KIBABII UNIVERSITY

*KNOWLEDGE FOR DEVELOPENT*

# UNIVERSITY EXAMINATIONS
# 2020/2021 ACADEMIC YEAR

## END OF SEMESTER EXAMINATIONS
## YEAR THREE SEMESTER TWO EXAMINATIONS

## FOR THE DEGREE OF
## (COMPUTER SCIENCE)

**COURSE CODE** : 362E

**COURSE TITLE** : **DATA ANALYSIS TECHNIQUES**

**DATE:** 08/10/2021 **TIME:** 02.00 P.M – 04.00 P.M

**INSTRUCTIONS TO CANDIDATES**

ANSWER QUESTIONS ONE AND ANY OTHER TWO.

**QUESTION ONE** (COMPULSORY) **[30 MARKS]**

a) Name the three main components of a statistical study. [3 marks]

b) The following statement refers to which aspect of a statistical study: "The average age of the students in a statistics class is 25 years"? [1 marks]

c) Drug X, a drug that claims to treat male pattern scalp hair loss (baldness), was administered for 12 months to over 1800 men aged 18 to 41 with mild to moderate amounts of ongoing hair loss. Whether they were receiving drug X or a placebo (a pill containing no medication), all men were given a medicated shampoo. In general, men who took drug X maintained or increased the number of visible scalp hairs; while scalp hair counts in men who took the placebo continued to decrease. This concluded that drug X is effective in maintaining or increasing the amount of scalp hair in men.

i. Which statement of this example can be referred to as descriptive statistics? [2 marks]

ii. Which statement can be characterized as inferential statistics? [2 marks]

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

e) Briefly explain why survey data typically consists of data for a sample and not for an entire population. [1 mark]

f) The relationship between the number of games won by a league football team and the average attendance at their home games is analyzed. A regression to predict the average attendance from the number of games won has an r= 0.73. Interpret this statistic. [2 marks]

g) Almost all of the acidity of soda pop comes from the phosphoric acid which is added to give them a sharper flavor. Is there an association between the pH of the soda and the amount of phosphoric acid (in grams)? The correlation between pH and phosphoric acid is -0.991. Describe the association. [2 marks]

h) Identify the values of the y-intercept a and the slope b of the following regression line.

$$\hat{y} = 4 - x$$ [2 marks]

i) A psychologist does an experiment to determine whether an outgoing person can be identified by his or her handwriting. She claims that the correlation of 0.89 shows that there is a strong causal relationship between personality type and handwriting. Explain what is wrong with her interpretation. [2 marks]

j) The test scores of 15 students are listed below. Find the five-number summary for the given data. [5 marks]

36 40 48 65 67
69 70 73 75 76
79 82 87 90 99

k) Some schools in a certain country carry out random testing for drugs on their students in a bid to curb drug abuse. A study was carried out to establish if testing for drugs by schools reduced use of drugs by students. The table below shows results from questionnaire on drug use by students in high school.

| Drug Use | | | |
|---|---|---|---|
| Drug Tests? | Yes | No | Total |
| Yes | 2,091 | 3,561 | 5,653 |
| No | 6,452 | 10,985 | 17,437 |

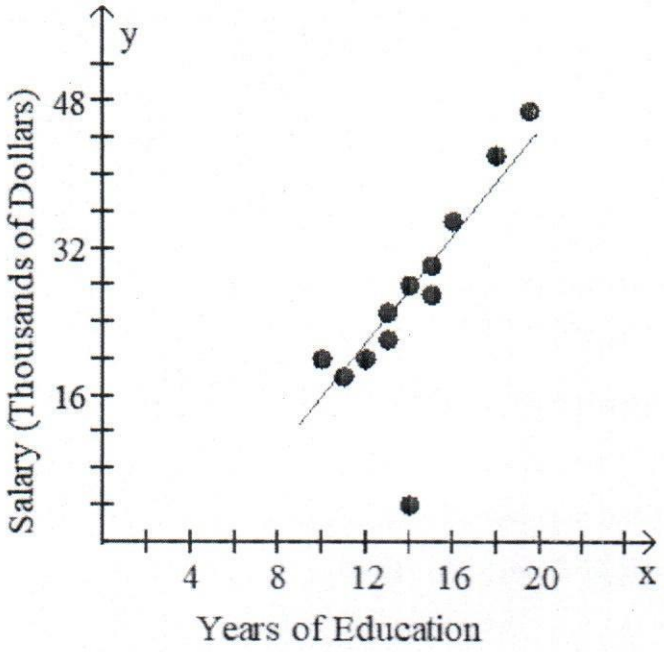
i. What were the response and explanatory variables? [2 marks]

ii. Was this an observational study or an experiment? [1 mark]

l) What is a *z-score*? What values of $z$ would be unusual if a distribution is bell shaped? [3 marks]

## QUESTION TWO [20 MARKS]

a) A researcher recorded for each of a number of recent high school and college graduates the number of years of education and their annual salary when they started their first job. The scatter plot is shown below together with the regression equation. The point (14, 4) was excluded when obtaining the regression equation.




Is the point (14, 4) an outlier on x? Is it an outlier on y? Is it a regression outlier? [3 marks]

b) A computer network manager wants to test the reliability of some new and expensive fiber-optic Ethernet cables that the computer department just received. The computer department received 4 boxes containing 30 cables each. The manager does not have the time to test every cable in each box. The manager will choose one box at random and test 6 cables chosen randomly within that box.

i. What is the population of interest? [2 mark]

ii. What is the sample? [2 mark]

c) In a clinical trial, 780 participants suffering from high blood pressure were randomly assigned to one of three groups. Over a one-month period, the first group received a low dosage of an experimental drug, the second group received a high dosage of the drug, and the third group received a placebo. The diastolic blood pressure of each participant was measured at the beginning and at the end of the period and the change in blood pressure was recorded.

i. Identify the explanatory variable. [2 mark]

ii. Identify the response variable. [2 mark]

d) It is believed that seat position within a bus may have some effect on whether one experiences motion sickness. The table below classifies each person in a random sample of bus riders by the location of his or her seat and whether nausea was reported.

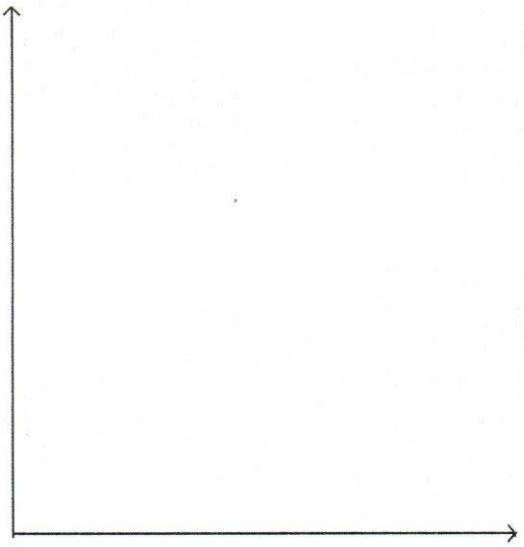
| | Nausea | No Nausea |
|---|---|---|
| Front | 58 | 870 |
| Middle | 166 | 1163 |
| Rear | 193 | 806 |

i. What is the response variable, and what is the explanatory variable? [2 marks]
ii. How do the proportions experiencing nausea compare for the 3 seat positions? [5 marks]
iii. What proportion of all sampled bus riders experienced nausea? [2 marks]

**QUESTION THREE [20 MARKS]**

a) The data below represent the numbers of absences and the final grades of 15 randomly selected students from a statistics class. Use a scatter plot to display the data. Is there a relationship between the students' absences and their final grades? [5 marks]

| Student | Number of Absences | Final Grade (Percent) |
|---|---|---|
| 1 | 5 | 79 |
| 2 | 6 | 78 |
| 3 | 2 | 86 |
| 4 | 12 | 56 |
| 5 | 9 | 75 |
| 6 | 5 | 90 |
| 7 | 8 | 78 |
| 8 | 15 | 48 |
| 9 | 0 | 92 |
| 10 | 1 | 78 |
| 11 | 9 | 81 |
| 12 | 3 | 86 |
| 13 | 10 | 75 |
| 14 | 3 | 89 |
| 15 | 11 | 65 |



b) Twenty-four workers were surveyed and asked how long it takes them to travel to work each day.
The data below are given in minutes.

20 35 42 52 65 20 60 49 24 37 23 24 22 20 41 25 28 27 50 47 58 30 32 48

i. Construct the stem-and-leaf plot for the data. [4 marks]
ii. What is the mode? [1 mark]

iii. What is the median? [2 marks]

c) The table below shows the unemployment rate in one city from 2003 to 2012.

| Year | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|---|---|---|---|
| Unemployment Rate (Percent) | 5.90 | 5.78 | 5.45 | 5.28 | 5.06 | 4.88 | 4.80 | 4.63 | 4.44 | 4.24 |

i. Construct a time plot for these data. [4 marks]

ii. Is there a trend? If so, what kind? [2 marks]

iii. Would a histogram more clearly describe the above dataset? Explain. [2 marks]

## QUESTION FOUR [20 MARKS]

a) Distinguish between the following pairs of variables: [4 marks]

i. Nominal variable and ordinal variable

ii. Categorical variable and quantitative variable

b) What are the main ways of *gathering data*, and what are their advantages and disadvantages? [4 marks]

c) An analysis of data for the 47 Kenyan counties on $y$ = violent crime rate (measured as number of violent crimes per 100,000 people in the county) and $x$ = poverty rate (percent of people in the county living at or below the poverty level) yielded the regression equation,

$$\hat{y} = 209.9 + 25.5x.$$

i. Interpret the slope. [3 marks]

ii. The county poverty rates ranged from 8.0 (for county X) to 24.7 (for county Y). Over this range, find the range of predicted values for the violent crime rate. [5 marks]

iii. Would the correlation between these variables be positive or negative? Why? [4 marks]

## QUESTION FIVE [20 MARKS]

a) The owners of a coffee shop conducted a taste test to determine whether its customers preferred a new coffee brand to the current one sold by the shop. Customers who were willing to participate were given small samples of each of the two brands in random order and were asked to select which one they preferred without knowing the brand. Of the 100 participating customers, 90% chose the new brand. Based on these results, the owners determined that a majority of their customers preferred the new brand and therefore switched their coffee supplier.

Identify aspects of this statistical that refer to: [3 marks]

i. Design

ii. Description

iii. Inference

b) A survey of 1004 American adults 18 years and older were asked, "Do you have a great deal of concern regarding global warming (the greenhouse effect)?" Of the 1004 adults surveyed, 40% said they worried about global warming a great deal.

i. Is the study above an example of an observational study or an experimental study? Explain. [2 marks]

ii. Calculate an approximate margin of error for the results of this study. How is it interpreted? [3 marks]

c) A stock broker has been following different stocks over the last month and has recorded whether the various stock values are up, unchanged, or down at the end of the month. The results were

| Stock performance | up | same | down |
|---|---|---|---|
| Count | 21 | 7 | 12 |

i. What is the variable of interest? [1 mark]

ii. Is the variable categorical or quantitative? [1 mark]

iii. Which response is the mode? [1 mark]

iv. Add proportions to this frequency table. [3 marks]

d) The enrollment for fall semester at University X is as follows.

| Enrollment | Count |
|---|---|
| Undergraduate | 24,814 |
| Graduate/Professional | 8386 |
| Independent Study | 20 |

i. Construct a bar graph for these data. [4 marks]

ii. Would a dot plot or a stem-and-leaf plot make sense for these data? Explain. [2 marks]