UNIVERSITI TEKNOLOGI MARA
FINAL EXAMINATION

COURSE	:	**STATISTICS FOR BUSINESS AND SOCIAL SCIENCES**
COURSE CODE	:	**STA404**
EXAMINATION	:	**JANUARY 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 four – 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 Chapter 1 - Introduction

A headmaster of a private higher learning institution is interested to study the relationship between the students' hours spending on social media and their academic performances. He believed that the more time students spent on social media, the more likely the students will fail in their academics. The institution has a total of 2500 students. Based on the previous semester examinations, the students' academic performance has been categorized as Excellent, Moderate and Low, whereby the number of students in each categories are 750, 1350, 400, respectively. A random sample of 100 students was selected for this study and the time spent on social media was recorded.

a) State the population and the sample for the above study. (2 marks)

Population = All 2500 students of a private higher learning institution
Sample = A random sample of 100 students

b) Identify whether the researcher conducted a census or sample survey. Give a reason for your answer. (2 marks)

Sample survey.
Reason: only 100 students were selected out of the total of 2500 students

c) Suggest another sampling method that can be used by the headmaster. Explain briefly the sampling method chosen. (3 marks)

Stratified Random Sampling.
By using this sampling design, the headmaster is able to make comparison across different categories of students.

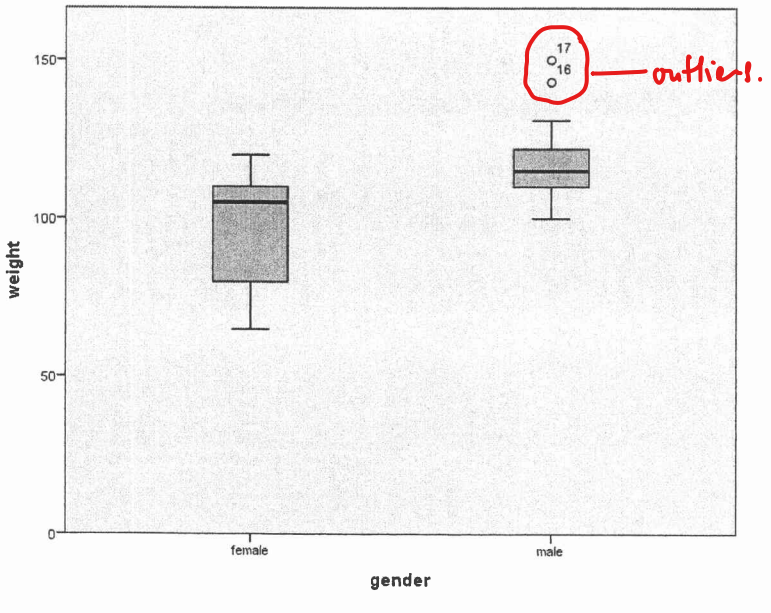
1. Divide the population into non-overlapping groups (Excellent, Moderate, Low)
2. Calculate the required samples needed in each category.
   Excellent = 750/2500 x 100 = 30, Moderate = 54, Low = 16
3. Apply SRS or systematic sampling to obtain the required sample calculated in 2.

QUESTION 2

Chapter 2 - Descriptive Statistics

The following chart summarizes the weight in kilogram of 17 male and female orang utans in the Reserves of Semonggok.

Gender	Mean	Std. Deviation	Minimum	Maximum
Male	117.94	13.32	100	150
Female	97.65	17.79	65	120



a) Name the diagram given.
Box and Whiskers Plot (1 mark)

b) Identify the outlier(s), if any.
There are two outliers which are both male orang utans (case #16 and #17) (2 marks)

c) Using an appropriate measure, determine which gender is more consistent in distribution of weight.
$CV = \frac{s}{\bar{x}} \times 100\%$
CV male = 13.32 / 117.94 x100% = 11.3%
CV female = 17.79 / 97.65 x 100% = 18.2%
Thus, the male orang utan has a more consistent weight distribution (5 marks)

QUESTION 3

Chapter 3 - Estimation

There was a claim that the price of *ikan kembung* sold in a certain market was different from the average RM17 per kg. A study was conducted to investigate the changing price per kg (in RM) of *ikan kembung*. Fifty stalls were selected at random and the results obtained are as follows:

One Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Price per kg (in RM)	50 ✓	17.2740	0.8354 *	0.1181 ✓

a) Show that the standard error of the mean is 0.1181.
$SE = s/\sqrt{n} = 0.8354/\sqrt{50} = 0.1181$ * same value (3 marks)

b) Construct a 99% confidence interval for the mean price of *ikan kembung*. $(\alpha = 0.01) \Rightarrow \alpha/2 = 0.005$
$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 17.2740 \pm 2.58\left(\frac{0.8354}{\sqrt{50}}\right) = (16.9692, 17.5788)$ (4 marks)

c) Based on the confidence interval in b), does the average price per kg (in RM) of *ikan kembung* in the market differ from RM17? Give a reason to support your answer.
There is enough evidence that the average price per kg (in RM) of ikan kembung do not differ from RM17 since this value is included in the confidence interval. (2 marks)

QUESTION 4 Chapter 4 - Difference between two means (dependent samples)

A researcher claims that after playing a certain type of interactive game, the memory capability of a group of autistic children had improved. In order to test her hypothesis, a sample of 20 autistic children was selected and the ability to memorize items out of 10 before and after playing the interactive game was recorded. The data were analyzed using SPSS and have the following outputs:

Paired Sample Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Memory1 (Before)	2.15	20	1.837	.310
	Memory2 (After)	4.95	20	1.317	.294

Paired Sample Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Memory1 – Memory2	**D**	1.196	.268	–3.360	–2.240	– 10.466	**E**	.000

a) Find the values of **D** and **E**. (2 marks)

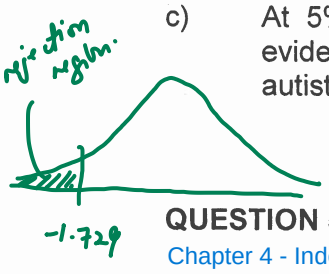
D = 2.15 - 4.95 = -2.8

E = 20 - 1 = 19

b) State the 95% confidence interval for the mean difference and explain. (3 marks)

(-3.36,-2.24). We are 95% confidence that the true population mean difference before and after the treatment (interactive game) lie between the mean score of -3.36 and -2.24.

c) At 5% level of significance, using the value of **t = –10.466,** is there sufficient evidence to indicate that the interactive game has improved the memory of the autistic children? (4 marks)



H0: Interactive game has not improved the memory of autistic children
H1: Interactive game has improved the memory of autistic children
Critical Value = -1.729 (one tail with alpha = 0.05, df = 19)
Decision - Reject null since t = -10.466 < CV = -1.729.
Conclusion: There is enough evidence that interactive game has improved the memory of autistic children

QUESTION 5

Chapter 4 - Independent Sample t test

The Ministry of Education wants to determine whether there is significant difference in the starting salary offer per month between those who hold a bachelor degree from the public and private universities. A survey was conducted on a randomly selected 100 employees and the statistics are shown in the following tables.

Group Statistics

	University	N	Mean	Std. Deviation	Std. Error Mean
Starting Salary Offer (RM)	Public	60	3889.70	32.19	4.156
	Private	40	3813.50	70.39	11.130

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff	Std. Error Diff	95% Confidence Interval of the Difference	
									Lower	Upper
Starting Salary Offer (RM)	Equal variances assumed	19.3	0.79	6.41	98	.001	76.20	11.88	16.20	126.20
	Equal variances not assumed			6.24	81.2	.075	76.20	12.21	20.20	130.20

a) State the null and alternative hypotheses for the above test.

Null Hypothesis: Average starting salary do not differ between those from public or private universities (2 marks)
Alternative Hypothesis: Average starting salary differ between those from public or private universities

b) Based on the Levene's Test, what can be concluded about the equality of variances? Use $\alpha = 0.05$.

H0: Variance are equal; H1: Variance are not equal
Since the Levene's Test p-value = 0.79 > 0.05, we failed to reject the H0. Thus, there is enough evidence that the variance for both group are assumed to be equal (3 marks)

c) Using the *p*-value, do the data provide sufficient evidence to indicate that there is significant difference in starting monthly salary between graduates from public and private universities? Use $\alpha = 0.05$.

Decision: Reject the null hypothesis if p-value < 0.05. Here we have p-value = 0.001 < 0.05. Hence, the null hypothesis was rejected. (3 marks)
Conclusion: There is enough evidence that there is significant difference in starting monthly salary between graduates from public and private university. Based on the result provided, average starting salary for graduates from public university (mean = RM3889.70) is higher than those from private university (mean = RM3813.50)

QUESTION 6

GrabbyGrab wishes to determine whether students or working individuals are more likely to use their service. A sample of 200 GrabbyGrab users was selected at random. The data analyzed using SPSS has produced the following statistics.

Working Status*Frequency Using GrabbyGrab Model Crosstabulation

			Frequency Using GrabbyGrab			Total
			Minimal	Moderate	Frequent	
Working Status	Students	Count	5	**A**	55	95
		Expected Count	23.8	35.6	35.6	95.0
	Workers	Count	45	40	20	105
		Expected Count	26.3	**B**	39.4	105.0
	Total	Count	50	75	75	200
		Expected Count	50.0	75.0	75.0	200.0

A = 75 - 40 = 35
B = (105)(75) / (200) = 39.4
or
75 - 35.6 = 39.4

Chi-Square Tests

	Value	df	Asymp. Sig.(2-sided)
Pearson Chi-Square	48.287	2	.000
Likelihood Ratio	53.625	2	.000
Linear-by-Linear Association	47.586	1	.001
N of Valid Cases	200		

a) Find the values of **A** and **B**. (4 marks)

A = 75 - 40 = 35
B = (105)(75) / (200) = 39.4, or, 75 - 35.6 = 39.4

b) State the hypotheses for this study. (2 marks)

H0: Working status is not related to the frequent usage of GrabbyGrab
H1: Working status is related to the frequent usage of GrabbyGrab

c) Based on the *p*-value, can we conclude that working status is related to the frequent usage of GrabbyGrab? Use $\alpha = 0.01$. (3 marks)

alpha = 0.01
decision: reject the null hypothesis if p-value < alpha. We reject the null hypothesis since the resulted p-value < 0.001 and less than the alpha value of 0.01.
conclusion: Therefore, we have enough evidence that working status is related to the frequent usage of GrabbyGrab.

QUESTION 7

Chapter 5 - Bivariate Analysis (Correlation and SLR)

The higher management of Spa Ayu is interested to assess the impact of advertising by looking at the monthly advertising cost (RM thousands) and the monthly income (RM millions) for eight consecutive months. The data were recorded and analyzed using SPSS. The results are as follows:

150,000 30,000 20,000 4,000,000

Monthly Advertising Cost (RM thousands)	150	120	100	80	50	40	30	20
Monthly Income (RM millions)	20	15	10	7	4	6	5	3

$\Sigma x = 590$
$\Sigma y = 70$
$\Sigma x^2 = 58,700$
$\Sigma y^2 = 860$
$\Sigma xy = 7010$
$n = 8$

Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig
1	(Constant)	-.221	1.354		-.164	.875
	Advertising cost	.122	.016	.953	7.697	.000

a. Dependent Variable: Monthly Income

a) Show that the Pearson's correlation coefficient is 0.953 and comment on the value. (4 marks)

$r = \frac{\Sigma xy - \Sigma x \Sigma y / n}{\sqrt{\cdots}} = \frac{7010 - (590)(70)/8}{\sqrt{\cdots}} = 0.9529 \approx 0.953$ (show)

b) Calculate the coefficient of determination. Interpret the value obtained. (3 marks)

r-square = 0.908.
Interpretation: 90.8% variations in monthly income can be explained by the variations in monthly advertizing cost.

c) State the slope and interpret the value in the context of the problem. (1 mark)

slope = 0.122
Interpretation: For each additional increase in monthly advertising cost by RM1000, the monthly income will increase by 0.122 (RM millions)

d) Estimate the monthly income (RM millions) if the monthly advertising cost is RM90,000. (2 marks)

x = 90
y = -0.221 + 0.122(90) = 10.759 (RM millions) @ RM10,759,000

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

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