INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

Mid-Autumn Semester 2017-18

Date of Examination: **19.09.2017** Session: **FN**

Subject No. : **CS40003** Subject: **Data Analytics**

Department: Computer Science & Engineering

Full Marks: 60 Time: 02 hours

Instructions:

- Non-programmable digital calculator can be used for calculations.
- Use statistical table, if you need.
- Answers to all problems should be given in the same order as they appear in the question paper.
- All the symbols, if not stated explicitly bears usual meaning.
- Clearly mention reasonable assumption(s), if any, while answering the questions.
- Answer to ALL questions.

Question No. 1 [10×2=20]

i. An observation related to an opinion poll regarding the abortion law is recorded, which is furnished in the contingency table shown below (Table 1).

Table 1: Q. No. 1(i)

Opinion	Male	Female	Transgender
For	82	70	62
Against	93	62	67
Neutral	25	18	21

(a) What is the size of the sample in the observation? **Ans. 500**

(b) What is the χ^2 value at $\alpha = 0.05$ to reject a hypothesis "Opinion concerning the abortion law are not dependent with genders"? Ans. 9.488

ii. MHRD (Ministry of Human Resource Development), Government of India takes an initiative to improve the country's human resources and hence set up 23 IIT's in the country. To measure the engineering aptitudes of graduates, MHRD conducts GATE (Graduate Aptitude test in Engineering) examination for a total mark of 500 every year. A sample of 300 students who have taken GATE examination in 2016 were selected at random and the mean is observed as 220. In this context, there is a need to statistically infer the mean mark of all GATE-2016 examinee.

(a) Write down the two hypotheses, which may be relevant in the hypothesis testing.
Ans. $H_0: \mu=220$ $H_1: \mu \neq 220$.

(b) The hypothesis test that you have mentioned, whether it comes under one-tailed or two-tailed test?
Ans. two-tailed.

iii. With reference to the hypothesis testing which of the following option(s) is(are) **not correct**?

(a) If the p value is 0.043, then the null hypothesis would be rejected when $\alpha = 0.01$.

(b) If the null hypothesis is rejected by a one-tailed test, then it will also be rejected by a two-tailed test.

(c) If the null hypothesis is rejected at 0.01 level of significance, then it will also be rejected at 0.05 level of significance.

(d) If the test statistics falls in the acceptance region, then the alternate hypothesis has been proven to be true.

Ans. (a) (b) and (d)

iv. Refer the data in Table 2 showing the relationship between x and y.

Table 2: Q. No. 1(iv)

x	0	1	2	3	4	5	6
y	1	4	5	3	2	3	4

The relationship between x and y can be better understood with

(a) Simple linear regression analysis.

(b) Simple non-linear regression analysis.

(c) Auto-regression analysis.

(d) It cannot be told precisely.

Select the most appropriate option(s) which you may think as true.

Ans. (d).

v. If covariance for a set of time series data $Y_1, Y_2, \dots\dots, Y_n$ are arranged in an $n \times n$ matrix V, then

(a) The entry on the $i-th$ row and $j-th$ column gives the $cov(Y_i, Y_j)$.

(b) The diagonal entries are variance values.

(c) Either lower (or upper) half of the above (or below) the diagonal provides useful information for the calculation of auto-correlation coefficients.

(d) The $p-th$ $(p = 1, 2, \dots\dots, n)$ entry in the first row is useful for the calculation of order p autocorrelation coefficient.

Select the correct option(s) from the above-mentioned alternatives.

Ans. (a), (b) and (c).

vi. For a given sample, the observation is as follows (see Table 3). Here, x denotes a sample value and $f(x)$ denotes the frequency of occurrence of x.

Table 3: Q. No. 1(vi)

x	1	2	3	4	5	6
$f(x)$	25	50	10	30	40	20

Find the five-point summary of the above data.

Ans.

1. Min= 1
2. Max=6
3. Median= 4
4. 1st Quartile(Q1)=2
5. 3rd Quartile(Q3)=5

vii. Find the location of mean given a sample whose distribution of values is shown in Fig. 1

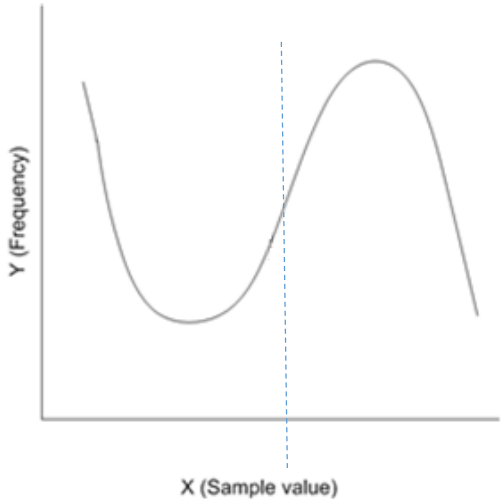

Fig. 1: Q. No. 1(vii)

viii. MapReduce is a technology meant for

(a) Data visualization.

(b) Massive parallel computing.

(c) Query reporting.

(d) Data storing in Cloud

Select the correct alternative(s).

Ans. (b).

ix. On which type of data : N, O, I, R, you can compute each of the following?

(a) Mean Ans. (I, R)

(b) Median Ans.(O, I, R)

(c) Mode Ans. (N, O, I, R)

(d) IQR Ans. (N, O, I, R)

x. The confusion matrix related to Hypothesis testing is shown in Table 4. Identify the entries for correct decisions and Type-I and Type-II errors in the table.

Table 4: Q. No. 1(x)

		Observation	
		H_0 is true	H_0 is false
Decision	H_0 is rejected	Type I error	
	H_0 is accepted		Type II error

Question No. 2 [3+4+3=10]

(a) Write down the 3V characteristics of Big data.

(b) NOIR topology is used to categorize different type of data. Give an example to each of NOIR data.

(c) Explain the data cube modelling and following operations in it.

i. Roll-up

ii. Drill-down

iii. Slice

You should illustrate your answers with appropriate diagrams.

Question No. 3 [2+2+3+3=10]

A sample is collected to learn the opinion on "Need of Facebook in our society" from a randomly selected people with varied ages. The frequency of opinion in favour of the motion is shown in the form of frequency histogram (see Fig. 2). Here, age is represented as grouped data.

Based on the data, answer the following questions.

(a) Data is skewed or not? If skewed, then whether it is positively or negatively skewed data? Ans. Skewed, Positive.

(b) Calculate the mean of the observation. Ans. mean= $\frac{\sum FiXi}{\sum Fi}$ =48.75

(c) Calculate the variance of the observation. Ans. Variance= $\frac{\sum Fi(Xi-\overline{X})2}{\sum Fi}$ = 265.09

(d) What is the coefficient of variance (CV) in this case? What does it mean?

Ans. CV= 18.39

For each calculation you should write the corresponding formula and all steps in your calculation.

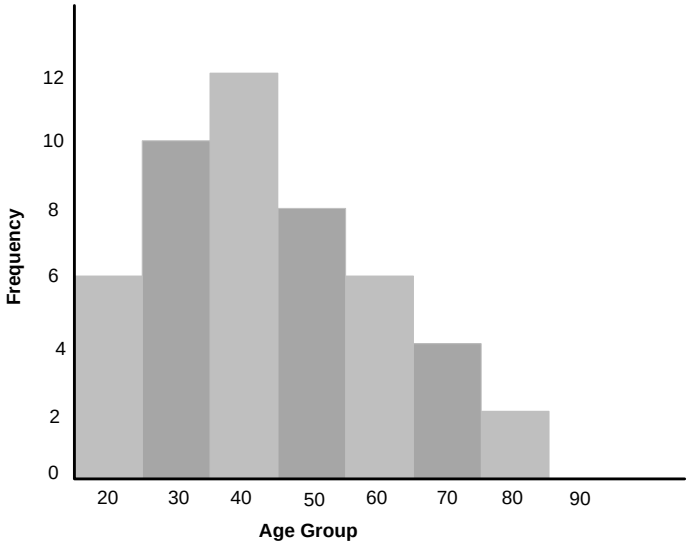

Fig. 2: Q. No. 3

Question No. 4 [10]

The state health monitoring department wants to monitor if the people in a municipal area are suffering from the malnutrition problem or not. The department declares an area affected if the haemoglobin count is below 13.0 mg/L for male and 12.0 mg/L for female. A sample of 16 female participants with different ages selected at random from different families is collected and the observation is listed in Table 5.

Table 5: Q. No. 4

12.08	11.71	11.89	11.72
12.00	11.90	11.77	11.81
12.33	11.67	11.79	11.79
11.94	11.84	12.17	11.87

Test the hypothesis that "Female inhabitants are not suffering from the malnutrition".

Assume the significance level $\alpha = 0.05$.

Clearly show each step of the hypothesis testing. State any assumption, if you make.

Ans.

t=2.47, H_0 is rejected

Assumption: Population follow normal distribution.

Question No. 5 [10]

A fitness club wants to infer if obesity is related to weight. The club selected 15 members who are visiting the club at random intervals and the observation is as shown in Table 6.

Table 6: Q. No. 5

Observation	Age	Severity of obesity
1	YY	VL
2	Y	L
3	M	VH
4	OO	VH
5	OO	H
6	M	H
7	Y	L
8	YY	L
9	Y	VL
10	Y	L
11	M	H
12	M	VH
13	O	VL
14	O	L
15	O	VH

Legends
Age
YY: Very Young
Y: Young
M: Middle-aged
O: Old
OO: Very Old
Obesity
VL: Low
L: Normal
H: High
VH: Very High

Test the hypothesis that "Severity of obesity is related to age". Assume the significance level $\alpha = 0.05$.

Clearly mention all the steps in your answer.

Answer:

Observation	Age	Rank	Obesity	Rank	Rank Difference	D^2
1	YY	4.5	VL	7.6	-3.1	9.61
2.	Y	7	L	8.2	-1.2	1.44
3.	M	8	VH	8.5	-0.5	0.25
4.	OO	4.5	VH	8.5	-4	16
5.	OO	4.5	H	7.3	-2.8	7.84
6.	M	8	H	7.3	0.7	0.49
7.	Y	7	L	8.2	-1.2	1.44
8.	YY	4.5	L	8.2	-3.7	13.69
9.	Y	7	VL	7.6	-0.6	0.36
10.	Y	7	L	8.2	-1.2	1.44
11.	M	8	H	7.3	0.7	0.49
12.	M	8	VH	8.5	-0.5	0.25
13.	O	14	VL	7.6	6.4	40.96
14.	O	14	L	8.2	5.8	33.64
15.	O	14	VH	8.5	5.5	30.25

$r = 1 - \frac{6\sum D2}{n3-n} = 0.718$

$t = r\sqrt{n-2}/(1-r^2) = 3.72$

Check with (n-2)=13 degree of freedom, it is rejected

---*---