(Knowledge for Development)

KIBABII UNIVERSITY

UNIVERSITY EXAMINATIONS

2017/2018 ACADEMIC YEAR

THIRD YEAR SECOND SEMESTER

MAIN EXAMINATION

FOR THE DEGREE OF BACHELOR OF SCIENCE MATHEMATICS

COURSE CODE: STA 348

COURSE TITLE: STATISTICAL COMPUTING

DATE: 08/08/18 **TIME:** 9 AM -11 AM

INSTRUCTIONS TO CANDIDATES

Answer Question One and Any other TWO Questions

TIME: 2 Hours

This Paper Consists of 7 Printed Pages. Please Turn Over.

QUESTION ONE (30 MKS)

Titanic - Details for passengers travelling on the Titanic when it sank:

Name	class	Survived 0 = died	Country of residence	Gender	Age	No. of siblings/ spouses on board	No. of parents/ children on board	price of ticket
Abbing, Anthony	3	0	USA	male	42	0	0	7.55
Abbott, Rosa	3	1	USA	female	35	1	1	20.25
Abelseth, Karen	3	1	UK	female	16	0	0	7.65

a) What variable type is each of the variables in the table above?(8 mks)
b) Which variables would you use to investigate the research question: 'Were wealthy people more likely to survive the sinking of the Titanic'?(4 mks)

(c).Prepare an R script that can generate the mean, standard deviation and correlation

Coefficients of variables in a data set. (6 marks)

(d) Describe how you can generate **Histograms** and **Box plots** in R Programming **(** 4marks**)**
(e) Suppose you have the following Research Question

- To what extent does weight of a car in pounds predict miles per gallon in a U.S. dataset of 398 models of cars?

The regression SPSS output of the above question was as below. Use it to answer the following questions

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.807[a]	.651	.650	4.622

a. Predictors: (Constant), Vehicle Weight (lbs.)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15794.632	1	15794.632	739.503	.000[a]
	Residual	8457.943	396	21.358		
	Total	24252.575	397			

a. Predictors: (Constant), Vehicle Weight (lbs.)

b. Dependent Variable: Miles per Gallon

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	45.492	.841		54.110	.000
	Vehicle Weight (lbs.)	-.007	.000	-.807	-27.194	.000

a. Dependent Variable: Miles per Gallon

i) Write out the model regression equation, state the assumptions assumed

ii) How much variance in Miles Per Gallon is explained by Car Weight?

iii) Is this variance explained significantly different to 0?

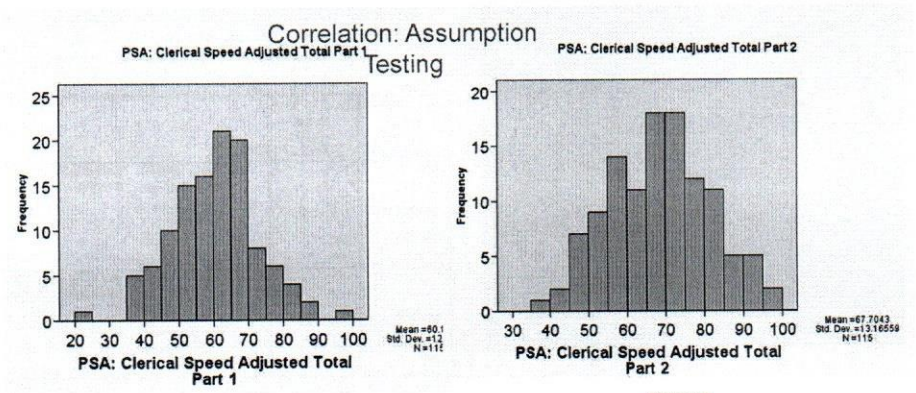
iv) Is the slope statistically significant? Interpret the result

v) If a car weighed 1000 pounds, what would be the predicted miles per gallon?

vi) What would be the approximate 95% confidence interval of our prediction

QUESTION TWO(20 MKS)

One of the ability tests was called the Clerical Speed and Accuracy test, Each Question involved selecting a two-letter pair out of five two-letter pairs that corresponded to the two-letter pair in the question booklet. The test was timed and split into two halves consisting of 100 questions each and lasting 3 minutes each

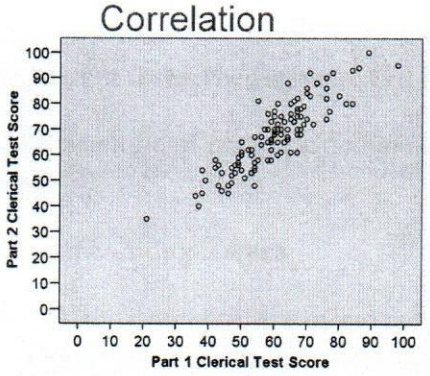


i. What are the two assumptions of correlation required in order to obtain accurate p-values for the significance test?
ii. In the histogram, does performance on the two variables look normally distributed?
iii. What is skewness / se skew for the two variables and does? Is it larger than 3 for either variable?

Statistics

		PSA: Clerical Speed Adjusted Total Part 2	PSA: Clerical Speed Adjusted Total Part 1
N	Valid	115	115
	Missing	1	1
Skewness		.075	.043
Std. Error of Skewness		.226	.226
Kurtosis		-.368	.775
Std. Error of Kurtosis		.447	.447

)



Correlations

		PSA: Clerical Speed Adjusted Total Part 2	PSA: Clerical Speed Adjusted Total Part 1
PSA: Clerical Speed Adjusted Total Part 2	Pearson Correlation	1	.882
	Sig. (2-tailed)		.000
	Sum of Squares and Cross-products	19759.948	16207.026
	Covariance	173.333	142.167
	N	115	115
PSA: Clerical Speed Adjusted Total Part 1	Pearson Correlation	.882	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	16207.026	17072.487
	Covariance	142.167	149.759
	N	115	115

i. Looking at the scatterplot, describe the relationship in terms of direction, form, and degree.
ii. What is the largest absolute value a correlation can take?
iii. Can we justifiably say that part 1 test score caused their part 2 test score?
iv. The sample was made up of roughly 25% first year psychology students and 50% friends of one of the researchers typically aged around 22 and 25% parents of friends of the researchers, do you think restricted range might be an issue here?
v. Are there any extreme scores?
vi. If instead of using the raw scores, we used the rank of each participant, and performed a correlation on these ranks, what type of correlation would we have?
vii. If we split part 1 scores into high and low groups and correlated it with the raw version of part 2, what type of correlation would we have?
viii. Is the correlation statistically significant? What is our df, what is critical r (two-tailed) in G&W?

QUESTION THREE (20 MKS)

(a).A function Y is given by y=sin (x) and another function z is given by z= cos(x)

Write m.scripts that can be executed in MATLAB to generate corresponding sine and cosine waves respectively (6 marks).

(b).Discuss logical and relational operators as used in programming (4 marks)

(d).The correlation coefficients of rows and columns in matrix B can be given by:

1.0000 -0.3149 -0.9685 0.4575

-0.3149 1.0000 0.1325 -0.2294

-0.9685 0.1325 1.0000 -0.5774

0.4575 -0.2294 -0.5774 1.0000

(i) . Describe how you can generate such correlation coefficients from matrix B using a MATLAB command .Explain statistical interpretations and importance of three sets of columns whose correlation coefficients are 1,0 and -1 respectively (10 marks)

QUESTION FOUR(20 MKS)

A market researcher is interested in the coffee drinking habits of males and females. He asks a sample of male and female office workers to record the number of cups of coffee they consume during a week.

a) Which parametric statistical technique could the researcher use to determine if males and females differ in terms of the number of cups of coffee consumed in a week? Justify your answer and describe how you would obtain this statistic using SPSS.

b) What are the key values you would look for in the output?

c) Explain the assumptions you should check for when using the technique that you chose in question (a), above.

d) What non-parametric technique could be used to address this research question?

QUESTION FIVE(20 MKS)

Consider the following SPSS output

Tests of Between-Subjects Effects

Dependent Variable: TEP: Average Text Editing Performance (Seconds) ANOVA Table

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	4371[a]	2	2185.5	16.2	.000
Intercept	114650	1	114650.5	849.3	.000
abilitycluster	4371	2	2185.5	16.2	.000
Error	9720	72	135.0		
Total	128741	75			
Corrected Total	14091	74			

a. R Squared = .310 (Adjusted R Squared = .291)

Which three rows correspond to our typical: between, within and total sums of squares?
If you did not know the sample size or number of groups, how could you work this out from this table?
What is the overall text editing standard deviation for the entire sample?
What two numbers can be added to give us corrected total?
What two numbers are used to obtain between treatments mean square?
What two numbers are used to obtain within mean square?
What two numbers are used to obtain the F value?
What is critical F at .05? What about .01? What effect does decreasing alpha have on critical F?
What effect does decreasing alpha have on statistical power?
If the sample size was 203, what would critical F .05 be? What about a sample size of 18?
What effect does increasing sample size have on critical F? What effect does it have on power
What if there were 4 groups? What if there were 5 groups? What would critical F be? What effect does increasing the number of groups have on critical F?
Is the obtained F statistically significant at .05? What about .01?
Can we conclude that all three groups differ? If not, what could we do to find out?

TEP: Average Text Editing Performance (Seconds)

Tukey HSD [a,b]

Ability Cluster	N	Subset 1	Subset 2
2 Higher Ability Group	25	31.1865	
3 Lower General Ability / Higher Psychomotor Abilty Group	25	36.6923	
1 Lower Ability Group	25		49.4159
Sig.		.222	1.000

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 134.997

Post hoc tests

1. What would Tukey's HSD .05 be using the exam formula sheet?
2. Which pairs of groups differ by more than this amount?
3. Looking at the homogenous subsets output, which groups are meant to be homogenous?

Multiple Comparisons

Dependent Variable: TEP: Average Text Editing Performance (Seconds)
Tukey HSD

(I) Ability Cluster	(J) Ability Cluster	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1 Lower Ability Group	2 Higher Ability Group	18.2295*	3.28629	.000	10.3650	26.0940
	3 Lower General Ability / Higher Psychomotor Abilty Group	12.7236*	3.28629	.001	4.8591	20.5881
2 Higher Ability Group	1 Lower Ability Group	-18.2295*	3.28629	.000	-26.0940	-10.3650
	3 Lower General Ability / Higher Psychomotor Abilty Group	-5.5059	3.28629	.222	-13.3704	2.3586
3 Lower General Ability / Higher Psychomotor	1 Lower Ability Group	-12.7236*	3.28629	.001	-20.5881	-4.8591
	2 Higher Ability Group	5.5059	3.28629	.222	-2.3586	13.3704

Based on observed means