Assignment 1 (Sol.)

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.

   Sol. (d)
   In general, we are interested in the properties of the population and not of any one sample of the population. However, it is normally not possible to perform measurements on the entire population. Hence, we perform measurements on the sample and then calculate statistics based on these measurements. However, the aim is still to use these measurements and statistics to infer the properties of the population.

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

   Sol. (a)
   We expect the average calorie intake to have an effect on growth and not the other way round.

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

Sol. (a)
In histograms, the variable on the X axis is quantitative not categorical as in this case with the different branches.

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

   Sol. (a)
   While fuel flow during emergency dumping will appear as outliers, they should not be ignored (as might happen when using the median), since they are neither errors in the data, nor are they not meaningful to the task at hand (you could imagine that with low probability such situations will happen again in future). Thus, we should use the mean.

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

   Sol. (b)
   The number of grade G2 employees is the maximum (6). Thus, the **mode** salary is Rs. 35,000 (the salary corresponding to grade G2).

   Listing out all salaries in ascending order, we have: 25,000, 25,000, 35,000, 35,000, 35,000, 35,000, 35,000, 35,000, 45,000, 45,000, 45,000, 45,000, 60,000, 60,000, 5,00,000, 5,00,000. Since there are an even number of items, we consider the middle two, which are 35,000 (eighth) and 45,000 (ninth). Taking the average of the two, we get

   $$\frac{35000 + 45000}{2} = 40000.$$

   Thus, the **median** salary is Rs. 40,000.

Summing the above list of salaries and dividing by 16 (the number of employees), we have

$$\frac{2 * (25000 + 60000 + 500000) + 4 * 45000 + 6 * 35000}{16} = 97500.$$

Thus, the **mean** salary is Rs. 97,500.

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

   Sol. (b) & (c)
   As we can observe in the solution to the previous problem, the mean is highly influenced by the very high salaries earned by only two of the company's 16 employees. Thus, we would prefer to use either the median or the mode to describe the salary of a typical employee.

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

   Sol. (d)
   Listing out all salaries in ascending order, we have: 25,000, 25,000, 35,000, 35,000, 35,000, 35,000, 35,000, 35,000, 45,000, 45,000, 45,000, 45,000, 60,000, 60,000, 5,00,000, 5,00,000. Since there are an even number of salaries, we split the list of salaries in half. The first quartile is the median of the first list of salaries, which in this case would be 35000. The third quartile is the median of the second list of salaries, which in this case is

   $$\frac{45000 + 60000}{2} = 52500.$$

   So, we have

   $$IQR = Q_3 - Q_1 = 52500 - 35000 = 17500.$$

   Thus, the IQR is Rs. 17,500.

   Form the solution to question 5 we know that the mean of the salaries is 97500. Calculating the standard deviation, we have

   $$\sqrt{\frac{2 * ((25000 - \bar{x})^2 + (60000 - \bar{x})^2 + (500000 - \bar{x})^2) + 4 * (45000 - \bar{x})^2 + 6 * (35000 - \bar{x})^2}{16}}$$

   $$= 152438.51$$

   where $\bar{x} = 97500$. Thus, the standard deviation of salaries is Rs 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.

   Sol. (a), (c) & (d)
   Since in calculating the IQR, only the middle 50% of the data is considered, the measure is less affected by extreme values/outliers. Also, note that in calculating the IQR the central value (around which deviation is being measured) is the median. This is especially clear to see in this example, since the mean value of 97500 does not even lie in the IQR range (i.e., between 35000 & 52500). Of course, the standard deviation is calculated considering the mean as the central value.

   Note that the effect of the squaring operation in the standard deviation calculation is cancelled by taking the square root and hence it does not contribute to the large difference that we observe in this example.

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

   Sol. (a)
   We know that the mean of the salaries is 97500. Thus, considering the mean absolute deviation around the mean, we have

$$\frac{2 * (|25000 - \bar{x}| + |60000 - \bar{x}| + |500000 - \bar{x}|) + 4 * |45000 - \bar{x}| + 6 * |35000 - \bar{x}|}{16}$$

$$= 100625$$

   where $\bar{x} = 97500$. Thus, the mean absolute deviation around the mean of salaries is Rs 1,00,625.

   Again, we know that the median of the salaries is 40000. To calculate the median absolute deviation around the median, we simply calculate the individual absolute deviations from the median and find the median of this list. Listing out all salaries in ascending order, we have: 25,000, 25,000, 35,000, 35,000, 35,000, 35,000, 35,000, 35,000, 45,000, 45,000, 45,000, 45,000, 60,000, 60,000, 5,00,000, 5,00,000. For a salary of 25000, the absolute deviation from the median is calculated as

$$|25000 - 40000| = 15000$$

where 40000 is the median salary.

Following the same process, we can list out the absolute deviations from the median in ascending order as: 5000, 5000, 5000, 5000, 5000, 5000, 5000, 5000, 5000, 5000, 15000, 15000, 20000, 20000, 460000, 460000. The median of this list is 5000, thus the median absolute deviation around the median is Rs. 5,000.

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

    Sol. (c)
    The first list represents a variable measured on the interval scale, whereas the second list represents a variable measured on the ratio scale. For the interval scale, the operations of multiplication and division (on the data points) are not meaningful. Thus, for the numbers in the first list, it is not meaningful to consider the geometric mean. This is not the case for the ratio type, for which a unique and non-arbitrary zero value exists, and hence for which, multiplication and division (and therefore, the geometric mean) does make sense.