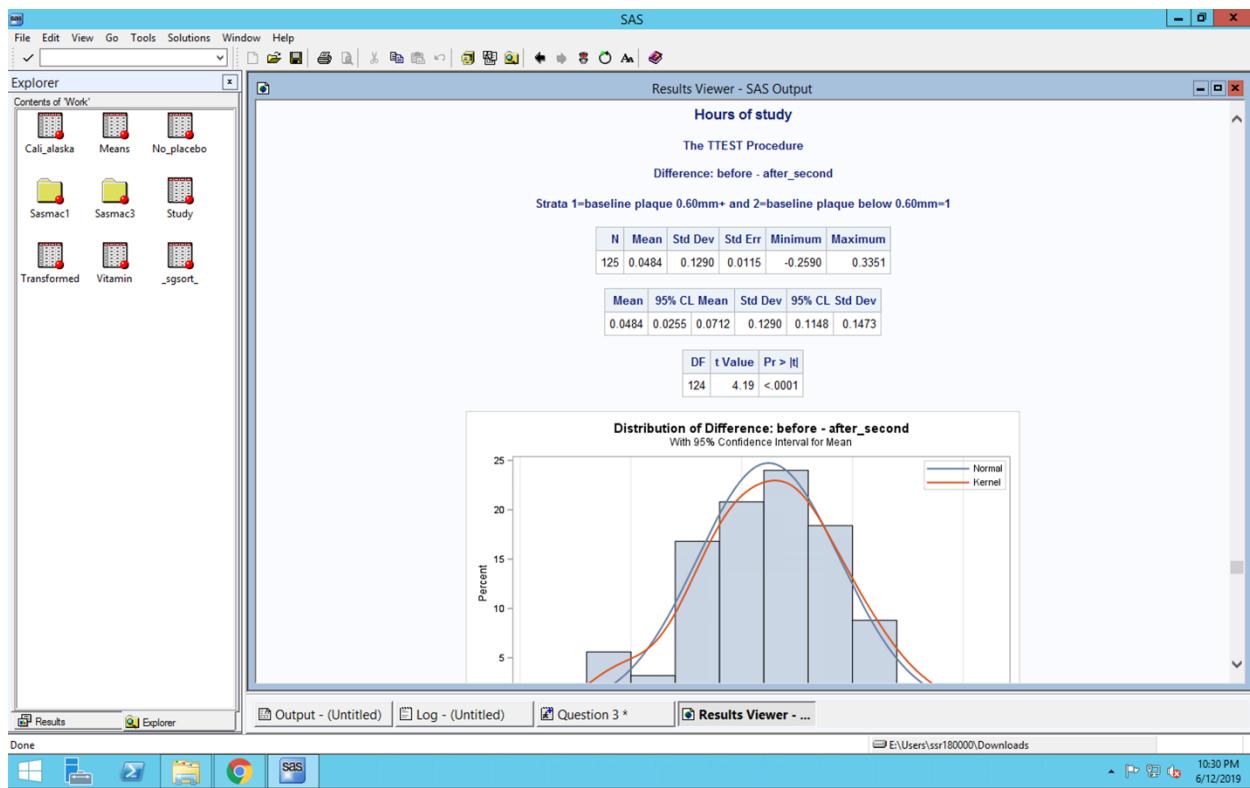
The screenshot shows the SAS software interface. The top menu bar includes File, Edit, View, Tools, Solutions, Window, and Help. The left pane is the Explorer, showing a folder structure under 'Contents of 'Work'' with items like Cali_alaska, Means, No_placebo, Sasmac1, Sasmac3, Study, Transformed, Vitamin, and _sgsort_. The right pane has three tabs: Log - (Untitled), Question 3 *, and Results Viewer - S... The Log tab displays PROCEDURE TTEST output. The Question 3 * tab contains the following SAS code:

```
libname ssr 'E:\Users\ssr18000\Downloads';
data vitamin;
set ssr.vite;
run;
proc contents data=vitamin;
run;
/* Part 1 - Assuming there were no placebo (i.e., control) group, How would you test whether there is a difference in plaque level before treatment and after the second visit? */

data no_placebo;
set ssr.vite(where=(Treatment=1));
run;
proc transpose data=no_placebo out=transformed(drop=_: col2 rename=(col1=before col3=after_second));
by ID Strata;
var Plaque;
run;
/* Type of test - paired sample t-test since the observations are drawn from a single sample and we will be considering the before and after effects for each person.
H0: Plaque measurement before treatment <= Plaque measurement after second visit
H1: Plaque measurement before treatment > Plaque measurement after second visit */

proc ttest data=transformed alpha=0.05;
paired before*after_second;
by Strata;
run;
/* From the ttest, we observe that the average difference is 0.0484 and 0.0112 when strata was 1 and 2 respectively which indicates the presence of a positive difference meaning that the plaque was greater before the treatment was started.
Also p-value<0.05 we can conclude that there is a difference in plaque level. */
```

The bottom status bar shows the path E:\Users\ssr18000\Downloads, the date 6/12/2019, and the time 10:29 PM.



The null is rejected. So, taking vitamin E has a significant effect on reducing the plaque level for treatment group after the second visit.

Now, considering the fact that there is indeed a control group in your dataset, change in the hypothesis test will be:

In many application, researchers divide individuals in control and treatment groups. Having the control group will insure that the effect is due to the factors we are interested in. From the results in part (a), one may conclude that reduction in plaque level is due to taking vitamin E. However, the reduction in plaque level may be due to some other factors which are not observed by researcher. Here is how the control group will help. We conduct a hypothesis test for control group as well, then two things might happen:

- 1- The plaque level for control group is the same before and after treatment. This implies that the reduction in plaque level for treatment group is indeed due to taking vitamin E. The reason is that if this reduction was due to some factors other than vitamin E, the plaque level for control group must have changed to.
- 2- The plaque level for control group is also changing. Here we cannot make an immediate conclusion. We need to know the amount of change. If the reduction in plaque level for treatment group is much more than that for control group, we may still conclude that taking vitamin E is effective. This method is known as Difference in Difference approach (Diff in Diff).

In summary, the null hypothesis that we need to conduct is:

H_0 : The reduction in plaque level for treatment group is less than that for control group

If we reject the hypothesis above, we can conclude that the reduction in plaque level is significantly greater for treatment group compared with control group, implying that taking vitamin E is significantly effective in reducing the plaque level.

The screenshot shows the SAS interface with two main windows. The top window is the 'Log - (Untitled)' window, which displays the following SAS code and its execution results:

```

152 set ssr.vite;
153 run;
NOTE: There were 1500 observations read from the data set SSR.VITE.
NOTE: The data set WORK.CONTROL_TREATMENT has 1500 observations and 12 variables.
NOTE: DATA statement used (Total process time):
      real time       0.01 seconds
      cpu time        0.01 seconds

154 proc ttest data=control_treatment alpha=0.05;
   class Treatment;
   var Plaque;
   by Strata;
run;

NOTE: PROCEDURE TTEST used (Total process time):
      real time       3.97 seconds
      cpu time        2.09 seconds

```

The bottom window is a code editor titled 'Question 3 *', containing the following SAS code and comments:

```

/* Part 2 - now that there exists a control group, how will you change your hypothesis test in part (a)
   Type of test - independent 2 sample t-test since we now have 2 different samples, one with a placebo and the rest were given
   Vitamin e treatment. Subjects in each of these groups were then recorded as per the visits.
   H0: Plaque measurement for placebo treatment <= Plaque measurement for vitamin e
   H1: Plaque measurement for placebo treatment > Plaque measurement for vitamin e */

data control_treatment;
set ssr.vite;
run;
proc ttest data=control_treatment alpha=0.05;
class Treatment;
var Plaque;
by Strata;
run;
/* its important to understand the variance of the population, we can observe that in the table for equal variances, p>alpha hence
we fail to reject the null ie they have equal variances. in the table just above que check their values for pooled method and notice
that p<alpha. thus we reject the null and can conclude Plaque measurement for placebo treatment > Plaque measurement for vitamin e */

```

The screenshot shows the SAS interface with the 'Results Viewer - SAS Output' window open. The window displays the output of the TTEST procedure for the variable 'Plaque' (Plaque measurement (mm)).

The TTEST Procedure

Variable: Plaque (Plaque measurement (mm))

Strata 1=baseline plaque 0.60mm+ and 2=baseline plaque below 0.60mm-

Treatment	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	375	0.7988	0.0806	0.00416	0.6003	0.9897
1	375	0.7743	0.0888	0.00459	0.6007	1.0808
Diff (1-2)	0.0245	0.0848	0.00619			

Treatment	Method	Mean	95% CL Mean	Std Dev	95% CL Std Dev		
0		0.7988	0.7906	0.8070	0.0806	0.0752	0.0868
1		0.7743	0.7653	0.7833	0.0888	0.0829	0.0957
Diff (1-2)	Pooled	0.0245	0.0123	0.0367	0.0848	0.0807	0.0893
Diff (1-2)	Satterthwaite	0.0245	0.0123	0.0367			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	748	3.96	<.0001
Satterthwaite	Unequal	741.03	3.96	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	374	374	1.21	0.0604

The screenshot shows the SAS software interface. The top menu bar includes File, Edit, View, Tools, Solutions, Window, and Help. The left sidebar has a 'Results' tree view with various nodes like Print: The SAS System, Means: Summary Statistics of Magnitude, and Test: hypothesis testing. The main window contains three panes:

- Log - (Untitled)**: Displays SAS log output. It shows the execution of PROC TTEST on the 'control_treatment' dataset with alpha=0.05, comparing 'Treatment' by 'Plaque' across 'Strata'. It notes 1500 observations and a total process time of 3.97 seconds.
- Question 3 ***: A text editor pane containing a question and its answer. The question asks about randomizing subjects. The answer explains that subjects were initially grouped by plaque measurement (0.60+ or 0.6-) and then further divided by treatment (vitamin E). It emphasizes that randomizing subjects between control and treatment groups is crucial for validity.
- Output - (Untitled)**: Shows the results of the TTEST procedure, including the test statistics and p-value.

The bottom status bar shows the file path E:\Users\ssr18000\Downloads, the date 6/12/2019, and the time 10:34 PM.

```

152 set ssr.vite;
153 run;

NOTE: There were 1500 observations read from the data set SSR.VITE.
NOTE: The data set WORK.CONTROL_TREATMENT has 1500 observations and 12 variables.
NOTE: DATA statement used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds

154 proc ttest data=control_treatment alpha=0.05;
155   class Treatment;
156   var Plaque;
157   by Strata;
158 run;

NOTE: PROCEDURE TTEST used (Total process time):
      real time          3.97 seconds
      cpu time           2.99 seconds

```

One of the critical factors in randomizing the subjects in control and treatment groups is to make sure that the subjects are perfectly randomized in all aspects. Using the last two columns (i.e., alcohol and cigarette usage) conduct two hypothesis tests to check whether subjects are randomized perfectly.

Following the arguments above, we need to make sure that the only difference between treatment and control group is taking/not taking vitamin E and they are similar in every other aspect. This implies that randomizing people in two groups has to be done very carefully. A bad randomizing, for example, is a case where a high portion of individuals in treatment group drink alcohol regularly, whereas, no one in control group drinks alcohol. This will ruin the results. In our example, we need to run the following two tests.

H₀: The average number of cigarettes smoked per day per individual is the same for both groups.

H₀: The average amount of alcohol drunk per day per individual is the same for both groups.

If we reject any of these tests, it means the randomization is not perfect. I don't put the results here. These two tests should be very straightforward to do. You can find the SAS codes posted. So, I'll leave it as your own exercise.

Remark: Ideally we want also the portions to be the same. In other words, we need to also make sure that: **The portion of people who smoke cigarettes/drink alcohol is the same for both groups.**

For the purpose of this assignment, doing only the first test is accepted.

SAS

File Edit View Tools Run Solutions Window Help

Results

Log - (Untitled)

```

NOTE: PROCEDURE MEANS used (Total process time):
      real time          0.01 seconds
      cpu time           0.01 seconds

170 proc ttest data=average_smoke alpha=0.05;
171   class Treatment;
172   var Smoke;
173   run;

NOTE: PROCEDURE TTEST used (Total process time):
      real time          1.04 seconds
      cpu time           1.01 seconds

```

Question 3 *

```

/* Part 4 - type of test - 2 sample ttest
H0: the average number of cigarettes smoked per day per individual is not equal for two groups
H1: the average number of cigarettes smoked per day per individual is equal for two groups */

proc sort data=control_treatment;
by Treatment;
run;

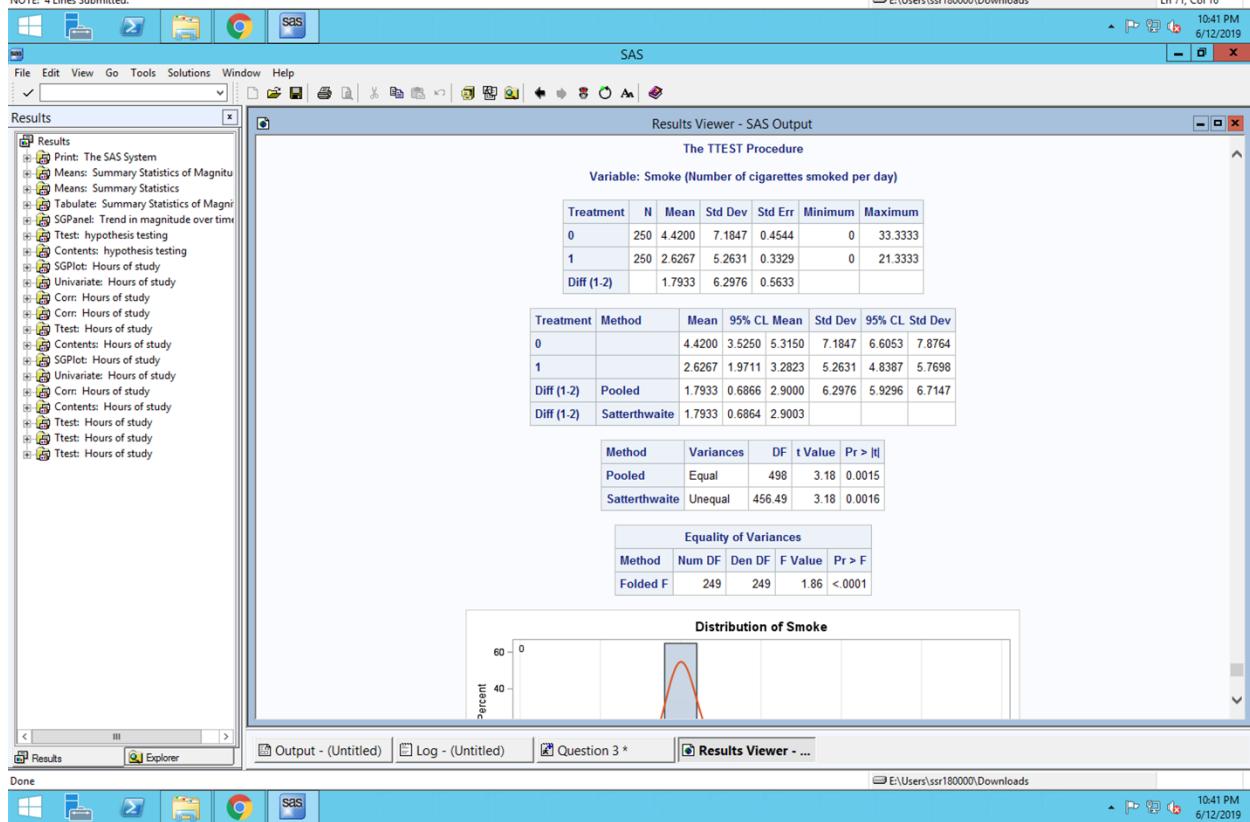
proc means data=control_treatment nway noprint;
var Smoke;
class ID;
by Treatment;
output out=average_smoke(drop=_:) mean=;
run;

proc ttest data=average_smoke alpha=0.05;
class Treatment;
var Smoke;
run;

/* We can observe that in the table for equal variances, palpha hence we reject the null ie consider unequal variances.
In the table just above we check the values for Satterthwaite method and notice that palpha. Thus we reject the null and
can conclude the average number of cigarettes smoked per day per individual is equal for two groups and thus the subjects are
randomized perfectly wrt smoke */

```

Output - (Untitled) Log - (Untitled) Question 3 * Results Viewer - S... E:\Users\ssr18000\Downloads Ln 71, Col 16 10:41 PM 6/12/2019



SAS

File Edit View Tools Solutions Window Help

Results

Log - (Untitled)

```

NOTE: PROCEDURE MEANS used (Total process time):
      real time       0.02 seconds
      cpu time        0.03 seconds

182 proc ttest data=average_alcohol alpha=0.05;
183   class Treatment;
184   var Alcohol;
185 run;

NOTE: PROCEDURE TTEST used (Total process time):
      real time       1.83 seconds
      cpu time        0.32 seconds

```

QUESTION 3 *

```

/* type of test - 2 sample ttest
H0: the average number of drinks per day per individual is not equal for two groups
H1: the average number of drinks per day per individual is equal for two groups */

proc means data=control_treatment nway noperint;
var Alcohol;
class ID;
by Treatment;
output
  out=average_alcohol(drop=_)
  mean;
run;

proc ttest data=average alcohol alpha=0.05;
class Treatment;
var Alcohol;
run;

```

/* We can observe that in the table for equal variances, p>alpha hence we fail to reject the null ie consider equal variances.

In the table just above we check the values for pooled method and notice that p>alpha. Thus we fail to reject the null and can conclude the average number of drinks per day per individual is not equal for two groups and thus the subjects are not randomized perfectly wrt alcohol */

Output - (Untitled) Log - (Untitled) Question 3 * Results Viewer - S...

E:\Users\ssr18000\Downloads 10:56 PM 6/12/2019

