Assignment 1

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. In inferential statistics, the aim is to:

   (a) learn the properties of the sample by calculating statistics on the sample.

   (b) learn the properties of the population by calculating statistics on the population.

   (c) learn the properties of the sample by calculating statistics on the population.

   (d) learn the properties of the population by calculating statistics on the sample.

   (e) none of the above.

2. In a medical study, doctors recorded the average calorie intake for a group of adolescents in a specific year and their corresponding average increase in height for the same period. If the doctors were to perform regression analysis on this data, which variable should they consider to be the independent variable and which one should they consider to be the dependent variable?

   (a) Independent variable: average calorie intake; dependent variable: average increase in height

   (b) Independent variable: average increase in height; dependent variable: average calorie intake

3. The placement office of a particular college wants to analyse the data on campus placements conducted in the college for a particular year. To compare results across different branches, they count the number of placement offers received by students in each branch. Is it appropriate to represent this information graphically in the form of a histogram, with the counts on the Y axis and the different branches on the X axis?

   (a) no

   (b) yes

4. A Boeing 747 passenger aircraft has different fuel consumption rates at different stages of flight (take-off, landing, etc.). In certain emergency situations, fuel needs to be dumped in the air. The fuel flow rate in this process is much higher than the fuel consumption rates during normal flight. To project the fuel requirements for the Boeing 747 fleet of your airlines, you are given the historical data with actual fuel consumption rates for different aircraft during the different phases of flight (along with corresponding duration data). To understand the “average” fuel consumption of the fleet, which measure of central tendency would you prefer?

   (a) mean

(b) median

(c) mode

(d) none of the above

5. In a small startup company, there are only 16 employees. To identify the different designations, the company uses a system of grades. There are two grade G1 employees, six grade G2 employees, four grade G3 employees, two grade G4 employees and two grade G5 employees. The salaries for the different grades are as follows: G1 - Rs. 25,000, G2 - Rs. 35,000, G3 - Rs. 45,000, G4 - Rs. 60,000, and G5 - Rs 5,00,000. What are the mean, median and mode salaries respectively, of the employees of this company?

   (a) 97,500, 35,000, 35,000

   (b) 97,500, 40,000, 35,000

   (c) 97,500, 45,000, 35,000

   (d) 97,500, 45,000, 45,000

6. For the above question, if we are mainly interested in the salary made by a typical employee of the company, which of the measures is/are suitable?

   (a) mean

   (b) median

   (c) mode

   (d) all of the above

7. For the data given in question 5, what is the inter quartile range and standard deviation of the salaries?
   Hint: Consider **Method 1** described here (link: https://en.wikipedia.org/wiki/Interquartile_range) to identify the quartiles in calculating the IQR.

   (a) 10,000, 1,57,437.82

   (b) 10,000, 1,52,438.51

   (c) 17,500, 1,57,437.82

   (d) 17,500, 1,52,438.51

8. Both IQR and standard deviation are measures of dispersion, characterising the deviation of the data from a central value. In the previous question we observe that there is a large difference between the IQR and the standard deviation calculated on the salaries data. Which among the following reasons do you think account for this large difference?

   (a) IQR is more robust to outliers than standard deviation.

   (b) In standard deviation, we are squaring the deviations leading to a large increase in the resultant value.

   (c) The central value around which IQR and standard deviation are being calculated are different.

   (d) In calculating IQR, we are ignoring part of the data where as for standard deviation, we consider all of the data.

9. Using the same salary data, what is the mean absolute deviation around the mean and the median absolute deviation around the median?

   (a) 1,00,625, 5,000
   (b) 1,00,625, 15,000
   (c) 1,07,333, 5,000
   (d) 1,07,333, 15,000

10. You are given two lists of numbers. The numbers in the first list represent a variable that does not have a non-arbitrary zero value (say, for example, date), but the relative differences between the values are meaningful. The numbers in the second list represent a variable that does have a meaningful, non-arbitrary zero value in addition to meaningful relative differences between the values (say, for example, speed). Which among the following measures is suitable for one of the lists of numbers but not for the other?

    (a) arithmetic mean
    (b) standard deviation
    (c) geometric mean
    (d) median
    (e) all measures are equally suitable for both lists