FINAL EXAMINATION

COURSE	**:**	**APPLIED STATISTICS**
COURSE CODE	**:**	**BUM2413**
COURSE COORDINATOR	**:**	**KU MUHAMMAD NA'IM KU KHALIF**
DATE	**:**	**10 OCTOBER 2020**
DURATION	**:**	**3 HOURS**
SESSION/SEMESTER	**:**	**SESSION 2019/2020 SEMESTER II**

INSTRUCTIONS TO CANDIDATES:

1. This examination paper consists of **SEVEN (7)** questions. Answer **ALL** questions.
2. All answers to a new question should starts on a new page.
3. All the calculations and assumptions must be clearly stated.
4. Candidates are not allowed to bring any materials other than those allowed by the invigilator into the examination room.
5. All calculations must be in **FOUR (4) decimal places**.

EXAMINATION REQUIREMENTS:

1. Statistical Tables and Formulae

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO

This examination paper consists of **NINE (9)** printed pages including front page.

QUESTION 1 [21 MARKS]

Wind energy is one of the cheapest forms of new electricity generation available today. A group of researchers in Sarawak conducted a study to develop two different wind turbine prototypes namely X and Y. The wind turbine prototype is supposed to have the capacity of 3 MegaWatt (MW). Six wind turbines from each prototype X and prototype Y were tested and the results on the wind turbine capacity (in MW) were recorded in **Table 1**.

Table 1: Wind Turbine Capacity

Prototype X	2.90	3.00	2.90	2.70	3.00	2.95
Prototype Y	2.65	3.05	3.00	2.80	3.00	2.60

Assume that the data for the capacity of wind turbine prototypes X and Y are normally distributed.

i) Find a 96% confidence interval for the ratio of population variances for capacity of wind turbine prototypes X and Y.

[7 Marks]

ii) Based on the confidence interval obtained in **i)**, test if the wind turbine prototypes X and Y have similar variability of capacity.

[4 Marks]

iii) Does the data suggest that wind turbine prototype Y has higher average of capacity compared to prototype X at 3% significance level? Use the assumption of the equality of population variances obtained in **ii)**.

[10 Marks]

QUESTION 2 [8 MARKS]

Biodiesel is an alternative fuel for diesel engine to replace or reduce the use of conventional diesel. An experiment was conducted to compare the emission of harmful gas such as unburn hydrocarbon (UHC) when using conventional diesel and biodiesel fuels. A random sample of eight diesel engines were injected with conventional diesel and the UHC emissions (in ppm) were recorded. The same procedure was repeated for the same eight diesel engines injected with biodiesel fuel. **Figure 1** presents the *Microsoft Excel* output for the related test on the UHC emissions of diesel engine using conventional diesel and biodiesel fuel at 10% significance level.

t-Test: Paired Two Sample for Means

	Biodiesel	*Conventional Diesel*
Mean	178.5	170.875
Variance	56.2857	109.8393
Observations	8	8
Pearson Correlation	-0.2462	
Hypothesized Mean Difference	40	
df	7	
t Stat	-6.3981	
P(T<=t) one-tail	1.8398E-04	
t Critical one-tail	1.4149	
P(T<=t) two-tail	3.6796E-04	
t Critical two-tail	1.8946	

Figure 1: *Microsoft Excel* Output

Assume that the data on UHC emission by conventional diesel and biodiesel fuels are normally distributed.

i) Determine whether biodiesel fuel produces more UHC emission by at most 40ppm compared to conventional diesel fuel.

[5 Marks]

ii) Based on your answer in **i)**, is it possible that the biodiesel fuel produces less UHC emission than the conventional diesel?

[1 Mark]

iii) Identify the type of error involved in **i)**. Justify your answer.

[2 Marks]

QUESTION 3 [15 MARKS]

A vehicle manufacturer wishes to test the ability of three types of steel-alloy panels to resist corrosion when three different paint types are applied. The corrosion data in millimetre (mm) represents the ability of the painted panels to resist weathering. A random sample of four different types of steel-alloy panel for different type of paints were recorded in **Table 2**. The data is analysed and the *Microsoft Excel* output is given in **Figure 2**.

Table 2: Corrosion Data

Paint Type	Steel-Alloy 1	Steel-Alloy 2	Steel-Alloy 3
A	40	51	56
	45	50	57
	35	53	58
	44	54	60
B	54	55	67
	53	58	65
	50	54	66
	51	53	66
C	47	43	55
	45	44	58
	34	46	69
	38	45	56

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	**P**	2	291.583	**U**	1.6452E-06	**V**
Columns	1662.1667	**Q**	831.083	64.713	5.0146E-11	3.3541
Interaction	124.6667	**R**	31.1667	2.4268	0.07230703	2.7278
Within	346.75	27	**T**			
Total	2716.75	**S**				

Figure 2: *Microsoft Excel* Output for ANOVA Table

i) Find the values of **P**, **Q**, **R**, **S**, **T**, **U** and **V** in **Figure 2**.

[5 Marks]

ii) Given that there is no interaction exist between the types of paint and types of steel-alloy panels on the effect of corrosion, test the marginal effect at $\alpha = 0.05$. Use the *P*-value approach.

[10 Marks]

QUESTION 4 [17 MARKS]

In Department of Educational Psychology & Counselling, University of Harimau Malaya, a researcher noted that people have more difficulty to sleep in a bright room than in a dark room. To study the relationship between the intensity of light (in watt) and time taken to fall asleep (in minute) for people, he took a sample of 12 staff from the university. The data is recorded and summarised as follows.

$$\sum_{i=1}^{12} x_i = 285, \quad \sum_{i=1}^{12} x_i^2 = 8561, \quad \sum_{i=1}^{12} y_i = 308, \quad \sum_{i=1}^{12} y_i^2 = 9030, \quad \sum_{i=1}^{12} x_i y_i = 8672$$

$$S_{xx} = 1792.25, S_{yy} = 1124.6667, S_{xy} = 1375, se\left(\hat{\beta}_1\right) = \sqrt{0.0047}$$

Based on the data summary, answer the following questions.

i) Identify the control and response variables.

[2 Marks]

ii) Calculate mean for the control and response variables.

[4 Marks]

iii) Test whether there is a significant relationship between the intensity of light and time it took to fall asleep at 6% significance level by using *t*-test approach.

[9 Marks]

iv) Can the same conclusion obtained in **iii)** be implemented if the samples are university student? Justify your answer.

[2 Marks]

QUESTION 5 [14 MARKS]

The unprecedented increase in CO_2 emissions has become an important global issue. An engineer used a secondary data from the Emission Department to study CO_2 emissions in Malaysian road transport sector. The data records the CO_2 emission (y), (in Million ton), the fuel consumption (x_1), (in ktoe), the average of fuel efficiency (x_2), (in km/L), and the average of distance travel (x_3), (in Bpkm) for emission readings obtained from a certain region. A multiple linear regression analysis is conducted and the *Microsoft Excel* output is given in **Figure 3**.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9420
R Square	0.8873
Adjusted R Square	0.8309
Standard Error	3.4707
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	568.9131	189.6377	15.7432	0.0030
Residual	6	72.2743	12.0457		
Total	9	641.1874			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	14.1432	11.0217	1.2832	0.2467	-12.8259	41.1122
Average of fuel consumption	0.0023	0.0007	3.3914	0.0147	0.0006	0.0039
Average of fuel efficiency	-0.7267	1.6262	-0.4469	0.6707	-4.7058	3.2524
Average of distance travel	0.0156	0.0145	1.0721	0.3249	-0.0200	0.0511

Figure 3: *Microsoft Excel* Output of Multiple Linear Regression Analysis

Based on **Figure 3**, answer the following questions.

i) How many samples of CO_2 emission readings are used in this study?

[1 Mark]

ii) Identify and interpret the coefficient of determination value.

[2 Marks]

iii) Interpret the estimated value of regression coefficients of the average of distance travel, $\hat{\beta}_3$.

[1 Mark]

iv) Identify the significant predictors based on the *P*-values in the coefficient table at 1% significance level.

[5 Marks]

v) The summary of the multiple linear regression analysis for the study is given by **Table 3**. Complete the table for the three predictors based on **Figure 3**. Then, select the best regression model for studying the CO_2 emission at 0.05 significance level. Justify your answer for the selection.

[3 Marks]

Table 3: Summary Table for Multiple Linear Regression Analysis

Predictor	*P*-value	r^2	Adjusted r^2	Regression equation
x_1	9.8E-05	0.8644	0.8474	$\hat{y} = 10.7771 + 0.0026x_1$
x_2	0.06421	0.3653	0.2859	$\hat{y} = -14.3883 + 5.1858x_2$
x_3	0.004853	0.6499	0.6061	$\hat{y} = 11.2613 + 0.05341x_3$
x_1x_2	8.9E-04	0.8657	0.8273	$\hat{y} = 13.5953 + 0.0027x_1 - 0.4229x_2$
x_1x_3	5.4E-04	0.8835	0.8503	$\hat{y} = 9.4078 + 0.0021x_1 + 0.0144x_3$
x_2x_3	0.02038	0.6712	0.5773	$\hat{y} = 0.7688 + 1.5743x_2 + 0.0461x_3$
$x_1x_2x_3$				

vi) Based on the best regression model in **v)**, predict the CO_2 emission when $x_1 = 2889\text{ktoe}$, $x_2 = 7.47\text{km/L}$ and $x_3 = 106.4\text{Bpkm}$.

[2 Marks]

QUESTION 6 [13 MARKS]

(a) In a goodness of fit test, what should be done if the value of expected frequency is less than five?

[1 Mark]

(b) In finding the critical value of Chi-square distribution, what should be the value of m if the value of the parameters are all given?

[1 Mark]

(c) An analysis of fat content (in %) of burgers in certain fast food restaurant has been investigated by a food inspector. A random sample of 175 burgers of particular grade resulted the information shown in **Table 4**.

Table 4: Fat Content of Burgers

Fat Content	Frequency	Probability
$26 \le X < 28$	7	0.043
$28 \le X < 30$	22	0.108
$30 \le X < 32$	36	0.214
$32 \le X < 34$	45	0.270
$34 \le X < 36$	33	0.214
$36 \le X < 38$	28	0.108
$38 \le X < 40$	4	0.043

i) The food inspector claimed that the fat content follows a Normal distribution with a mean of 33% and standard deviation of 2.91%. Test the food inspector's claim at 2.5% significance level.

[9 Marks]

ii) Based on your answer in **i)**, does the data have a good fit to the model? Give a reason.

[2 Marks]

QUESTION 7 [12 MARKS]

Three groups of drivers where 500 drivers in each group are observed for a 3-year period. The frequencies of accidents of the three groups of drivers are recorded and depicted in **Table 5**.

Table 5: Number of Accidents Recorded

Number of accidents	Drivers		
	Group 1	Group 2	Group 3
0	209	280	240
More than 1	291	220	260

i) Identify one of the variables under study.

[1 Mark]

ii) Identify the level of measurement associated with the variable in **i)**.

[1 Mark]

iii) Test at 0.5% significance level whether the number of accidents is the same among the groups of drivers.

[10 Marks]

END OF QUESTION PAPER