(Knowledge for Development)

KIBABII UNIVERSITY

UNIVERSITY EXAMINATIONS
2017/2018 ACADEMIC YEAR

END OF SEMESTER EXAMINATIONS
SPECIAL/SUPPLEMENTARY EXAM

FOR THE DEGREE OF
BACHELOR OF SCIENCE COMPUTER SCIENCE

COURSE CODE : CSC 360E

COURSE TITLE : DATA ANALYSIS TECHNIQUES

DATE: 4/10/2018 TIME: 3.00PM – 5.00PM

INSTRUCTIONS TO CANDIDATES

ANSWER QUESTIONS ONE AND ANY OTHER TWO.

Question 1 (30 marks)

a) A recent poll asked 1,245 registered voters nationwide, "Which party do you think can do a better job of dealing with corruption in Kenya?"31% of the respondents answered "NASA coalition". With a margin of error of±3%, it is estimated that between 28 and 34 percent of registered voters nationwide feel that NASA can do a better job of handling corruption. Identify which part of this example is inferential. [4 marks]

b) Define the terms population and sample. [2 marks]

c) Write Python code to build a word frequency table of sentence stored in the variable `sentence`, using a dictionary object. [6 marks]

d) Explain the difference between the mean, median and mode. [3 marks]

Use the following table to answer questions e) - h)

AGE OF RESPONDENT

Age	Frequency
28	21
29	49
30	52
31	54
32	74
Total	250

e) What is the mode? [2 mark]

f) What is the median? [2 mark]

g) What is the mean age? [2 mark]

h) Is this distribution normal, negatively skewed, or positively skewed? How did you determine this? [4 marks]

i) A numerical summary of a population is called _____while a numerical summary of a sample is called_____. [2 marks]

j) With the help of an example, explain the difference between a categorical variable and a quantitative variable. [3 marks]

Question 2 (20 marks)

A researcher is investigating the association between blood pressure and "workaholism" in a certain population. She classifies someone who works more than 60 hours per week as a workaholic. She records the income level and blood pressure (high or normal) for each participant and whether or not they can be classified as a "workaholic". The data are summarized in the table below.

	Income Group					
	Low		Middle		High	
Workaholic	HBP	NBP	HBP	NBP	HBP	NBP
Yes	25	75	23	87	26	134
No	25	80	18	72	9	51

a) Consider each income group separately and find the proportion with high blood pressure among

i. the workaholics [4 marks]

ii. non-workaholics [4 marks]

b) Determine

i. the proportion of workaholics overall who have high blood pressure. [4 marks]

ii. the proportion of non-workaholics who have high blood pressure. [4 marks]
(ignoring information about income group in both cases)

c) How would you explain what is responsible for this result? [4 marks]

Question 3 **(20 marks)**

a) Write a Python program to open the file `romeo.txt` and read it line by line. For each line, split the line into a list of words using the `split` function. [8 marks]

b) For each word in question (a) above, check to see if the word is already in a list. If the word is not in the list, add it to the list. [6 marks]

c) When the program completes, sort and print the resulting words in alphabetical order. [6 marks]

Question 4 **(20 marks)**

a) A sample of randomly selected doctors was asked to indicate the category that best described how often they used the Internet. The results follow.

Internet Usage Pattern	Count
Never	31
Rarely(about 3 times per year)	15
Occasionally	52
Often	109
Daily	117

i. Construct a pie chart for these data [8 marks]

ii. In creating a bar graph of these data, would it be more useful to list the patterns as given in the table above or in the order of a Pareto? [4 marks]

b) Construct a frequency table using the following ages: 30, 30, 30, 33, 33, 34, 37, 37, 37, 37, 37, 40, 40, 45, 45, 45, 45, 45 [8 marks]

Question 5 **(20 marks)**

Illustrate the use of R to compute the following from a data set:

a) Mean [4 marks]

b) Correlation Coefficients between two variables [4 marks]

c) Give Statistical Explanation and significance of correlation Analysis [5 marks]

d) Design a script that can be run to generate time series graph [7 Marks]