UNIVERSITI TEKNOLOGI MARA
FINAL EXAMINATION

COURSE	:	STATISTICS FOR BUSINESS AND SOCIAL SCIENCES
COURSE CODE	:	STA404
EXAMINATION	:	DECEMBER 2018
TIME	:	2 HOURS

INSTRUCTIONS TO CANDIDATES

1. This question paper consists of seven (7) questions.
2. Answer ALL questions in the Answer Booklet. Start each answer on a new page.
3. Do not bring any material into the examination room unless permission is given by the invigilator.
4. Please check to make sure that this examination pack consists of :
   i) the Question Paper
   ii) a five – page Appendix 1
   iii) an Answer Booklet – provided by the Faculty
   iv) a Statistical Table – provided by the Faculty
5. Answer ALL questions in English.

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

This examination paper consists of 6 printed pages

QUESTION 1 Unit 1 - Introduction to Statistics

In the automobile industry, customer service is a crucial factor affecting car sales. The management of a reputed automobile company is interested in determining the level of customers' satisfaction with the service provided by the company's service centers. The company has altogether 40 service centers throughout Malaysia. A sample of eight centers was selected at random. Questionnaires are disseminated to all customers who service their cars at these eight selected service centers on one selected day (the day of the survey). One of the questions asked is satisfaction level on the services provided (using rating: good, fair, poor).

a) State the population of the study.

All of the company service centers customers

(1 mark)

b) Name the variable of interest for the above study. State its type and its level of measurement.

variable: customers satisfaction level
type: qualitative variable
level of measurement: ordinal

(3 marks)

c) Identify the sampling technique used in the survey. Explain briefly how the sample is selected.

cluster sampling.
1. divide the population into non-overlapping groups (40 service centers)
2. select at random 8 service centers (can apply SRS or systematic sampling)
3. disseminate questionnaire to all ...

(3 marks)

QUESTION 2 Unit 2 - Descriptive Statistics

The descriptive statistics for the life span (in years) of brand AA washing machine are summarized as below.

Descriptive statistics

	N	Mean	Mode	Std. Deviation	Minimum	Maximum
Life span (years)	89	6.59	7.04	0.74	5.2	8.1

a) Calculate the coefficient of skewness. Hence, comment on the shape of the distribution.

$$\text{Skewness} = \frac{\bar{x} - \hat{x}}{s} = \frac{6.59 - 7.04}{0.74} = -0.608$$

Comment:
Since the skewness value is -0.608 < 0, therefore the distribution is skewed to the left.

(2 marks)

b) Explain the meaning of the mode value.

mode = 7.04 years.
This means that, most of the washing machines have a life span of approximately 7.04 years

(1 mark)

c) Given the mean and variance for the life span (in years) of brand BB were 7.1 and 12.3 respectively. Using an appropriate measurement, determine which brand has a more consistent life span.

$$CV = \frac{s}{\bar{x}} \times 100\%$$

$$CV_{AA} = \frac{0.74}{6.59} \times 100\% = 11.2\%$$

$$CV_{BB} = \frac{\sqrt{12.3}}{7.1} \times 100\% = 49.4\%$$

(4 marks)

Since CV for AA brand is smaller than CV for BB brand, Brand AA has a more consistent distribution of life span (in years)

QUESTION 3 Unit 3 - Estimation

A study was done to estimate the average monthly electricity bill (in RM) for a company. The manager claims that the average monthly electricity bill (in RM) is RM700. Analysis of previous 20 monthly bills gives the following result.

One Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Monthly_bill (RM)	20	705.8	114.5695	**K**

a) Show that **K** is 25.6185. $k = s/\sqrt{n} = \frac{114.5695}{\sqrt{20}} = 25.6185$ (2 marks)

$\alpha = 10\%$ or $0.1 \rightarrow \alpha/2 = 0.05$

b) Calculate a 90% confidence interval for the mean monthly electricity bill. (3 marks)

$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \Rightarrow 705.8 \pm 1.729(25.6185) = (661.5, 750.1)$

c) Construct a 95% confidence interval for the mean monthly electricity bill. (3 marks)

$t_{0.025,19} = 2.093 \Rightarrow 705.8 \pm 2.093(25.6185) = (652.2, 759.4)$

d) Based on the confidence interval in b), is the manager's claim true? Give a reason to support your answer. (2 marks)

Yes. There is enough evidence to support the managers claim that the average monthly electricity bill is RM700 since the value is included in the confidence interval.

e) What can be concluded based on the confidence interval conducted in b) and c) in term of the size of the interval? (2 marks)

When the confidence level increase from 90% to 95%, the size of intervals become larger. In other words, the confidence level is directly related to the size of CI.

QUESTION 4 Unit 4 - Difference Between Two Means (Dependent Samples)

The manager of an insurance company hired a marketing executive officer to advice the best advertising strategies to improve the sales of the insurance policies under the company. In order to investigate whether the sales had improved, the manager recorded the sales of the insurance policies a month before and after the officer was hired. The data was analyzed using SPSS and the result is as follows. (one-tail)

Paired Samples Test

	Paired Differences					t	df
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference			
				Lower	Upper		
Pair 1 Sales before - Sales after	- 4.94	3.901	1.745	- 9.784	- .096	- 2.83	4

a) Show how the value of test statistic, *t* is obtained. (2 marks)

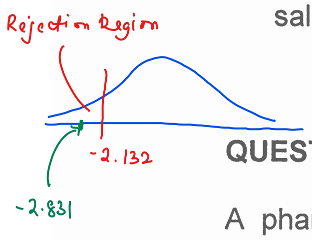
$t = \frac{\bar{d} - d_0}{s_d/\sqrt{n}} = \frac{-4.94 - 0}{1.745} = -2.831$

b) State the null and alternative hypotheses. (1 mark)

Null: There is no improvement in the sales of insurance policies after hiring the officer
Alternative: There is improvement in the sales of insurance policies after hiring the officer

c) At the 5% level of significance, do the data provide sufficient evidence to indicate that the sales had improved? (4 marks)

t critical value = -2.132
t statistics = -2.831
Decision: reject null hypothesis, if t statistic < t critical value.
Since -2.831 < -2.132, we reject the null hypothesis.
Conclusion: There is enough evidence that the sales had improved



QUESTION 5 Unit 4 - One Way ANOVA

A pharmacist at Dungun Hospital conducted a study to determine whether there is a significant difference in time taken for the three brands of medicine to provide relief from headache. The following data depicts the time (in minutes) taken by patients to get relief from headache after taking the medicine.

$\bar{x}_1 = \frac{9+10+12+15+11}{5} = 11.4$

Brand 1	Brand 2	Brand 3
9	15	14
10	20	19
12	12	14
15	16	18
11	10	11
11.4	14.6	15.2

GM 13.7

The data was analyzed using SPSS and the result is shown in the following ANOVA table.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	Sig.
Between	**A**	2	**C**	2.032	0.174
Within	**B**	12	**D**		
Total	164.933	14			

a) Using the Sum of Square Between (SSB) formula, show that **A** = 41.733. Then, determine the values of **B**, **C** and **D**. (5 marks)

$A = 5(11.4 - 13.73)^2 + 5(14.6 - 13.73)^2 + 5(15.2 - 13.73)^2 = 41.733$
B = 164.933 - 41.733 = 123.2
C = 41.733/2 = 20.8665; D = 123.2/12 = 10.2667

b) State the null and alternative hypotheses. (1 mark)

Null: There is no difference in average time taken for three brands of medicine to provide relief from headache
Alternative: at least one average time taken differ from others

c) Based on the output, is there any evidence to indicate there is difference in time taken for the three brands of medicine to provide relief from headache? Use α = 0.05. (3 marks)

alpha = 0.05
decision: failed to reject null hypothesis since the p-value = 0.174 > alpha = 0.05
conclusion: there is no evidence to indicate there is difference in average time taken for three brands of medicine to provide relief from headache

QUESTION 6 Unit 4 - Chi square independence test

A lecturer at Faculty of Hotel and Tourism Management wishes to determine whether the program of the students is related to their smoking habit. The results obtained are shown below.

Program * Smoking Habit Crosstabulation

			Smoking habit			Total
			None	Light	Heavy	
Program	HM240	Count	46	36	13	95
		Expected Count	**P**	29.6	24.8	95.0
	HM242	Count	31	20	34	85
		Expected Count	36.4	26.4	22.2	85.0
Total		Count	77	56	47	180
		Expected Count	77.0	56.0	47.0	180.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	**Q**	2	.000
Likelihood Ratio	18.975	2	.000
Linear by Linear Association	11.257	1	.001
N of Valid Cases	180		

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

a) Find the values of **P** and **Q**. Show all your calculations.
P = (row total x column total) / grand total = (95 x 77) / 180 = 40.6
P = 77 - 36.4 = 40.6
Q = 16.3412
(3 marks)

b) State the null and alternative hypotheses.
Null: The program of the students is not related to their smoking habit
Alternative: The program of the students is related to their smoking habit
(2 marks)

c) At the 1% level of significance, is there sufficient evidence to conclude that the program of the students is related to their smoking habit?
(3 marks)

p-value method
alpha = 0.01
decision: reject null hypothesis since p-value < 0.001 which is smaller than the alpha value of 0.01
conclusion: hence, there is sufficient evidence that the program of the students is related to their smoking habit

traditional method
Chi square CV = 9.210
test statistic = 16.3412
decision: reject the null hypothesis since the test statistic value > chi square critical value
conclusion: hence, there is sufficient evidence that the program of the students is related to their smoking habit

QUESTION 7 Unit 5 - Bivariate Analysis (Correlation & Regression)

A study was conducted to determine the relationship between working experience (in years) and the monthly salary (RM'00) of teachers at a primary school in Kuantan. The data were analyzed using SPSS and produced the following tables.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**M**	.878	.869	2.732

Coefficients

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	23.470	1.414	16.602	.000
	Experience	1.194	.123	9.685	.000

Answer the following questions based on the given output.

a) State the dependent and independent variables.
independent = working experience
dependent = monthly summary
(2 marks)

b) State the value of the coefficient of determination. Hence, interpret the value.
r-square = 0.878
interpretation: 87.8% variations in teacher's monthly salary can be explained by the variations of their working experience (in years)
(2 marks)

c) Determine the value of **M**. Comment on the value.
M = 0.937
comment: there exist a very strong positive relationship between teacher's monthly salary and their working experience.
(2 marks)

d) Find the regression line for the above study. Explain the value of the slope.
y = 23.47 + 1.194 x
slope = 1.194. For each additional increase in working experience by 1 year, the teacher's monthly salary will increase by 1.194 (RM '00) or RM119.40
(2 marks)

e) Estimate the monthly salary for a teacher with 15 years of working experiences.
x = 15
y = 23.47 + 1.194(15) = 41.38 (RM '00) or RM4,138
(2 marks)

END OF QUESTION PAPER

SAMPLE MEASUREMENTS

Mean	$\bar{x} = \frac{\sum x}{n}$
Standard deviation	$s = \sqrt{\frac{1}{n-1}\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]}$ or $s = \sqrt{\frac{1}{n-1}\left[\sum (x-\bar{x})^2\right]}$
Coefficient of Variation	$CV = \frac{s}{\bar{x}} \times 100\%$
Pearson's Measure of Skewness	Coefficient of Skewness = $\frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$ OR $\frac{\text{mean} - \text{mode}}{\text{standard deviation}}$

CONFIDENCE INTERVAL

Parameter and description	A (1 - α) 100% confidence interval
Mean μ, for large samples	$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$
Mean μ, for small samples, variance σ^2 unknown	$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$; $df = n - 1$
Difference in means of two normal distributions $\mu_1 - \mu_2$, variances $\sigma_1^2 = \sigma_2^2$ and unknown	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$; $df = n_1 + n_2 - 2$ $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
Difference in means of two normal distributions $\mu_1 - \mu_2$, variances $\sigma_1^2 \neq \sigma_2^2$ and unknown	$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$; $df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$
Mean difference of two normal distributions for paired samples, μ_d	$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$; $df = n - 1$ where n is no. of pairs

HYPOTHESIS TESTING

Null Hypothesis	Test statistic
$H_0 : \mu = \mu_0$ σ^2 known, large samples	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ or $z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
$H_0 : \mu = \mu_0$ σ^2 known, small samples	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$; $df = n - 1$
$H_0 : \mu_1 - \mu_2 = 0$ $\sigma_1^2 = \sigma_2^2$ and unknown	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$; $df = n_1 + n_2 - 2$ $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
$H_0 : \mu_1 - \mu_2 = 0$ $\sigma_1^2 \neq \sigma_2^2$ and unknown	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ $df = \dfrac{\left[\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right]^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$
$H_0 : \mu_d = 0$	$t = \dfrac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$; $df = n - 1$, where n is no. of pairs
Hypothesis for categorical data	$\chi^2 = \sum \dfrac{(o_{ij} - e_{ij})^2}{e_{ij}}$

ANALYSIS OF VARIANCE FOR A COMPLETELY RANDOMIZED DESIGN

Let:

k	=	the number of different samples (or treatments)
n_i	=	the size of sample i
T_i	=	the sum of the values in sample i
n	=	the number of values in all samples
	=	$n_1 + n_2 + n_3 + \ldots$
$\sum x$	=	the sum of the values in all samples
	=	$T_1 + T_2 + T_3 + \ldots$
$\sum x^2$	=	the sum of the squares of values in all samples

Degrees of freedom for the numerator = $k - 1$
Degrees of freedom for the denominator = $n - k$

Total sum of squares: $SST = \sum x^2 - \frac{(\sum x)^2}{n}$

Between-samples sum of squares:

$$SSB = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \ldots\right) - \frac{(\sum x)^2}{n}$$

Within- samples sum of squares = SST - SSB

Variance between samples: $MSB = \frac{SSB}{(k-1)}$

Variance within samples: $MSW = \frac{SSW}{(n-k)}$

Test statistic for a one-way ANOVA test: $F = \frac{MSB}{MSW}$

SIMPLE LINEAR REGRESSION

Sum of squares of *xy*, *xx*, and *yy*:

$$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \quad \text{and} \quad SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

Least Square Regression Line:

Y = a + bx

Least Squares Estimates of *a* and *b*:

$$b = \frac{SS_{xy}}{SS_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

Total sum of squares: $SST = \sum y^2 - \frac{(\sum y)^2}{n}$

Linear correlation coefficient: $r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$