Week 2 - Questions

Questions

1. The table below contains students' grades and the number of students who achieved that grade for a specific unit of study. What is the data type of the "Grades" column and what type of chart would you use to analyse the distribution of the dataset?

Grades	Frequency
HD	11
DI	24
CR	28
P	16
F	6

2. Given the following data set of years in the current job: [2, 10, 8, 4, 45], calculate the mean, median, and range.

3. Calculate the standard deviation and variance of the following dataset: [5, -3, 12, 7, -1, 9, 15, -4, 6, 2].

4. Identify the data type (nominal, ordinal, interval, or ratio) for each of the following variables and state which measures of central tendency would be appropriate.

    a) Education level (High school, Bachelor's, Master's, PhD)

    b) Annual income in dollars

    c) Customer satisfaction rating (1-5 stars)

    d) Zip codes

    e) Temperature in degrees Celsius

Answers

1. The table below contains students' grades and the number of students who achieved that grade for a specific unit of study. What is the data type of the "Grades" column. What type of chart would you use to analyse the distribution of the dataset and why?

Grades	Frequency
HD	11
DI	24
CR	28
P	16
F	6

   a. In the table above, the data type of the "Grades" column is "Ordinal Data" because the grades mentioned have a specific ordering /ranking to them where the ranks of each grade is defined as follows: HD > DI > CR > P > F. The best type of graph that can be used to analyse the distribution of the dataset is to create a bar chart of the table above because a bar chart can help visualise some key aspects of the data, such as which grade was the most common and which was the least. It can also help in comparing data points in different grade categories which will provide a holistic view of how the students performed. (For simplicity and clarity in showing the frequency distribution of a single categorical variable, a bar chart remains the most straightforward choice. However, any other solution with a valid reason will be evaluated.)

2. Given the following data set of years in the current job: [2, 10, 8, 4, 45], calculate the mean, median, and range.

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

a. Mean
   i. Mean = $\frac{sum\ of\ all\ values\ in\ dataset}{total\ number\ of\ values\ in\ dataset} = \frac{2+10+8+4+45}{5} = \mathbf{13.8\ years}$

b. Median
   i. Step 1 – Sort the dataset: [2, 4, 8, 10, 45]

   ii. Step 2 – determine m.
      1. Number of values in dataset = m = 5. Therefore, odd numbers in dataset. Therefore, m is odd.

   iii. Step 3 – calculate r.
      1. When m is odd, $m = 2r + 1$
      2. Therefore, $r = \frac{m-1}{2}$. Therefore, $r = \frac{5-1}{2} = 2$

   iv. Step 4 - Find the median.
      1. Since m is odd, the median of the dataset is in position $r + 1$.
      2. $r + 1 = 2 + 1 = 3$.
      3. Number in position 3 of the sorted dataset is 8.
      4. ***Therefore, median = 8.***

c. Range
   i. $Range = Largest\ Number\ in\ Dataset - Smallest\ Number\ in\ Dataset$

   ii. $\mathbf{Range = 45 - 2 = 43\ years}$

3. Calculate the standard deviation and variance of the following dataset: [5, -3, 12, 7, -1, 9, 15, -4, 6, 2].

   a. $Variance = \frac{\sum(X_i - mean)^2}{N-1}, N = number\ of\ values\ in\ dataset$

      i. Step 1 – calculate mean.

         1. $\frac{5+(-3)+12+7+(-1)+9+15+(-4)+6+2}{10} = 4.8$

      ii. Step 2 – calculate variance:

$$\frac{(5-4.8)^2+(-3-4.8)^2+(12-4.8)^2+(7-4.8)^2+(-1-4.8)^2+(9-4.8)^2+(15-4.8)^2+(-4-4.8)^2+(6-4.8)^2+(2-4.8)^2}{(10-1)}$$

$$= 39.96$$

      iii. Therefore, ***variance = 39.96***

   b. Standard Deviation = $\sqrt{variance}$

      i. **Standard Deviation** $= \sqrt{\mathbf{39.96}} \approx \mathbf{6.32}$

4. a) Ordinal data - appropriate measures: median, mode
   b) Ratio data - appropriate measures: mean, median, mode
   c) Ordinal data - appropriate measures: median, mode
   d) Nominal data - appropriate measure: mode only
   e) Interval data - appropriate measures: mean, median, mode