FINAL EXAMINATION

COURSE	**:**	**APPLIED STATISTICS**
COURSE CODE	**:**	**BUM2413**
COURSE COORDINATOR	**:**	**SITI ROSLINDAR YAZIZ**
DATE	**:**	**3 JANUARY 2020**
DURATION	**:**	**3 HOURS**
SESSION/SEMESTER	**:**	**SESSION 2019/2020 SEMESTER I**

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 **TEN (10)** printed pages including front page.

QUESTION 1 [20 MARKS]

(a) A supervisor of a factory claims that the bottling machine is operating properly if at least 95% of the processed bottles are full. Based on a daily-checking on the machine, out of 300 bottles selected, it was found that 20 bottles were not full.

i) Identify the appropriate variable in the study.

[1 Mark]

ii) Conduct a test at 10% level of significance to determine if the machine is operating properly based on data from the daily-checking.

[8 Marks]

(b) BMA Sdn. Bhd. is the frozen food manufacturer of Product A and Product B. The management of the manufacturer claims that there is a difference in demand between Product A and Product B in rural and urban areas. In a recent study of customers' product demand, a survey is done in rural and urban areas to test the claim by asking their customers' preference on the products. In a random sample of 100 customers in rural area, 70 of them preferred Product A than Product B. While, in a random sample of 100 customers in urban area, 50% of them chose Product A.

i) Do the data provide sufficient evidence that the proportions of all demands for Product A are different between rural and urban areas? Conduct an appropriate test for the study at $\alpha = 0.03$ by applying the confidence interval approach.

[10 Marks]

ii) Identify the type of hypothesis test in **(i)**, either as a two-tailed test, a right-tailed test or a left-tailed test.

[1 Mark]

QUESTION 2 [9 MARKS]

Variability in service time is an important statistical measure in banking facilities. In comparing the performance of Bank X and Bank Y regarding their service to customers, the service times data are collected from 100 customers of each bank. **Figure 1** presents the *Microsoft Excel* output for the related statistical test on variability in both banks.

F-Test Two-Sample for Variances

	Bank X	*Bank Y*
Mean	10.2330	9.6590
Variance	8.2255	8.3822
Observations	100	100
df	99	99
F	0.9813	
P(F<=f) one-tail	0.4627	
F Critical one-tail	0.6244	

Figure 1

i) Identify the parameter involved in the analysis.

[1 Mark]

ii) Using 7% significance level, do the data support the claim that Bank X provides more efficient service time than Bank Y? Use the *P*-value approach.

[5 Marks]

iii) Based on the result in **(ii)**, which bank provides less efficient service time?

[1 Mark]

iv) Based on the result in **(ii)**, identify the type of error involved in this study and give a reason.

[2 Marks]

QUESTION 3 [15 MARKS]

(a) Four different types of solar energy panels were measured to test the performance on energy absorption. Each type was observed at five randomly chosen times, and the power (in watts) was measured. **Figure 2** presents the *Microsoft Excel* output based on the recorded data.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.6135				2.7734E-05	3.2389
Within Groups			0.0310			
Total	2.1095	19				

Figure 2

i) How many treatment(s) involved in the study?

[1 Mark]

ii) Copy and complete the *Microsoft Excel* output in **Figure 2**.

[4 Marks]

iii) Based on **Figure 2**, can you conclude that the mean power differs for different type of panels?

[5 Marks]

(b) Adding glass particles to clay brick may improve the structural properties of the brick. An article in Civil Engineering describes experiments in which the compressive strength (in MPa) was measured for bricks with varying amounts of glass content (in %) and glass particle size. *Microsoft Excel* output in **Figure 3** shows all the effects involved between the amounts of glass content and glass particle size on strength of the brick.

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	6322.0858	2	3161.0429	25.7968	0.0000	6.0129
Columns	1890.3750	1	1890.3750	15.4271	0.0010	8.2854
Interaction	1305.0025	2	652.5013	5.3250	0.0153	6.0129
Within	2205.6500	18	122.5361			
Total	11723.1133	23				

Figure 3

Since there is no interaction effect exists between the amount of glass content and glass particle size, test the effect of glass particle size on the strength of the brick at 1% significance level.

[5 Marks]

QUESTION 4 [10 MARKS]

A staff of a college bookstore must place an order of a book two months before a semester starts. He believes that the number of books that will be sold for any particular courses is related to the number of students registered for the course when the books are ordered. He would like to develop a linear regression model to predict the number of books to order. The number of students registered and the number of books actually sold for a course over 12 different semesters are recorded. The data summary is given as follows.

$$\sum_{n=1}^{12} x = 905,\ \sum_{n=1}^{12} x^2 = 71027,\ \sum_{n=1}^{12} y = 826,\ \sum_{n=1}^{12} y^2 = 61684,\ \sum_{n=1}^{12} xy = 64478$$

$$S_{xx} = 2774.9167,\ S_{yy} = 4827.6667,\ S_{xy} = 2183.8333,\ se\left(\hat{\beta}_1\right) = \sqrt{0.1120}$$

Based on the data summary, answer the following questions.

i) Identify the control variable.

[1 Mark]

ii) Test whether there is a significant relationship between the number of books that will be sold and the number of students registered at 6% significance level by using *t*-test approach.

[9 Marks]

QUESTION 5 [20 MARKS]

Nowadays, stroke is one of the top five leading causes of death and one of the top ten causes for hospitalisation in Malaysia. A researcher interested to investigate the major factor that increase the risk of having stroke. Risk (in %) is interpreted as the probability that the patient will have a stroke over the next 10-year period. The researcher collects data on predictor variables such as age (in year), blood pressure (in mm/Hg), and smoking habit (number of cigarettes per day) which are related to the risk of having stroke. An analysis of relationship between the risk of stroke with the predictor variables produces the *Microsoft Excel* output as shown in **Figure 4**.

Regression Statistics								
Multiple R	0.970132592							
R Square	0.941157245							
Adjusted R Square	0.930124229							
Standard Error	3.925931753							
Observations	20							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	3	3944.342958	1314.781	85.30371	4.67115E-10			
Residual	16	246.6070421	15.41294					
Total	19	4190.95						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-24.9597	17.0114	-1.4672	0.1617	-61.0223	11.1029	-61.0223	11.1029
AGE	0.2679	0.1991	1.3460	0.1971	-0.1540	0.6899	-0.1540	0.6899
BLOOD PRESSURE	0.0694	0.0474	1.4641	0.1625	-0.0311	0.1699	-0.0311	0.1699
SMOKER	3.3697	0.2263	6.0530	0.0000	0.8900	1.8494	0.8900	1.8494

Figure 4

Based on **Figure 4**, answer the following questions.

i) How many patients are involved in this study?

[1 Mark]

ii) Interpret the coefficient of determination value.

[2 Marks]

iii) Is there any evidence to conclude that at least one of the predictor variables is related to the response variable?

[5 Marks]

iv) Based on the *P*-value in the coefficients table, identify the significant predictors.

[4 Marks]

v) The summary of the multiple linear regression analysis for the study is given by **Table 1**. If only one predictor is used to predict the risk of having stroke, which single variable is the best? Give your reason. Then, select the best regression model for predict the risk of having stroke at 1% significance level. Justify your answer for the selection.

[5 Marks]

Table 1

Predictor(s)	***P*-value**	r^2	**Adjusted** r^2	**Regression Model**
x_1	0.0002	0.6228	0.5907	$\hat{y} = -42.7965 + 1.0043x_1$
x_2	0.0908	0.1506	0.1035	$\hat{y} = -0.8056 + 0.1767x_2$
x_3	0.0000	0.9328	0.9291	$\hat{y} = -0.2127 + 1.6613x_3$
x_1, x_2	0.0000	0.8064	0.7836	$\hat{y} = -110.9420 + 1.3150x_1 + 0.2964x_2$
x_1, x_3	0.0000	0.9332	0.9254	$\hat{y} = -2.5335 + 0.0404x_1 + 1.6317x_3$
x_2, x_3	0.0000	0.9344	0.9267	$\hat{y} = -2.8503 + 0.0196x_2 + 1.6344x_3$
x_1, x_2, x_3	0.0000	0.9411	0.9301	$\hat{y} = -24.9597 + 0.2679x_1 + 0.0694x_2 + 3.3697x_3$

vi) Mr. K is a 45-year old patient who smokes regularly 20 cigarettes/day and has 180 mmHg blood pressure. Predict the risk of Mr. K having stroke based on the best multiple linear regression model in **(v)**.

[3 Marks]

QUESTION 6 [13 MARKS]

(a) What should be done if the value of expected frequency of goodness of fit test less than five?

[1 Mark]

(b) Metal plates are evaluated according to their surface finish and placed into four categories; premium, conforming, downgraded and unacceptable. A quality engineer claims that the proportions of metal plates in the four categories are 10%, 70%, 15% and 5%, respectively. In a sample of 200 metal plates, 19 were classified as premium, 133 as conforming, 35 as downgraded and 13 as unacceptable.

(i) Use 0.5% significance level to test whether the engineer's claim is true.

[11 Marks]

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

[1 Mark]

QUESTION 7 [13 MARKS]

Two different design configurations, namely A and B, are being considered for a particular component in computer system. There are three possible failure modes, namely X, Y and Z, for the component. An engineer obtained the data of number of failures in each mode for each of the two configurations as recorded in **Table 2**.

Table 2

		Failure Mode		
		X	**Y**	**Z**
Configuration	**A**	20	44	17
	B	4	17	7

i) Identify the variable under study.

[1 Mark]

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

[1 Mark]

iii) At 2.5% significance level, does the configuration appear to have an effect on mode of failure?

[11 Marks]

END OF QUESTION PAPER