(Knowledge for Development)

KIBABII UNIVERSITY

UNIVERSITY EXAMINATIONS

2019/2020 ACADEMIC YEAR

THIRD YEAR SECOND SEMESTER

SPECIAL/SUPPLEMENTARY EXAMINATIONS

FOR THE DEGREE OF BACHELOR OF SCIENCE (INFORMATION TECHNOLOGY)

COURSE CODE: BIT 323

COURSE TITLE: DATA ANALYSIS TECHNIQUES

DATE: 03/02/2021 **TIME:** 11.00A.M. -1.00 P.M.

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 COMPULSORY [30 MARKS]

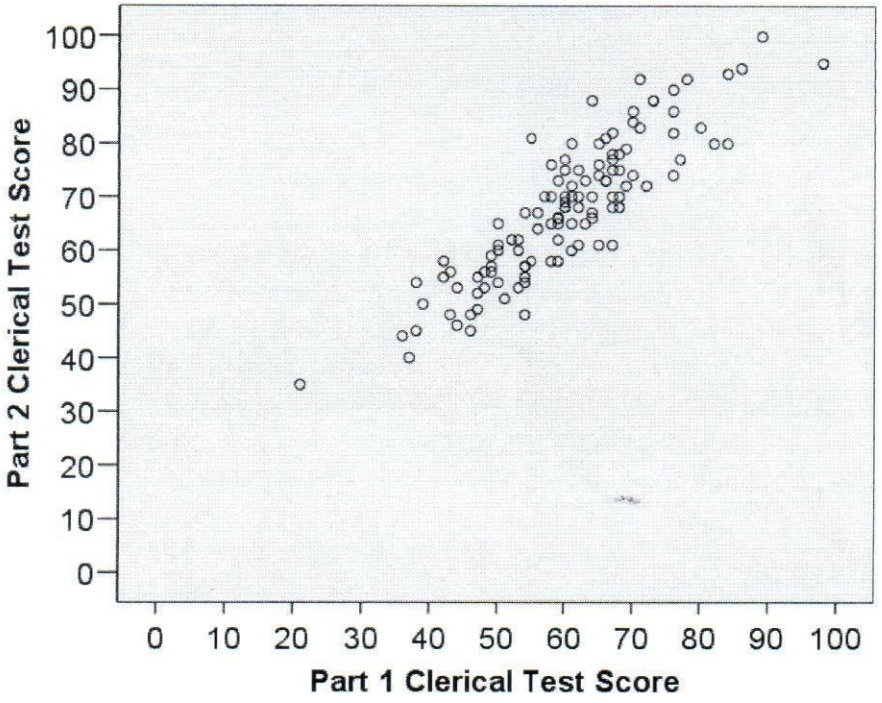
a) Briefly describe how you can carry out the following procedure using SPSS
   (i) Descriptive
   (ii) Cross-tabulation
   (iii) Regression

b) When talking about statistical inference, what is alpha? What are the typical sizes of alpha? What does it mean?

c) You want to test the following hypotheses. What type of statistical test do you run? Explain the procedure of running out test in SPSS.

i) Research hypothesis #1: Males have a higher mean level of income than females. Income is measured in thousands of dollars earned per year.
ii) Research hypothesis #2: Males are more likely to be alcoholics than females. "Alcoholic" is coded "yes/no."

d) Consider the following SPSS output



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 TWO (20 MARKS)

The SPSS output sheets for this question shows a crosstabulation of the variables SEX (gender) and **Illegal Drugs (Marijuana, Cocaine)** , taken from a random sample of Americans. Both variables are coded as shown in the table. Using the information presented to you, do the following:

Respondent's Sex * Illegal Drugs (Marijuana, Cocaine) Crosstabulation

Count

		Illegal Drugs (Marijuana, Cocaine)		Total
		Yes	No	
Respondent's Sex	Male	23	393	416
	Female	7	589	596
Total		30	982	1012

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	16.149[a]	1	.000		
Continuity Correction[b]	14.670	1	.000		
Likelihood Ratio	16.190	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	16.133	1	.000		
N of Valid Cases	1012				

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	.126	.029	4.047	.000[c]
Ordinal by Ordinal	Spearman Correlation	.126	.029	4.047	.000[c]
N of Valid Cases		1012			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

i. Identify the independent and dependent variable.
ii. State a plausible research hypothesis for these two variables.
iii. State the null hypothesis.
iv. Test the null hypothesis and make a decision, reporting the information you used to make that decision.

(e) interpret the results

QUESTION THREE [20 MARKS]

a) 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?

i) Are the two variables discrete or continuous?

ii) Are the two variables nominal, ordinal, interval or ratio scales?

iii) Which statistical procedure could we use to test the research question?

iv) What is the null hypothesis?

v) What is your expectation?

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

ii) Is this variance explained significantly different to 0?

iii) What is the constant

iv) What is the slope?

v) Is the slope statistically significant?

vi) Write out the model regression equation

vii) What is the standardised regression coefficient for vehicle weight?

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

ix) What is the standard error of the estimate?

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

QUESTION FOUR [20MARKS]

Sample questionnaire

1. Sex ___ Male ___ Female

2. Age in years _____

3. Education level (please indicate the highest level of schooling that you completed)

_____ Year 10 _____ Year 12 _____ University or College

4. Are you currently on a diet to lose weight? ____ Yes ____ No

Please indicate how much you either agree or disagree with each of the following statements.

Write a number from 1 to 4 on the line next to each statement.

strongly disagree 1 2 3 4 strongly agree

1. _____ I have little control over the things that happen to me
2. _____ I can do just about anything I really set my mind to
3. _____ There is really no way I can solve some of the problems I have
4. _____ There is little I can do to change many of the important things in my life
5. _____ What happens to me in the future mostly depends on me
6. _____ I often feel helpless in dealing with the problems of life
7. _____ Sometimes I feel that I'm being pushed around in life

a) Prepare a codebook for this questionnaire, detailing each of the variable names and codesto be used to prepare the data for entry into SPSS.
b) Using the codebook you developed, describe how you would create value labels for Question 3 (Education level). Describe the steps you would use in SPSS.

c) Using SPSS, create a new data file for this questionnaire. Enter some pretend data in this data file (for at least five cases).

QUESTION FIVE [20 MARKS]

a) The term "Vector" can have a mathematical Definition or as a data type. Explain (4 marks)

b) Discuss inputs in R and how they can be **read** in R (10 marks)

c) The weight of children at Kibabii clinic is between 2.1: 5 kilograms. Explain how you can generate a Histogram from the data set (3 marks)

d. What are "Box plots" and how are they used in statistics? (3 Marks)

Illustrate the use of R programming to generate box plots (2 Marks)