(Knowledge for Development)

KIBABII UNIVERSITY

(KIBU)

UNIVERSITY EXAMINATIONS
2020/2021 ACADEMIC YEAR

END OF SEMESTER EXAMINATIONS
YEAR TWO SEMESTER TWO EXAMINATIONS

FOR THE DEGREE OF
BACHELOR OF SCIENCE
(INFORMATION TECHNOLOGY)

COURSE CODE : BIT 225

COURSE TITLE : DATA ANALYSIS TECHNIQUES

DATE: 05/10/2021 **TIME: 9.00 A.M-11.00 A.M**

INSTRUCTIONS TO CANDIDATES

ANSWER QUESTIONS ONE AND ANY OTHER TWO.

QUESTION ONE (COMPULSARY) [30 MARKS]

1. With which of the data classes Nominal, Ordinal, Interval/Ratio can a mean be used?
   A. Interval/Ratio B. Nominal, Ordinal, and Interval/Ratio
   C. Ordinal and Interval/Ratio D. None of the above
2. With which of the data classes Nominal, Ordinal, Interval/Ratio can a mode be used?
   A. Interval/Ratio B. Nominal, Ordinal, and Interval/Ratio
   C. Ordinal and Interval/Ratio D. None of the above
3. With which of the data classes Nominal, Ordinal, Interval/Ratio can a media be used?
   A. Interval/Ratio B. Nominal, Ordinal, and Interval/Ratio
   C. Ordinal and Interval/Ratio D. None of the above
4. Measures of central tendency are
   A. inferential statistics that identify the best single value for representing a set of data.
   B. descriptive statistics that identify the best single value for representing a set of data.
   C. inferential statistics that identify the spread of the scores in a data set.
   D. descriptive statistics that identify the spread of the scores in a data set.
5. The mean is
   A. the statistical or arithmetic average.
   B. the middlemost score.
   C. the most frequently occurring score.
   D. the best representation for every set of data.
6. A measure of central tendency tells us, using a single value, the best representation for an entire set of scores. A measure of variability tells us
   A. if the high and low extreme scores cancel each other out.
   B. if the mean is greater than the mode.
   C. how well the measure of central tendency represents the entire set of scores.
   D. whether or not to compute percentiles.
7. The sum of squared deviations is a good measure of variability except
   A. it uses the mean in its calculation. B. it can only be used with ratio data.
   C. it does not take N into account. D. it can not be used in any other statistical analyses.
8. The mean of ten numbers is 58. If one of the numbers is 40, what is the mean of the other nine?
   A. 18 B. 60 C. 162 D. 540
9. If you are comparing the number of girls to the number of boys in a classroom, you are gathering
   A. Interval Data B. Nominal data C. Ratio data D. Ordinal data
10. What kind of data is a temperature reading?
    A. Interval Data B. Nominal data C. Ratio data D. Ordinal data
11. The order that runners cross the finish line is an example of:
    A. a nominal scale B. an ordinal scale C. an interval scale D. a ratio scale
12. A specific characteristic of a population is known as a
    A. sample B. parameter. C. statistic. D. descriptive statistic.
13. Both descriptive and inferential statistics are used to change data into information which in turn is converted into _____ that leads to better decision making.
    A. knowledge B. a process C. a forecast D. a parameter.
14. An observed set of the population that has been selected for analysis is called
    A. a sample B. a process C. a forecast D. descriptive statistic.
15. Inferential statistics is a process that involves all of the following EXCEPT
    A. estimating a parameter B estimating a statistic C. test a hypothesis D. analyse relationships
16. What does a significant test statistic tell us?
    A. There is an important effect
    B. That the test statistic is larger than we would expect if there were no effect in the population.
    C. The hull hypothesis is false D. All of the above
17. Which of the following statements is true?
    A. The standard error is calculated solely from sample attributes.
    B. The standard deviation is calculated only from sample attributes.
    C. The standard error is a measure of central tendency.

D. All of the above.

18. If you drew all possible samples from some population, calculated the mean for each of the samples, and constructed a line graph (showing the shape of the distribution) based on all of those means, what would you have?
    A. A population distribution B. A sample distribution C.
    C. A sampling distribution D. A parameter distribution
19. What does it mean when you calculate a 95% confidence interval?
    A. The process you used will capture the true parameter 95% of the time in the long run
    B. You can be "95% confident" that your interval will include the population parameter
    C. You can be "5% confident" that your interval will not include the population parameter
    D. All of the above statements are true
20. What is the difference between a bar chart and a histogram?
    A. Bar charts represent numbers, whereas histograms represent percentages.
    B. A histogram does not show the entire range of scores in a distribution.
    C. There are no gaps between the bars on a histogram.
    D. Bar charts are circular, whereas histograms are square.
21. What does the term 'outlier' mean?
    A. A score that is left out of the analysis because of missing data
    B. The arithmetic mean
    C. A type of variable that cannot be quantified
    D. An extreme value at either end of a distribution
22. Which of the following is NOT a technique that is commonly used to do data cleaning?
    A. Frequency Analyses B. Analysis of variance C. Cross-tabulations D. Plots
23. What does the operation "Recode Into Different Variables" do to the data?
    A. Replaces missing data with some random scores.
    B. Reverses the position of the independent and dependent variable on a graph.
    C. Redistributes a range of values into a new set of categories and creates a new variable.
    D. Represents the data in the form of a pie chart.

For the next 4 questions: The simple linear regression equation can be written as $\hat{y} = b_0 + b_1 x$

24. In the simple linear regression equation, the symbol $\hat{y}$ represents the
    A. average or predicted response B. estimated intercept
    C. estimated slope D. explanatory variable
25. In the simple linear regression equation, the term b_0 represents the
    A. estimated or predicted response B. estimated intercept
    C. estimated slope D. explanatory variable
26. In the simple linear regression equation, the term b_1 represents the
    A. estimated or predicted response B. estimated intercept
    C. estimated slope D. explanatory variable
27. In the simple linear regression equation, the symbol x represents the
    A. estimated or predicted response B. estimated intercept
    C. estimated slope D. explanatory variable
28. A scatterplot is a
    A. one-dimensional graph of randomly scattered data.
    B. two-dimensional graph of a straight line.
    C. two-dimensional graph of a curved line. D. two-dimensional graph of data values.
29. Two variables have a positive association when
    A. the values of one variable tend to increase as the values of the other variable increase.
    B. the values of one variable tend to decrease as the values of the other variable increase.
    C. the values of one variable tend to increase irregardless of how the values of the other variable change.
    D. the values of both variables are always positive.
30. A researcher reports that the correlation between two quantitative variables is $r = 0.8$. Which of the following statements is correct?
    A. The average value of y changes by 0.8 when x is increased by 1.
    B. The average value of x changes by 0.8 when y is increased by 1.

Work out the value of χ 2 (Chi Square). $\chi^2 = \sum \frac{(f_e - f_o)^2}{f_e}$ **[6 mark**

d. Explain three important characteristics of a Chi Square Test. **[6 marks]**

QUESTION FIVE [20 MARKS]

a. The following table shows the relationship between the number of workers and the amount of time in minutes it takes them to harvest the sugar cane in a particular field.

Workers (n)	3	4	5	6	8	9	10
Time (t)	799	703	645	570	422	322	241

i. Find the equation of the regression line of t on n. **[3 marks]**

ii. Find the value of the Pearson's product–moment correlation coefficient, r. **[3 marks]**

iii. Use the regression equation to find how long it would take seven workers to harvest the sugar cane. **[2 marks]**

b. Compare and contrast Correlation and Simple Linear Regression. **[6 marks]**

c. Compare and contrast ANOVA and T Test. **[6 marks**