UNIVERSITY OF PERADENIYA

DEPARTMENT OF STATISTICS AND COMPUTER SCIENCE

END SEMESTER EXAMINATION SEMESTER I – 2019/2020

AS 404 - DATA INTEGRATION AND MANAGEMENT

Answer **ALL** questions. Time Allowed: **Two hours**.

Reg. Number -

1. Select the most suitable answer from the given options and encircle the letter of your choice.

[*2.5 * 10 = 25 marks*]

i. Characteristics of a population are represented by ________, while those of a sample are represented by _________.

a) statistics; measures b) statistics; parameters c) statistics; variables
d) parameters; statistics e) variable; parameter

ii. Which one of the followings is a continuous variable?

a) age in years of a university student
b) monthly income group of a government servant
c) the length of time required to answer a telephone call
d) number of questions answered in a question paper
e) number of children in a family

iii. What are information, which can be obtained from a box-plot about a given set of values?

a) Minimum, Maximum, Median, Mean
b) Minimum, Maximum, Median, Variance
c) First Quartile , Third Quartile, Median, Range
d) Minimum, Maximum, Mean, Variance
e) First Quartile , Outliers , Median, Variance

iv. What is the most suitable plot to show the association between two variables?

a) Scatter Plot b) Line Chart c) Bar Chart
d) Histogram e) Pie Chart

v. ___________ is **not** an atomic class of object in R.

a) Integer b) Matrix c) Complex
d) Logic e) Character

vi. Let, vecX is a vector and the length of vecX can be obtained by,

a) dim(vecX) b) Length(vecX) c) len(vecX)
d) length(vecX) e) vecX.length()

vii. Which of the following command you would use to determine the current working directory of an R session?

a) getwd() b) getDir() c) workDir()
d) getWD() e) getWDir()

viii. Bottles of surgical spirit are supposed to contain 100 ml of surgical spirit. An inspector takes a random sample of 20 bottles and measures the volumes of their contents. If the inspector wants to carry out a hypothesis test to determine whether the manufacturer correctly fills the bottles, which of the following alternative hypothesis he should use?

a) $H_1: \mu = 100$ b) $H_1: \mu < 100$ c) $H_1: \mu > 100$

d) $H_1: \mu > 101$ e) $H_1: \mu \neq 100$

ix. A machine is supposed to produce paper with a mean thickness of 0.06mm. Fifteen random measurements of a paper gave a mean of 0.059 mm with a standard deviation of 0.003 mm. If the thickness of the paper has a normal distribution, which of the following test you would use to determine whether the machine produces papers as expected?

a) Z test b) T test c) F test

d) Chi-square test e) Q test

x. A manufacturer claims that 6 out 10 children prefer his band of chocolate to any other. In a random sample of 200 children, it was found that 114 of them prefer that chocolate. Which of the alternative hypothesis you would use to test the manufacturer's claim?

a) $H_1: p \neq 134$ b) $H_1: p < 0.6$ c) $H_1: p > 134$

d) $H_1: p \neq 0.6$ e) $H_1: p > 0.6$

2. a) i. Define the two terms "Population" and "Sample". [*8 marks*]

ii. State two advantages and two disadvantages of Primary data over Secondary Data. [*4 marks*]

iii. An exam has 8 score bands; 5.5, 6.0, 6.5, 7.0, 7.5, 8.0, 8.5 and 9.0. If someone claims that ***Score band of a candidate of this exam*** is an example for a discrete variable, do you agree with him? Give the reasons for your answer. [*4 marks*]

b) The process of data analysis is an iterative process which involves many steps.

i. List the five main stages of the data analysis process. [*5 marks*]

ii. Briefly describe one of the stages you have listed in part i. [*4 marks*]

3. a) The birth weights of babies are known to be normally distributed with a standard deviation of 0.4 kg. In a study, it was found that the average weight of 25 randomly selected babies was 3.4 kg.

i. Construct a 95% confidence interval for the mean birth weight of a baby. [*5 marks*]

ii. Interpret the result in part i. [*4 marks*]

b) Jars of honey are filled by a machine. It has been found that the quantity of honey in a jar follows a normal distribution with a mean of 640.3 g and a standard deviation of 3.2 g. It is believed that the machine controls have been altered in such a way that, although the standard deviation is unaffected, the mean quantity has changed. A random sample of 60 jars is taken, and the mean quantity of those 60 jars is found to be 641.2 g. The quality control office wants to find whether the sample provides evidence at 5% level of significance that the mean quantity has changed.

i. State the suitable null and alternative hypotheses. [*4 marks*]

ii. State the appropriate test statistic and distribution of the test statistic. [*3 marks*]

iii. Calculate the test statistic under the null hypothesis. [*3 marks*]

iv. Determine the critical value of the test statistic. [*2 marks*]

v. State the decision of the test. [*2 marks*]

vi. Interpret the decision of the test in part v. [*2 marks*]

Hint : If X has a normal distribution with mean zero and standard deviation one,
$P[X < -1.96] = P[X > 1.96] = 0.025$

4. The data below are the final exam scores of 10 randomly selected Engineering students and the number of hours they studied per week.

Hours (X)	3	5	2	8	2	4	4	5	6	3
Scores (Y)	65	80	60	88	66	78	85	90	90	71

A data analyst wishes to use a simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ to describe the changes in final exam scores based on the number of hours students have spent on their studies per week.

i. State the least squares estimates of β_0 and β_1. [*7 marks*]

ii. Estimate the regression coefficients β_0 and β_1 of the regression line, Y on X for the given data. [*8 marks*]

iii. Interpret your results in part ii. [*6 marks*]

iv. Predict expected test score of a student who has studied 7 hours per week. [*4 marks*]

[Hint : $\sum x = 42, \sum y = 773, \sum x^2 = 208, \sum y^2 = 60875, \sum xy = 3406$]

*****