CS40003
Data Analytics

Mid-Autumn Semester Test
(Session 2016-2017)

Full Marks: 60 Time: 120 minutes

Instructions

- This is a question-cum-answer booklet. No separate sheet is required for solving any problem and answering.
- There are two parts in the question paper. Answer to both parts.
- To give your answers, use the ANSWER SHEET given in the Page 11 of the booklet. Don't give answer anywhere else. Put a CIRCLE on the option you have chosen as correct.
- **There is NO NEGATIVE marking**.

Part A

All questions in this part are of multiple choice type questions.
For a question, there may be one or more option(s) is(are) correct.
For question with more than one correct options, credit will be given on pro-rata basis.
No credit will be given, if wrong options(s) is(are) chosen.
Give answer on the ANSWER SHEET (Page 11) only.

1. Present size of the digital universe is in the order of

   (a) Terabyte (TB)
   (b) Petabyte (PB)
   (c) Exabyte (EB)
   (d) Zetabyte (ZB)

2. Which is/are the source of data in data analytics?

   (a) Scientific instruments
   (b) Social media
   (c) Mobile devices
   (d) Sensor networks

3. Elastic is a tool for

   (a) Strong big data
   (b) Processing data with scalable architecture
   (c) A distributed file system
   (d) Cloud security

4. MapReduce is meant for

   (a) Data visualizations
   (b) Massive parallel programming

(c) Query reporting
(d) Data storage in Cloud

5. Which data scale uses "zero point as origin"?

(e) Nominal
(f) Ordinal
(g) Interval
(h) Ratio

6. Which is/are **not true**?

(a) Interval data can be transformed to "Categorical" data
(b) Interval data can be transformed to "Categorical" data and vice-versa
(c) Interval data can be transformed to ratio data
(d) Interval data can be transformed to ratio data and vice-versa

7. Which operation **cannot** be carried out on "Ordinal" data?

(a) To find the minimum
(b) To find the mean
(c) To find the mode
(d) To find the median

8. The data type that can be used to store both interval and ratio data are

(a) Integer
(b) Float
(c) Character
(d) Double

9. If the interquartile range is zero, you can conclude that

(a) the range must also be zero
(b) the mean is also zero
(c) at least 50% of the observations have the same value
(d) all of the observations have the same value

10. The "average" type of grass used in Kharagpur campus lawns is best described by

(a) the mean
(b) the median
(c) the mode.
(d) the standard deviation

11. Identify which of the following is a measure of dispersion

(a) median
(b) 90th percentile
(c) interquartile range
(d) mean

12. What is the primary characteristic of a set of data for which the standard deviation is zero?

(a) All values of the variable appear with equal frequency.

(b) All values of the variable have the same value.
(c) The mean of the values is also zero.
(d) None of the above is correct.

13. The median is a better measure of central tendency than the mean if

(a) the variable is discrete
(b) the distribution is skewed
(c) the variable is continuous
(d) the distribution is symmetric

14. A measurable characteristic of a population is

(a) a parameter
(b) a statistic
(c) a sample
(d) an experiment

15. The degrees of freedom of n numbers is

(a) n-1
(b) n+1
(c) n
(d) 0 (zero)

16. Central limit theorem is applicable to

(a) Only continuous probability distribution
(b) Only discrete probability distribution
(c) Only normal distribution
(d) Any probability distribution

17. Which of the following probability distributions is/are discrete distribution(s)

(a) Binomial distribution
(b) Poisson's distribution
(c) Hypergeometric distribution
(d) Weibull distribution

18. If μ and σ denote the mean and standard deviation of a population, then the standard normal distribution is better described as

(a) $f(x: A, B) = \begin{cases} \frac{1}{B-A} & A \leq x \leq B \\ 0 & Otherwise \end{cases}$

(b) $f(x: \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$

(c) $f(z: 0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty < z < \infty$

(d) $f(x: \mu, \sigma) = \begin{cases} \frac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$

19. Which of the following statements is/are **not** correct

(a) $\frac{1}{\sigma^2}\sum(x_i-\mu)^2$ is a Chi-square distribution with *n*-degrees of freedom

(b) $\frac{(n-1)S^2}{\sigma^2}$ is a Chi-square distribution with (*n*-1) degrees of freedom

(c) $\frac{(\bar{x}-\mu)^2}{\sigma^2/n}$ is Chi-square distribution with 1 degree of freedom

(d) None of the above

20. Which of the following statement is/are correct?

(a) χ^2-distribution is used to describe the sampling distribution of S^2
(b) t-distribution is used when population mean μ is known and standard deviation of sample S is known
(c) F distribution is used when variance of populations ${\sigma_1}^2$, ${\sigma_2}^2$ and samples ${S_1}^2$, ${S_2}^2$ are known
(d) All of the above

Part B

This part includes 20 concept level questions.
You can solve a question in the space provided in the booklet.
Don't use any extra sheet for problem solving.
Few options are given as the answers. Select the correct option and put a circle on the option you have chosen on the ANSWER SHEET (Page 11).
Do not give answer elsewhere.

21. Map from entries in Column A to appropriate entries in Column B in the following table.

	Column A		**Column B**
(p)	Pig	(w)	Data storage
(q)	HDFS	(x)	Data process server
(r)	EC_2	(y)	Tool from parallel programming
(s)	ZooKeeper	(z)	Data analysis technique

(a) (p)-(y), (q)-(w), (r)-(x), (s)-(z)
(b) (p)-(x), (q)-(w), (r)-(y), (s)-(z)
(c) (p)-(w), (q)-(y), (r)-(z), (s)-(x)
(d) (p)-(y), (q)-(x), (r)-(z), (s)-(w)

22. Consider the data about all students in a course stored with the following structure.

Table Q.22

Name	Roll No	Category*	Mark1	Mark2	Total	Grade
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

**Category denotes whether a student belongs to UG or PG*

If the structure is used to store the data of 100 students, then the dimension of the data is

(a) 2
(b) 7
(c) 100
(d) 200
(e) 700

23. According to NOIR classification, the attribute "Category" in Table. Q22 can be categorized as

(a) Categorical
(b) Symmetric binary
(c) Asymmetric binary
(d) Ordinal

24. Consider the following data for two variables, X and Y given in the distribution graph (Figure Q.24)

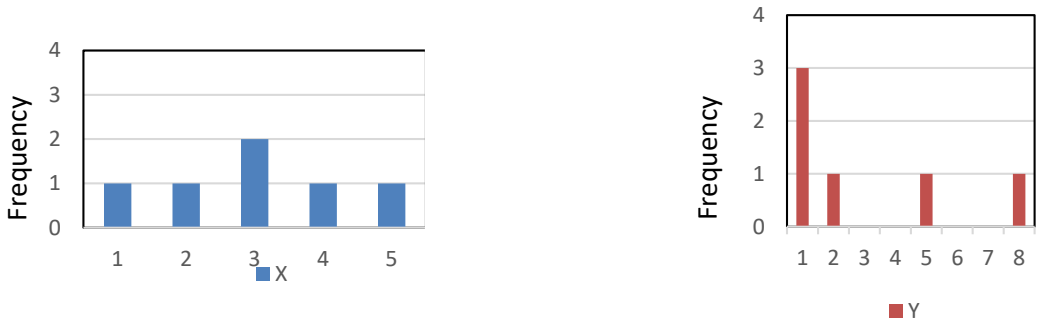


Figure Q.24

The Mean(X), Median(X), Mean(Y), Median(Y).

(a) Mean(X) = 3.0, Median(X) = 1.0, Mean(Y) = 2.0, Median(Y) = 2.5
(b) Mean(X) = 3.0, Median(X) = 2.0, Mean(Y) = 2.5, Median(Y) = 2.0
(c) Mean(X) = 3.0, Median(X) = 3.0, Mean(Y) = 3.0, Median(Y) = 1.5
(d) Mean(X) = 3.0, Median(X) = 4.0, Mean(Y) = 3.5, Median(Y) = 1.0

25. Consider a data related to male and female population of size 200 related to their options as engineering and medical profession. A contingency table is given summarizing a survey

Table Q.25

	Male	Female	Total
Engineering	52	08	60
Medical	28	112	140
Total	80	120	200

With reference to the Table. Q25, the degrees of freedom of the sample is

(a) 1
(b) 2
(c) 3

(d) 4

26. Which of the correct formulation for calculating coefficient variation? All symbols bear usual meanings.

(a) $\frac{1}{n} \sum_{i=1}^{n}(x_i - \bar{x})^2$

(b) $\frac{1}{n-1} \sum_{i=1}^{n}(x_i - \bar{x})^2$

(c) $\frac{1}{n-1} \sum_{i=1}^{n}(\bar{x} - x_i)^2$

(d) $\frac{\sigma}{\bar{x}} \times 100$

27. A set of data points follow a simple linear relation $y = 3x + 2$, where x is any integer number. The mean of the values of y for all values of x in the range [1 ... 100] (equally probable) is

(a) 50
(b) 50.5
(c) 152
(d) 153.5

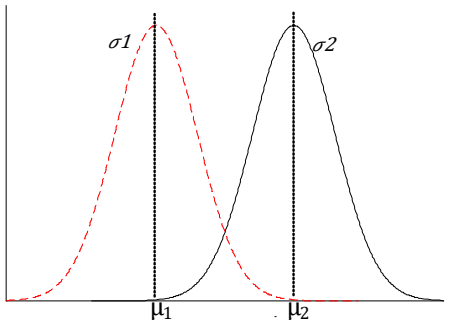
28. In the following Table. Q28, Column A lists some sampling distributions, whereas Column B lists the name of sampling distributions. All symbols bear theor usual meanings. The matching from Column A and Column B is

(A)	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	(W)	Normal distribution
(B)	$\frac{\bar{X} - \mu}{S/\sqrt{n}}$	(X)	Chi-squared distribution
(C)	$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$	(Y)	t-distribution
(D)	$\frac{(n-1)S^2}{\sigma^2}$	(Z)	F distribution

(a) (A)-(Z), (B)-(X), (C)-(Y), (D)-(Z)
(b) (A)-(X), (B)-(Z), (C)-(W), (D)-(W)
(c) (A)-(Y), (B)-(W), (C)-(X), (D)-(Y)
(d) (A)-(W), (B)-(Y), (C)-(Z), (D)-(X)

29. With reference to Figure Q.29, which option correctly represents the two normal distributions?

Figure Q.29



(a) $\sigma_1 \leq \sigma_2$, $\mu_1 = \mu_2$
(b) $\sigma_1 = \sigma_2$, $\mu_1 \geq \mu 2$
(c) $\sigma_1 \leq \sigma_2$, $\mu_1 = \mu 2$
(d) $\sigma_1 = \sigma_2$, $\mu_1 \leq \mu 2$

30. Suppose frequency distribution of two samples (I and II) are shown in Figure Q.30. Then

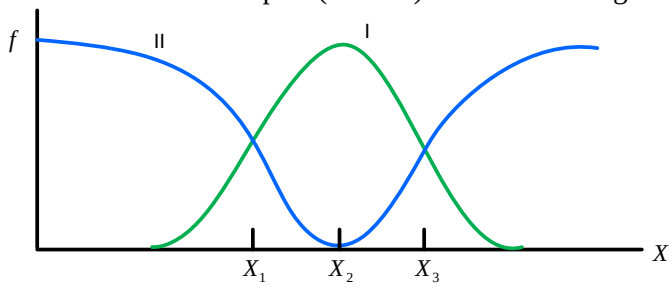


Figure Q.30

(a) the means, medians and modes for both I and II will be located at X2.
(b) the means of both I and II are at X1 and median and mode of II are at X1 and X3, respectively.
(c) the means of both I and II are at X1 and mode and median of II are at X1 and X3, respectively.
(d) data II does not have either median or mean.

31. A sample of pounds lost in a given week by individual members of a weight reducing clinic produced the following statistics.

mean = 5 pounds,	first quartile = 2 pounds
median = 7 pounds,	third quartile = 8.5 pounds
mode = 4 pounds,	standard deviation = 2 pounds

Identify the correct statement.

(a) One-fourth of the members lost less than 2 pounds.
(b) The middle 50% of the members lost between 2 and 8.5 pounds.
(c) The most common weight loss was 4 pounds.
(d) All of the above are correct.

32. In the following context, which denotes a random variable?

There is a box containing 100 balls: 30 red, 20 blue and 50 black balls.

(a) Probability that we draw two blue balls or two red balls from the box.
(b) Drawing any 5 balls at random.
(c) The number of red, blue and black balls drawn from the box.
(d) Drawing the number of five red and six black balls from the box.

33. Which denotes the **Type-I** error according to the error classification in a hypothesis testing?

	H_0 is false	H_0 is true
H_0 is rejected	**A**	**B**
H_0 is accepted	**C**	**D**

(a) A
(b) B
(c) C
(d) D

34. Following are two compositions for a hypothesis testing. Choose the incorrect statement.

H-I: $H_0: \mu = 100$
$H_1: \mu > 100$

H-II: $H_0: \mu \leq 100$
$H_1: \mu > 100$

(a) The two hypotheses are equivalent.
(b) H-I is valid, whereas H-II is not.
(c) H-I is a one-tailed test, whereas H-II is a two-tailed test.
(d) H-II satisfies exclusive and exhaustive property, whereas H-I does not.

35. In statistical inference, if the value of α increases, then

(a) The rejection region for H_0 increases.
(b) The rejection region for H_0 decreases.
(c) The acceptance region for H_1 increases.
(d) The acceptance region for H_1 decreases.

36. For a two-tailed hypothesis test,

(a) Sample mean $\mu > \mu_{H_0}$ and $\mu < \mu_{H_0}$ under a given α
(b) Sample mean $\mu > \mu_{H_0}$ or $\mu < \mu_{H_0}$ under a given α
(c) Sample mean $\mu > \mu_{H_0}$ under a given α
(d) Sample mean $\mu < \mu_{H_0}$ under a given α

37. Which of the following statement is not true in the context of any continuous probability distribution function?

(a) $f(x) \geq 0, \; for \; all \; x \in R$
(b) $\int_{-\infty}^{\infty} f(x)dx = 1$
(c) $\mu = \int_{-\infty}^{\infty} xf(x)\,dx$
(d) $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$

38. The mean of a sample is 8. If three other samples with mean 7, 8 and 9 are added to the sample, then the mean of the resultant sample is

(a) 8
(b) 8, if all samples are of equal size.
(c) Cannot be determined from the given data.

(d) The sample mean will be decided by the population mean, if the sizes of all samples are at least 30.

39. From the tabulation of marks of students participated in four courses C_1, C_2, C_3 and C_4, box-plots are drawn, which is shown in Figure Q.39.

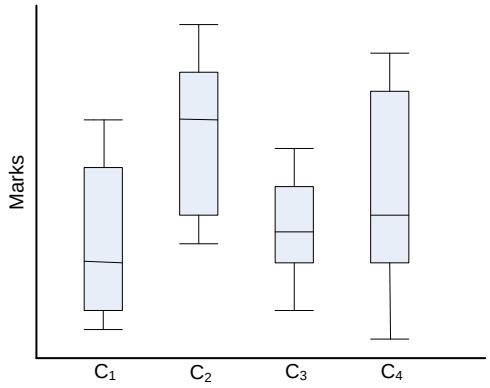


Figure Q.39

The course in which students perform better is

(a) C_1
(b) C_2
(c) C_3
(d) C_4

40. The GM of the following data will be calculated as

X = [50, 125, 70, 56, 49, 98]

(a) 70
(b) 74
(c) 100
(d) 101

Name:_____ Roll No._____

ANSWER SHEET

1.	A	B	C	**D**
2.	**A**	**B**	**C**	**D**
3.	A	**B**	C	D
4.	A	**B**	C	D
5.	A	B	C	**D**
6.	A	**B**	C	**D**
7.	A	**B**	C	D
8.	**A**	**B**	C	**D**
9.	A	B	C	**D**
10.	A	B	**C**	D
11.	A	B	**C**	D
12.	A	**B**	C	D
13.	A	**B**	C	D
14.	A	**B**	C	D
15.	**A**	B	C	**D**
16.	A	B	C	**D**
17.	**A**	**B**	**C**	D
18.	A	B	**C**	D
19.	A	B	C	**D**
20.	A	B	C	**D**

21.	**A**	B	C	D
22.	**A**	B	C	D
23.	**A**	B	C	D
24.	A	B	**C**	D
25.	**A**	B	C	D
26.	A	B	C	**D**
27.	A	B	C	**D**
28.	A	B	C	**D**
29.	A	B	C	**D**
30.	**A**	**B**	**C**	**D**
31.	A	B	C	**D**
32.	A	B	**C**	D
33.	A	**B**	C	D
34.	**A**	B	C	**D**
35.	**A**	B	C	D
36.	**A**	B	C	D
37.	**A**	B	C	D
38.	A	**B**	**C**	**D**
39.	A	B	**C**	D
40.	**A**	B	C	D

Signature of the student with date

FOR OFFICIAL USE

Part A			**Part B**		
Correct Answer	Wrong answers	Total (X)	Correct Answer	Wrong answers	Total (Y)
		Grand Total (X×1+Y×2)			

SPACE FOR ROUGH WORK

SPACE FOR ROUGH WORK

SPACE FOR ROUGH WORK

SPACE FOR ROUGH WORK