BUAN 6337 Homework 2_Group 9

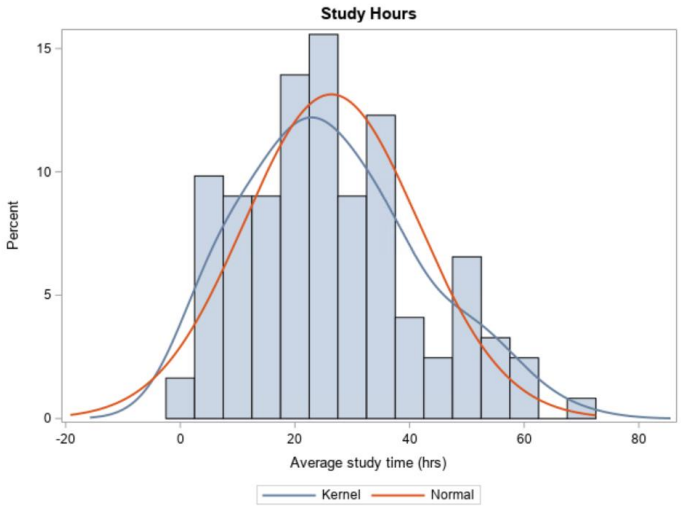
Question 1

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, and asks students to record their study time each week. At the end of the term the instructor compares students' 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) Plot 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?
b) Now, suppose you want to test the normality not just by eyeballing. Conduct a statistical test to check whether the hours of study follows a normal distribution. (Hint: You can use the Univariate procedure)
c) Conduct a hypothesis test to check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 1. What is your conclusion? What variable you think may cause the other?

Answer

a) From the below visualization, Study hours seem to approximate normal distribution as kernel distribution is closer in resemblance to that of the normal curve



b)

Null Hypothesis: Hours of study variable is normally distributed

Alternate Hypothesis: Hours of study variable is not normally distributed

For testing normality, in this case Shapiro Wilk Test is preferable since it performs well for small sample size (here 122 obs). Here p-value of 0.0079 < 0.05 (significance level) and hence we can reject null hypothesis of normality. This implies that distribution is not normal

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.969928	Pr < W	0.0079
Kolmogorov-Smirnov	D	0.067436	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.113863	Pr > W-Sq	0.0765
Anderson-Darling	A-Sq	0.853444	Pr > A-Sq	0.0279

C)

Null Hypothesis: There is not a significant correlation between the units enrolled, hours of study, and GPA.
Alternative Hypothesis: There is a significant correlation between the units enrolled, hours of study, and GPA.

First, we ran a Test for Normality for the three variables (units enrolled, hours of study, and GPA). In each variable we rejected the null hypothesis, which implies the variables are not normally distributed. Thus, we chose to use a Spearman Correlation test.

Based on the Spearman Correlation test, we do not see any strongly correlated combinations. The strongest correlation observed is moderate in strength at .41254 for Average time spent studying and Number of Units Enrolled.

Units Enrolled & Average Time Spent Studying – Correlation .41254 moderate positive correlation | P-Value - 0.0013 Reject the null hypothesis.

The cause of this positive correlation could be that more units enrolled would require, for the average student, more time spent studying. Based on the P-Value, we reject the null hypothesis that there is no correlation.

Average Time Spent Studying & GPA – Correlation -.34888 moderate negative correlation | P-Value 0.0073 Reject the null hypothesis

As average time spent studying increases GPA would be decreasing and vice versa which is unexpected. One would think the more time spent studying would cause GPA to increase. Perhaps the cause is that poorer performing students need to study more to retain the information, or better performing

students study less because they have more efficient methods or might be due to confounding variables. Based on P-value, we reject the null hypothesis which is these two data points are not correlated.

GPA & Units Enrolled – Correlation -0.20791 weak negative correlation | P-Value 0.1173 Fail to reject the null hypothesis

The cause of this negative correlation could possibly be explained by as units enrolled goes up GPA goes down due to the workload making it more difficult to score higher and vice versa. However, based on P-value we fail to reject the null hypothesis which is these data points are not correlated.

The SAS System

The CORR Procedure

Section of course=01

3 Variables:	Units AveTime GPA

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
Units	58	13.79310	3.15538	14.00000	9.00000	19.00000	Number of units enrolled
AveTime	58	29.68670	14.46548	27.57104	0.77286	69.00683	Average study time (hrs)
GPA	58	3.30138	0.39409	3.33500	2.42000	3.94000	GPA

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

	Units	AveTime	GPA
Units Number of units enrolled	1.00000	0.41254 0.0013	-0.20791 0.1173
AveTime Average study time (hrs)	0.41254 0.0013	1.00000	-0.34888 0.0073
GPA GPA	-0.20791 0.1173	-0.34888 0.0073	1.00000

Question 2

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.

(a) Assume there were no placebo (i.e., control) group in your data set. Conduct a test to see whether there is a difference in plaque level before treatment and after the second visit?
(b) 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.
(c) Which of the tests in part (a) and (b) is more reliable? Explain.

(d) 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), conduct two tests to check whether subjects are randomized perfectly.

Answer

a)

Before vs After study of Treatment Group:

Null Hypothesis: Difference in means of plaque level before and after second visit for treatment group with Vitamin E is zero

Alternate Hypothesis: Difference in means of plaque level before and after second visit for treatment group with Vitamin E is not zero

Paired t-test:

Difference: plaque_baseline - plaque_after_second_visit

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

With paired t-test since p value < alpha (0.05), we can reject the null hypothesis and infer that there is a significant difference in means of plaque level before and after second visit for treatment group with Vitamin E without considering placebo into account

b)

Difference-in-difference study:

Null Hypothesis: Difference in difference of means of plaque level before and after second visit for treatment group with Vitamin E and that of control group is zero

Alternate Hypothesis: Difference in difference of means of plaque level before and after second visit for treatment group with Vitamin E and that of control group is not zero

Variable: plaque_reduction

Treatment	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		250	0.0130	0.0990	0.00626	-0.2709	0.2577
1		250	0.0298	0.1182	0.00748	-0.2590	0.3351
Diff (1-2)	Pooled		-0.0168	0.1091	0.00975		
Diff (1-2)	Satterthwaite		-0.0168		0.00975		

Treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		0.0130	0.000628	0.0253	0.0990	0.0911	0.1086
1		0.0298	0.0150	0.0445	0.1182	0.1087	0.1296
Diff (1-2)	Pooled	-0.0168	-0.0360	0.00236	0.1091	0.1027	0.1163
Diff (1-2)	Satterthwaite	-0.0168	-0.0360	0.00236			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	498	-1.72	0.0855
Satterthwaite	Unequal	483.15	-1.72	0.0855

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	249	249	1.43	0.0053

From the equality of variance test, p value of 0.0053 < 0.05 (recommended alpha) this implies that the variances are unequal. Hence by Satterthwaite test, we fail to reject the null hypothesis since p-value of 0.0855 > 0.05 (alpha). Therefore, we conclude that there doesn't seem to be any significant difference in reduction of plaque level between treatment group and that of control group by difference-in-difference study. Provided the 2 samples are identical or in other words randomized, inference from Difference-in-difference study would tend to be insightful. Randomization study in this case could help us know if the groups are randomized perfectly w.r.t other factors like smoking, drinking habits etc which might as well intervene with medical treatment

c)

The test in part (b) is more reliable than the test in part (a). This is because the part (b) test includes a comparison against the control group whereas the test in part (a) only observes the treatment group. The control group is used as a basis of what should occur if there is no medicine used. The treatment group can then be compared to the control group and, if there is a clear distinction in results, it will be much easier to identify. Without the control group present, there is a degree of assumption and speculation to the true impact of the data in visit 0 versus visit 2. Other variables could cause an increase or a decrease in plaque level and comparing against the control group helps to factor these out.

d)

Randomization study for Smoke:

Null Hypothesis: Difference in means of average cigarettes smoked per day between treatment and control group is zero i.e., Groups are identical or randomized w.r.t smoke variable

Alternate Hypothesis: Difference in means of average cigarettes smoked per day between treatment and control group is not zero i.e., Groups are not identical or not randomized w.r.t smoke variable

The SAS System

The TTEST Procedure

Variable: Smoke (Number of cigarettes smoked per day)

Treatment	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		250	4.4200	7.1847	0.4544	0	33.3333
1		250	2.6267	5.2631	0.3329	0	21.3333
Diff (1-2)	Pooled		1.7933	6.2976	0.5633		
Diff (1-2)	Satterthwaite		1.7933		0.5633		

Treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		4.4200	3.5250	5.3150	7.1847	6.6053	7.8764
1		2.6267	1.9711	3.2823	5.2631	4.8387	5.7698
Diff (1-2)	Pooled	1.7933	0.6866	2.9000	6.2976	5.9296	6.7147
Diff (1-2)	Satterthwaite	1.7933	0.6864	2.9003			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	498	3.18	0.0015
Satterthwaite	Unequal	456.49	3.18	0.0016

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	249	249	1.86	<.0001

From the equality of variance test, p value < 0.05 (recommended alpha) this implies that the variances are unequal. Hence by Satterthwaite test, we reject the null hypothesis since p-value of 0.0016 > 0.05 (alpha). Therefore, we conclude that there is a significant difference in means of average amount of cigarettes smoked per day between treatment and control group. Also, from the above table the maximum number of cigarettes smoked is way higher in control group than in treatment group. Hence, we conclude that the samples are not randomized w.r.t smoke variable.

Randomization study for Alcohol:

Null Hypothesis: Difference in means of average amount of alcoholic drinks per day between treatment and control group is zero i.e., Groups are identical or randomized w.r.t alcohol variable

Alternate Hypothesis: Difference in means of average amount of alcoholic drinks per day between treatment and control group is not zero i.e., Groups are not identical or randomized w.r.t alcohol variable

Variable: Alcohol (Number of alcoholic drinks per day)

Treatment	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		250	0.8133	1.2705	0.0804	0	6.0000
1		250	0.6440	1.1807	0.0747	0	6.0000
Diff (1-2)	Pooled		0.1693	1.2264	0.1097		
Diff (1-2)	Satterthwaite		0.1693		0.1097		

Treatment	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		0.8133	0.6551	0.9716	1.2705	1.1681	1.3929
1		0.6440	0.4969	0.7911	1.1807	1.0854	1.2943
Diff (1-2)	Pooled	0.1693	-0.0462	0.3849	1.2264	1.1548	1.3076
Diff (1-2)	Satterthwaite	0.1693	-0.0462	0.3849			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	498	1.54	0.1233
Satterthwaite	Unequal	495.34	1.54	0.1233

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	249	249	1.16	0.2476

From the equality of variance test, p value of 0.2476 > 0.05 (recommended alpha) this implies that the variances are equal. Hence by pooled variance method, we fail to reject the null hypothesis since p-value of 0.1233 > 0.05 (alpha). Therefore, we conclude that there isn't any significant difference in the average amount of alcoholic drinks per day between treatment and that of control group. Also, from the above table the minimum and maximum range of alcoholic drinks (from 0 to 6) is equal in both control and treatment groups. Hence, we conclude that the samples are randomized w.r.t alcohol variable.