UTHM
Universiti Tun Hussein Onn Malaysia

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

FINAL EXAMINATION
SEMESTER II
SESSION 2014/2015

COURSE NAME	:	DATA ANALYSIS
COURSE CODE	:	BWA 21003
PROGRAMME	:	2 BWA
EXAMINATION DATE	:	JUNE 2015/JULY 2015
DURATION	:	3 HOURS
INSTRUCTION	:	ANSWER **ALL** QUESTIONS

THIS QUESTION PAPER CONSISTS OF **NINE (9)** PAGES

Q1 (a) Explain two differences between primary data and secondary data.
(4 marks)

(b) List 5 steps of statistical analysis.
(5 marks)

(c) Given the SPSS output of '*mpg*' as presented in **Table Q1(c)** below.

Table Q1(c) Descriptive summary of '*mpg*'

			Statistic	Std. Error
mpg	Mean		20.9231	.93302
	95% Confidence Interval for Mean	Lower Bound	19.0015	
		Upper Bound	22.8447	
	5% Trimmed Mean		20.6026	
	Median		21.0000	
	Variance		22.634	
	Std. Deviation		4.75750	
	Minimum		14.00	
	Maximum		35.00	
	Range		21.00	
	Interquartile Range		6.25	
	Skewness		.935	.456
	Kurtosis		1.793	.887

(i) Interpet the skewness of '*mpg*'.
(2 marks)

(ii) Describe the pattern of kurtosis.
(2 marks)

(iii) Interpret the 95% confidence interval of mean.
(2 marks)

(d) Analyse the box plot in **Figure Q1(d)**. Do the diagnostic checking and give your recommendations.

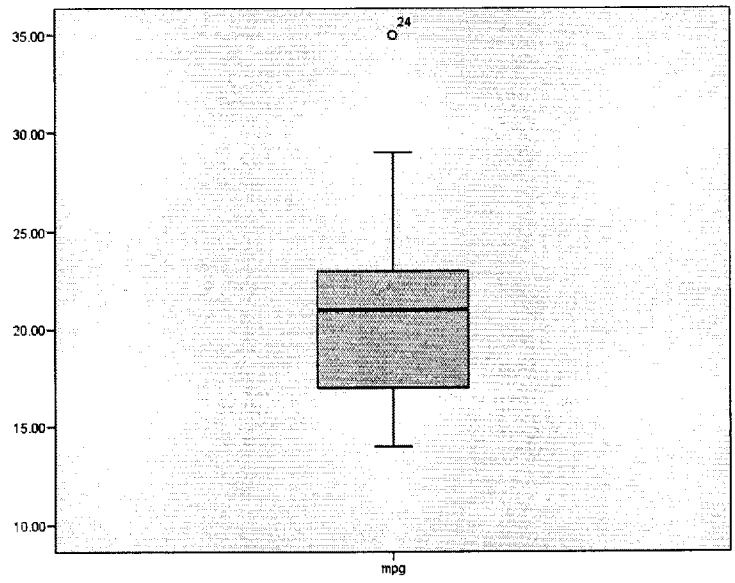


Figure Q1(d) Box Plot of '*mpg*'

(5 marks)

Q2 (a) In testing the hypotheses $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$, both F-test and t-test can be used. Show that both tests are algebraically equivalent.

(4 marks)

(b) Verify that the total sum of squares (SST) is a decomposition of error sum of squares (SSE) and regression sum of squares (SSR) by showing that the left hand side of the following equation is the same as the right hand side.

$$\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$$

$$\left[\text{Hint: } \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left[\left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)\right]^2\right]$$

(6 marks)

(c) A study was conducted to determine the relationship between starting salaries (RM thousands) for recent statistics graduates and their grade point averages in the major course. A linear regression model was fitted to the data and the estimates regression function was obtained. Part of the computer output for the above analysis is given below in **Table Q2(c)**:

Table Q2(c) MINITAB computer output

ANOVA					
Model	Sum of Squares	df	Mean Square	F	Sig.
Regression		1		147.28	.000
Error	734.9		40.828		
Total	6748.2				

Coefficients

Model	Unstandardized Coefficients		t	Sig.
	β	Std. Error		
Constant	-8.42	3.395	-2.48	0.011
GPA	3.007	0.2477	12.14	0.000

(i) Complete the ANOVA in **Table Q2(c)**.

(4 marks)

(ii) Write the estimated regression function.

(2 marks)

(iii) Determine the coefficient of determination for the model and interpret its meaning.

(4 marks)

Q3 (a) Give your comments of the following statements.

(i) Value of R^2 is close to 1.

(2 marks)

(ii) Value of correlation coefficient is close to zero.

(2 marks)

(b) The following table shows the mean speeds of 12 motorcycles and the amount of traffic fines (RM) they pay during a period of a month.

Table Q3(b) Mean speeds and traffic fines

Mean Speed (km/h)	Trafic fines (RM)
100	300
120	350
115	320
112	315
108	310
105	305
116	340
121	360
125	380
118	345
115	320
122	330

Assuming that both mean speeds and the traffic fines are normally distributed.

(i) Obtain the linear regression of the fines against the mean speeds.

(5 marks)

(ii) At the signficant level of 0.05, test the hypothesis that there exists a positive relation of the fines against the mean speeds.

(7 marks)

(iii) Find the Pearson correlation coefficient between the mean speeds and fines. Interpet your finding.

(4 marks)

Q4 (a) A QC engineer is testing a power supply used in producing notebook components. The complete table of observed frequencies is as follows:

Table Q4(a) Frequency table of power supply

Class interval	Observed frequencies O_i
$x < 4.948$	12
$4.948 \leq x < 4.986$	14
$4.946 \leq x < 5.014$	12
$5.014 \leq x < 5.040$	13
$5.040 \leq x < 5.066$	12
$5.066 \leq x < 5.094$	11
$5.094 \leq x < 5.132$	12
$x \geq 5.132$	14

Test whether the output voltage is adequately described by a normal distribution with mean 5.04V and standard deviation 0.08V at a significance level of 0.05.

(10 marks)

(b) Then, the QC engineer took a set of sample data to determine whether the proportions of output of notebook components for two shifts produced by machine A, B and C were the same. The following data were collected:

Table Q4(b) Work shifts and machines

Shift	Machine		
	A	B	C
1	100	120	180
2	120	180	100

Use a 0.05 level of significance to determine if the proportions of components for shift 1 are the same for all three machines.

(10 marks)

Q5 Three set of five mice were randomly selected to be placed in a standard maze but with different color doors. The response is the time required to complete the maze as seen below. Refer to the SPSS output as given in Appendix 1.

Table Q5 Time required to complete the maze

Colour	Time				
Red	9	11	10	9	15
Green	20	21	23	17	30
Black	6	5	8	14	7

(a) Briefly explain on the normality assumption based on Normal P-P plot of maze time for each different colour doors.

(3 marks)

(b) Conduct a test whether the variances are equal by using 0.01 level of significant.

(4 marks)

(c) Perform the appropriate analysis to test if there is an effect due to door colours. Then, if there is a difference in the mean times to complete the maze based on the door colours, perform Tukey-Kramer multiple comparison test. Use 0.01 level of significant for both analysis.

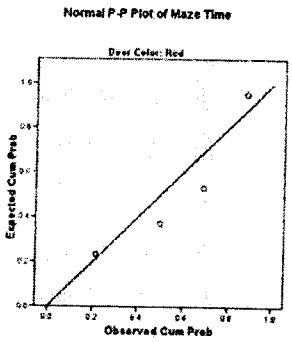
(13 marks)

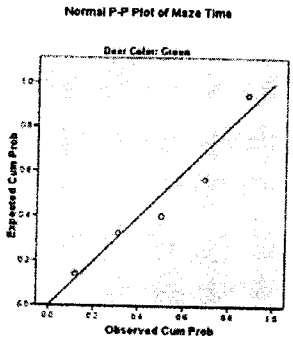
- END OF QUESTION -

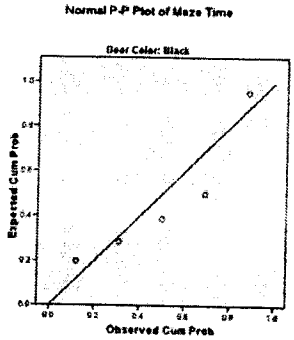
FINAL EXAMINATION

SEMESTER / SESSION : SEM II / 2014/2015
COURSE : DATA ANALYSIS
PROGRAMME : 2 BWA
COURSE CODE : BWA 21003

APPENDIX 1(1)







Oneway

Test of Homogeneity of Variances

Maze Time

Levene Statistic	df1	df2	Sig.
.6522	2	12	.5384

ANOVA

Maze Time

	df	Sum of Squares	Mean Square	F	Sig.
Between Groups	2	565.7333	282.8667	20.0142	.0002
Within Groups	12	169.6000	14.1333		
Total	14	735.3333			

APPENDIX 1(2)

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Maze Time

Tukey HSD

(I) Door Color	(J) Door Color	Mean Difference (I-J)	Std. Error	Sig	99% Confidence Interval	
					Lower Bound	Upper Bound
Red	Green	-11.4000*	2.3777	.0012	-19.8836	-2.9164
	Black	2.8000	2.3777	.4879	-5.6836	11.2836
Green	Red	11.4000*	2.3777	.0012	2.9164	19.8836
	Black	14.2000*	2.3777	.0002	5.7164	22.6836
Black	Red	-2.8000	2.3777	.4879	-11.2836	5.6836
	Green	-14.2000*	2.3777	.0002	-22.6836	-5.7164

*. The mean difference is significant at the .01 level.

Homogeneous Subsets

Maze Time

Tukey HSD[a]

Door Color	N	Subset for alpha = .01	
		1	2
Black	5	8.0000	
Red	5	10.8000	
Green	5		22.2000
Sig.		.4879	1.0000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 5.000.