Register Number:
Date:

ST. JOSEPH'S COLLEGE (AUTONOMOUS), BANGALORE – 27
M.Sc.(BIG DATA ANALYTICS) – I SEMESTER
SEMESTER EXAMINATION – OCTOBER 2021
(Examination conducted in January-March 2022)

BDA 1121: BASIC STATISTICAL METHODS

TIME: 2.5 HOURS **MAX MARKS: 70**

This Paper contain FOUR printed pages and THREE parts

ANSWER ALL QUESTIONS **20 X 1 = 20**

1. Determine the median of the following series
   77 73 72 70 75 79 78

   (a) 76
   (b) 75
   (c) 77
   (d) 78
2. The arithmetic mean of 9 observations is 100 and that of 6 is 80, the combined mean of all the 15 observations will be
   (a) 100
   (b) 80
   (c) 90
   (d) 92
3. One of the measures of dispersion which is more useful in case of open-ended distributions
   (a) Range
   (b) Mean Deviation
   (c) Quartile Deviation
   (d) Standard Deviation
4. When all the values in a series occur the same number of times, then one must not compute the value of
   (a) Mean
   (b) median
   (c) mode
   (d) weighted mean
5. If in a set of discrete values of observations, 50 per cent values are greater than 25 , the Q_2 is :
   (a) 20
   (b) 25
   (c) 50
   (d) 75

6. Chebyshev's theorem says that 99 percent of the values will lie within ±3 standard deviations from the mean for
   (a) bell-shaped distributions
   (b) positively skewed distributions
   (c) negatively skewed distributions
   (d) all distributions
7. The S.D of a set of 50 observations is 8. If each observations is multiplied by 2 , then the new value of standard deviation will be
   (a) 4
   (b) 8
   (c) 16
   (d) none of the above
8. Why are variance and standard deviation the most popular measures of variability?
   (a) They are the most stable and are foundations for more advanced statistical analysis
   (b) They are the simplest to calculate with large data sets
   (c) They provide nominally scaled data
   (d) (d) None of the above
9. A scatter diagram
   (a) is a statistical test
   (b) must be linear
   (c) must be curvilinear
   (d) is a graph of x and y values
10. There is a high inverse association between measures 'overweight' and 'life expectancy'. A correlation coefficient consistent with the above statement is
   (a) r =0.80
   (b) r=0.20
   (c) r=-0.20
   (d) r=-0.80
11. The strength of a linear relationship between two variables x and y is measured by
   (a) r
   (b) r^2
   (c) R^2
   (d) b_{xy} or b_{yx}
12. Data visualisation tools provide an accessible way to see and understand …. in data
   (a) trends
   (b) outliers
   (c) patterns
   (d) All the above
13. Plotting a histogram of the residuals helps you determine:
   (a) If the error terms are normally distributed
   (b) If the error terms are centred around zero
   (c) If there are any visible patterns in the error terms
   (d) Both (a) and (b)
14. If unexplained variation between variables x and y is 0.25, then r^2 is
   (a) 0.25
   (b) 0.50
   (c) 0.75
   (d) none of these

15. Which of the following is used to calculate the p-value for a particular beta coefficient?
    (a) The standard error of the beta coefficient
    (b) The t-statistic of the beta coefficient
    (c) The null hypothesis for the beta coefficient
    (d) None of the above
16. Of the following measurement levels, which is required for the valid calculation of the Spearman's correlation coefficient?
    (a) nominal
    (b) ordinal
    (c) interval
    (d) ratio
17. The line of 'best fit 'to measure the variation of observed values of dependent variable in the sample data is
    (a) regression line
    (b) correlation coefficient
    (c) standard error
    (d) none of these
18. Which Package contains most fundamental functions to run R?
    (a) root
    (b) child
    (c) base
    (d) parent
19. Use of Central Limit Theorem relates to
    (a) the shape of the sampling distribution
    (b) the sample statistics to estimate population parameters
    (c) large sample size more than 30 observations
    (d) all of these
20. What does it mean if you fail to reject the Null hypothesis in the case of simple linear regression?
    (a) β1 and thus, the independent variable it is associated with is significant in the prediction of the dependent variable.
    (b) β0 and thus, the independent variable it is associated with is significant in the prediction of the dependent variable.
    (c) β0 and thus, the independent variable it is associated with is insignificant in the prediction of the dependent variable
    (d) None of the above

PART B

Answer ANY SIX questions **6 X 5 = 30**

21. Explain the different methods of dealing with missing data and their limitations.
22. Why is data visualization important? Mention a few techniques of it.
23. Compute the 35th percentile, the 55th percentile, Q1, Q2, and Q3 for the following data.

    16 8 29 13 17 20 11 34 32 27 25 30 19 18 33
24. How would you account for the predominant choice of arithmetic mean as a measure of central tendency? Under what circumstances would it be appropriate to use mode or median?
25. Write a note on the following terms: (a) Qualitative and Quantitative data
    (b) Cross sectional and Time Series Data (c) Primary and Secondary Data

26. Differentiate between correlation and regression.
27. Write a note on the different scales of measurement using examples.
28. Construct a box-and-whisker plot on the following data. Do the data contain any outliers? Is the distribution of data skewed?

540 690 503 558 490 609 379 601 559 495 562 580 510 623 477 574 588 497

527 570 495 590 602 541

PART C

Answer ANY TWO questions **2 X 10 = 20**

29. In order to estimate how much water will need to be supplied to a locality in East Delhi area during the summer of 2015, the Minister asked the General Manager of the water supply department to find out how much water a sample of families currently uses. The sample of 20 families used the following number of gallons (in thousands) in the past years

9.3 19.6 14.5 17.8 14.7 15.0 13.9
12.7 10.0 13.0 25.0 16.3 11.2 20.2
15.4 11.6 16.5 11.0 12.2 10.9

(a) What is the mean and median amount of water used per family?

(b) Suppose that 10 years from now, the government expects that there will be 1800 families living in that colony. How many gallons of water will be needed annually, if rate of consumption per family remains the same?

(c) In what ways would the information provided in (a) and (b) be useful to the government? Discuss.

(d) Why might the Government have used the data from a survey rather than just measuring the total consumption in Delhi?

30. A company wants to study the relationship between R&D expenditure (in Rs 1000's) and annual profit (in Rs 1000's) . The following table presents the information for the last 8 years.

Year	1988	1987	1986	1985	1984	1983	1982	1981
R&D	9	7	5	10	4	5	3	2
Annual Profit	45	42	41	60	30	34	25	20

(a) Estimate the sample correlation coefficient
(b) Test the significance of correlation coefficient at α at 5% level of significance.

31. Explain the concept of multiple regression and try to find out an example in the practical field where multiple regression analysis is likely to be helpful.

The equation of a regression line is $y(estimated)=50.506-1.646x$ and the data are as follows:

x	5	7	11	12	19	25
y	47	38	32	24	22	10

Solve for residuals and graph a residual plot. Do these data seem to violate any of the assumptions of regression?