Statistics Assignment 2

1. **What are the three measures of central tendency?**

   Ans: Mean, Median, Mode

2. **What is the difference between the mean, median, and mode? How are they used to measure the central tendency of a dataset?**

   Ans: Mean is the sum of all data points divided by number of data points. It is commonly referred to as the average. Median is the middle value of the data set when data points are arranged by ascending or descending order. If there is an even number of data points then median is the average of two middle numbers.The mode is the value that appears most frequently in a dataset. A dataset can have more than one mode if multiple values appear with the same highest frequency (bimodal, multimodal), or no mode if all values are unique.

3. **Measure the three measures of central tendency for the given height data: [178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]**

   Ans: Mean :177.01875
   Median: 177.0
   Mode: 178

4. **Find the standard deviation for the given data: [178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]**

   Ans: Standard Deviation: 1.789

5. **How are measures of dispersion such as range, variance, and standard deviation used to describe**
   the spread of a dataset? Provide an example.

   Ans:

   Range :

   Range used to describe the spread of a data set. It provides a quick sense of how spread out the data is by measuring the distance between the smallest & largest value in the data set. A large range indicates that the data points are spread out over a wide interval, suggesting more variability or diversity within the data set. A small range indicates that the data points are close to each other, suggesting less variability and that the data is more consistent or clustered.

Example

Consider two data sets:

- **Data Set A**: {3,4,5,6,7}\{3, 4, 5, 6, 7\}{3,4,5,6,7}
- **Data Set B**: {1,4,7,10,13}\{1, 4, 7, 10, 13\}{1,4,7,10,13}

For Data Set A:

- **Minimum Value**: 3
- **Maximum Value**: 7
- **Range**: 7−3=47 - 3 = 47−3=4

For Data Set B:

- **Minimum Value**: 1
- **Maximum Value**: 13
- **Range**: 13−1=1213 - 1 = 1213−1=12

Variance:

Variance measures the average squared deviation of each data point from mean. It gives a sense of how much the data varies from the mean.

Standard Deviation:

Standard deviation is the square root of variance, providing a measure of dispersion in the same units as the data.

Imagine you're tracking the number of steps you take each day over a week:

- **Day 1**: 4,000 steps
- **Day 2**: 6,000 steps
- **Day 3**: 8,000 steps
- **Day 4**: 10,000 steps
- **Day 5**: 12,000 steps

Step 1: Calculate the Mean (Average)

First, find the mean number of steps:

Mean=4000+6000+8000+10000+120005=8,000 steps\text{Mean} = \frac{4000 + 6000 + 8000 + 10000 + 12000}{5} = 8,000 \text{ steps}Mean=54000+6000+8000+10000+12000=8,000 steps

Step 2: Calculate the Variance

Variance measures how much each step count deviates from the mean on average, but it uses squared differences to avoid negative values.

Example:

Imagine you're tracking the number of steps you take each day over a week:

Day 1: 4,000 steps

Day 2: 6,000 steps

Day 3: 8,000 steps

Day 4: 10,000 steps

Day 5: 12,000 steps

Step 1: Calculate the Mean (Average)

First, find the mean number of steps:

Mean = (4000 + 6000 + 8000 + 10000 + 12000) / 5 = 8,000 steps

Step 2: Calculate the Variance

Variance measures how much each step count deviates from the mean (8,000 steps) on average, but it uses squared differences to avoid negative values.

1. Calculate each deviation from the mean:

- Day 1: 4,000 - 8,000 = -4,000
- Day 2: 6,000 - 8,000 = -2,000
- Day 3: 8,000 - 8,000 = 0
- Day 4: 10,000 - 8,000 = 2,000
- Day 5: 12,000 - 8,000 = 4,000

2. Square each deviation:

- (-4,000)^2 = 16,000,000
- (-2,000)^2 = 4,000,000

- (0)^2 = 0

- (2,000)^2 = 4,000,000

- (4,000)^2 = 16,000,000

3. Find the average of these squared deviations:

Variance = (16,000,000 + 4,000,000 + 0 + 4,000,000 + 16,000,000) / 5 = 8,000,000

Step 3: Calculate the Standard Deviation

Standard deviation is the square root of the variance. It brings the measure back to the original units .

6. **What is a Venn Diagram?**

Ans: A Venn diagram is a graphical representation of sets and their relationships to each other. It uses overlapping circles (or other shapes) to show the relationships between different sets. Each circle represents a set, and the overlap between the circles represents the intersection of those sets, showing elements that are common to both. Venn diagrams are commonly used in probability, logic, statistics, and set theory.

7. **For the two given sets A = {2, 3, 4, 5, 6, 7} & B = {0, 2, 6, 8, 10}, find:**

(i) **A ∩ B (Intersection of A and B)**:
The intersection of two sets includes all elements that are common to both sets.
A ∩ B = {2, 6}

(ii) **A ∪ B (Union of A and B)**:
The union of two sets includes all elements that are in either set, without duplicates.
A ∪ B = {0, 2, 3, 4, 5, 6, 7, 8, 10}

8. **What do you understand about skewness in data?**

Skewness refers to the asymmetry or lack of symmetry in the distribution of data. A distribution is skewed if one of its tails is longer or fatter than the other. There are two types of skewness:

- **Right (Positive) Skewness**: The tail on the right side (positive side) of the distribution is longer or fatter. The mean is greater than the median.
- **Left (Negative) Skewness**: The tail on the left side (negative side) of the distribution is longer or fatter. The mean is less than the median.

9. **If a data is right-skewed, what will be the position of the median with respect to the mean?**

   In a right-skewed distribution, the mean is typically greater than the median. This is because the longer right tail pulls the mean to the right, making it larger than the median.

10. **Explain the difference between covariance and correlation. How are these measures used in statistical analysis?**

    Covariance and **correlation** both measure the relationship between two variables, but they do so in different ways:

    - **Covariance**: This measures the direction of the linear relationship between two variables. If the variables tend to increase together, the covariance is positive; if one tends to increase while the other decreases, the covariance is negative. However, covariance does not indicate the strength of the relationship, and its value is not standardized, making it difficult to interpret without additional context.
    - **Correlation**: This measures both the direction and the strength of the linear relationship between two variables. It is a standardized measure, with values ranging from -1 to +1. A correlation of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

    Usage in Statistical Analysis:

    - **Covariance** is used to determine whether two variables change together, but because its value depends on the units of the variables, it is less interpretable on its own.
    - **Correlation** is more commonly used because it provides a standardized measure of the strength and direction of the relationship, making it easier to compare relationships between different pairs of variables.

11. **What is the formula for calculating the sample mean? Provide an example calculation for a dataset.**

The sample mean is calculated using the formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Where:

- X bar is the sample mean,

- xi represents each value in the sample,
- n is the number of observations in the sample.

Example Calculation: For the dataset: 4, 8, 6, 5, 3

Example Calculation: For the dataset: 4, 8, 6, 5, 3

$$\bar{x} = \frac{4 + 8 + 6 + 5 + 3}{5} = \frac{26}{5} = 5.2$$

The sample mean is 5.2.

Q12. For a normal distribution data, what is the relationship between its measure of central tendency?

In a normal distribution:

- The **mean**, **median**, and **mode** are all equal.
- They are located at the center of the distribution.
- This symmetry indicates that the distribution is perfectly balanced around its center.

Q13. How is covariance different from correlation?

Covariance and **correlation** both measure the relationship between two variables, but they do so in different ways:

- **Covariance**: This measures the direction of the linear relationship between two variables. If the variables tend to increase together, the covariance is positive; if one tends to increase while the other decreases, the covariance is negative. However, covariance does not indicate the strength of the relationship, and its value is not standardized, making it difficult to interpret without additional context.
- **Correlation**: This measures both the direction and the strength of the linear relationship between two variables. It is a standardized measure, with values ranging from -1 to +1. A correlation of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Usage in Statistical Analysis:

- **Covariance** is used to determine whether two variables change together, but because its value depends on the units of the variables, it is less interpretable on its own.

- **Correlation** is more commonly used because it provides a standardized measure of the strength and direction of the relationship, making it easier to compare relationships between different pairs of variables.

Q14. How do outliers affect measures of central tendency and dispersion? Provide an example.

Outliers are extreme values that differ significantly from the rest of the data. They can have a strong impact on measures of central tendency and dispersion:

- **Central Tendency**: Outliers can skew the mean, making it less representative of the data. For example, in the dataset {1, 2, 3, 4, 100}, the mean is 22, which is much higher than most of the data due to the outlier (100). However, the median (3) is not affected by the outlier and is a better representation of the dataset's center.
- **Dispersion**: Outliers increase the range, variance, and standard deviation, making the data appear more spread out than it is. In the same dataset {1, 2, 3, 4, 100}, the range is 99, and the variance is much higher due to the outlier.

Outliers can distort the interpretation of data, so it's important to consider them when analyzing datasets.