# 7 March Basic Statistics Assignment

**Q1. What are the three measures of central tendency?**

**ANSWER:-**
The three measures of central tendency are:

1. Mean: The mean is the most commonly used measure of central tendency. It is calculated by summing up all the values in a dataset and dividing the sum by the total number of values. The mean is affected by outliers since it takes into account all the values in the dataset.

2. Median: The median is the middle value in an ordered dataset. To calculate the median, the data must be arranged in ascending or descending order. If there is an odd number of values, the median is the middle value. If there is an even number of values, the median is the average of the two middle values. The median is less affected by outliers compared to the mean.

3. Mode: The mode is the value that appears most frequently in a dataset. It is the only measure of central tendency that can be used with nominal or categorical data. A dataset can have one mode (unimodal), two modes (bimodal), or more than two modes (multimodal). Unlike the mean and median, the mode is not influenced by outliers since it only considers the frequency of values. If all values occur with the same frequency, the dataset is said to be "unimodal" without a specific mode.

**Q2. What is the difference between the mean, median, and mode? How are they used to measure the central tendency of a dataset?**
**ANSWER:-**

The mean, median, and mode are all measures of central tendency used to describe the "typical" or "central" value in a dataset. However, they differ in terms of their calculation methods and the information they provide about the data.

1. Mean: The mean is calculated by summing up all the values in a dataset and dividing the sum by the total number of values. It takes into account every value in the dataset and provides an average value. The mean is sensitive to extreme values or outliers because it incorporates all values in the calculation. It is commonly used when the data is normally distributed or when there are no significant outliers.

2. Median: The median is the middle value in an ordered dataset. To calculate the median, the data must be arranged in ascending or descending order. If there is an odd number of values, the median is the middle value. If there is an even number of values, the median is the average of the two middle values. The median is less influenced by extreme values or outliers because it focuses on the middle values. It is often used when the data is skewed or when there are outliers that might significantly affect the mean.

3. Mode: The mode is the value that appears most frequently in a dataset. Unlike the mean and median, the mode can be used with categorical or nominal data. It provides information

about the most common value or category in the dataset. The mode is not affected by outliers since it only considers the frequency of values. It is useful for identifying the most prevalent category or value in a dataset.

In summary, the mean represents the average value, the median represents the middle value, and the mode represents the most frequently occurring value. The mean is affected by outliers, the median is less influenced by outliers, and the mode is not affected by outliers. The choice of which measure to use depends on the nature of the data and the presence of outliers.

**Q3. 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]**

**ANSWER:-**
To find the three measures of central tendency (mean, median, and mode) 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], we can perform the following calculations:

1. Mean:
To find the mean, we sum up all the values in the dataset and divide by the total number of values:
Mean = (178 + 177 + 176 + 177 + 178.2 + 178 + 175 + 179 + 180 + 175 + 178.9 + 176.2 + 177 + 172.5 + 178 + 176.5) / 16
Mean ≈ 176.94

2. Median:
To find the median, we need to sort the dataset in ascending order first:
172.5, 175, 175, 176, 176.2, 176.5, 177, 177, 178, 178, 178, 178.2, 178.9, 179, 180
The dataset has 16 values, so the median will be the average of the two middle values:
Median = (176 + 176.2) / 2
Median ≈ 176.1

3. Mode:
The mode is the value(s) that appear most frequently in the dataset. In this case, there is no value that appears more than once, so the dataset does not have a mode.

Therefore, the three measures of central tendency for the given height data are approximately as follows:
Mean ≈ 176.94
Median ≈ 176.1
Mode = 178(3)  177(3)

**Q4. 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]**
**ANSWER:-**

To 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], you can follow these steps:

1. Find the mean of the dataset:
   Mean = (178 + 177 + 176 + 177 + 178.2 + 178 + 175 + 179 + 180 + 175 + 178.9 + 176.2 + 177 + 172.5 + 178 + 176.5) / 16
   Mean ≈ 176.94

2. Subtract the mean from each data point and square the result:
   (178 - 176.94)^2, (177 - 176.94)^2, (176 - 176.94)^2, (177 - 176.94)^2, (178.2 - 176.94)^2, (178 - 176.94)^2, (175 - 176.94)^2, (179 - 176.94)^2, (180 - 176.94)^2, (175 - 176.94)^2, (178.9 - 176.94)^2, (176.2 - 176.94)^2, (177 - 176.94)^2, (172.5 - 176.94)^2, (178 - 176.94)^2, (176.5 - 176.94)^2

3. Find the mean of the squared differences obtained in step 2.

4. Take the square root of the mean squared difference to get the standard deviation.

the standard deviation for the given data is approximately 1.661.

**Q5. How are measures of dispersion such as range, variance, and standard deviation used to describe the spread of a dataset? Provide an example.**
**ANSWER:-**

Measures of dispersion, such as range, variance, and standard deviation, are used to describe the spread or variability of a dataset. They provide information about how the data points are distributed and how much they deviate from the central tendency measures.

1. Range: The range is the simplest measure of dispersion and represents the difference between the maximum and minimum values in a dataset. It gives an idea of the total spread of the data. However, it only considers the extreme values and does not take into account the distribution of the rest of the data points. For example, if you have a dataset of exam scores ranging from 60 to 90, the range would be 90 - 60 = 30, indicating a spread of 30 points.

2. Variance: Variance measures the average squared deviation of each data point from the mean. It provides an understanding of how the data points vary around the mean. To calculate the variance, you subtract the mean from each data point, square the result, sum up all the squared deviations, and divide by the total number of data points. A higher variance indicates a greater spread or dispersion. For example, if you have a dataset of exam scores with a variance of 100, it suggests that the scores deviate from the mean by an average of 10 points.

3. Standard Deviation: The standard deviation is the square root of the variance. It measures the average deviation of each data point from the mean, but in the same unit as the original data. It provides a more intuitive understanding of the spread compared to the variance. A higher standard deviation implies a greater dispersion of data points. For example, if the standard deviation of a dataset is 5, it suggests that the data points deviate from the mean by an average of 5 units.

In summary, measures of dispersion like range, variance, and standard deviation are used to quantify the spread of a dataset, providing insights into the variability and distribution of the data points. They complement the measures of central tendency by giving a more comprehensive description of the dataset.

**Q6. What is a Venn diagram?**
**ANSWER:-**
A Venn diagram is a visual representation that uses overlapping circles or other shapes to illustrate the relationships between different sets of objects or concepts. It was introduced by John Venn, a British logician and philosopher, in the late 19th century.

In a Venn diagram, each circle or shape represents a set, and the overlapping areas depict the common elements or relationships between the sets. The non-overlapping regions represent the unique elements specific to each set. The size of the circles or shapes can be proportional to the number of elements in the corresponding sets, although it is not always necessary.

Venn diagrams are commonly used to illustrate set theory concepts, including unions, intersections, and complements of sets. They provide a visual tool to analyze the similarities and differences between sets and aid in logical reasoning and problem-solving.

Venn diagrams can be used in various fields such as mathematics, logic, statistics, computer science, and even in everyday scenarios to depict relationships, groupings, or overlaps between different categories or concepts. They are an effective way to visualize complex relationships and make comparisons between different sets.

**Q7. For the two given sets A = (2,3,4,5,6,7) & B = (0,2,6,8,10). Find:**
**(i) A B**
**(ii) A ∪ B**
**ANSWER:-**

To find the requested operations for the given sets A = {2, 3, 4, 5, 6, 7} and B = {0, 2, 6, 8, 10}, we can perform the following calculations:

(i) A ∩ B (Intersection of A and B):
The intersection of two sets is the set of elements that are common to both sets. In this case, the common element between sets A and B is 2 and 6.
A ∩ B = {2, 6}

(ii) A ∪ B (Union of A and B):
The union of two sets is the set of all elements present in either set, without duplication. In this case, the union of sets A and B consists of all the unique elements from both sets.
A ∪ B = {0, 2, 3, 4, 5, 6, 7, 8, 10}

Therefore:
(i) A ∩ B = {2, 6}

(ii) A ∪ B = {0, 2, 3, 4, 5, 6, 7, 8, 10}

**Q8. What do you understand about skewness in data?**
**ANSWER:-**

Skewness in data refers to the measure of the asymmetry or lack of symmetry in the distribution of a dataset. It provides insights into the shape of the distribution and the concentration of data points on either side of the central tendency.

A dataset can exhibit three types of skewness:

1. Positive Skewness (Right-skewed): A distribution is positively skewed when the tail of the distribution extends towards the right side. In a positively skewed distribution, the majority of data points are concentrated on the left side of the distribution, and the tail extends towards the right. The mean is typically greater than the median in a positively skewed distribution.

2. Negative Skewness (Left-skewed): A distribution is negatively skewed when the tail of the distribution extends towards the left side. In a negatively skewed distribution, the majority of data points are concentrated on the right side of the distribution, and the tail extends towards the left. The mean is typically less than the median in a negatively skewed distribution.

3. Zero Skewness (Symmetric): A distribution is symmetric or exhibits zero skewness when it is perfectly balanced and evenly distributed around the central tendency measures (mean, median, and mode). In a symmetric distribution, the left and right tails are mirror images of each other, and the mean, median, and mode are roughly equal.

Skewness can provide insights into the underlying characteristics of the dataset. For example, in finance, analyzing the skewness of investment returns can help identify the presence of extreme positive or negative returns, indicating higher risk or potential opportunities. In data analysis, understanding skewness can help in selecting appropriate statistical methods or transformations to handle skewed data and make accurate inferences.

Skewness is often measured using statistical measures such as the skewness coefficient, which quantifies the degree and direction of skewness in a dataset.

**Q9. If a data is right skewed then what will be the position of median with respect to mean?**
**ANSWER:-**

If a data set is right-skewed, the position of the median with respect to the mean will typically be lower than the mean. In a right-skewed distribution, the tail of the distribution extends towards the right side, indicating the presence of a few extreme values on the higher end of the data range.

Since the mean is influenced by these extreme values, it gets pulled towards the higher end of the distribution. On the other hand, the median represents the middle value of the data set and is less affected by extreme values. It tends to be closer to the values in the bulk of the data, which are concentrated on the left side of the distribution.

Therefore, in a right-skewed distribution:
- The mean will be greater than the median.
- The median will be located to the left of the mean.

**Q10. Explain the difference between covariance and correlation. How are these measures used in**
**statistical analysis?**
**ANSWER:-**

Covariance and correlation are both measures used in statistical analysis to quantify the relationship between two variables. While they are related, there are important differences between them.

Covariance:
Covariance measures how two variables vary together. It indicates the direction and magnitude of the linear relationship between two variables. A positive covariance indicates that the variables tend to move in the same direction (when one increases, the other also tends to increase), while a negative covariance indicates an inverse relationship (when one increases, the other tends to decrease). The magnitude of covariance is not standardized and depends on the units of the variables. Therefore, it is difficult to interpret the value of covariance alone.

Correlation:
Correlation measures the strength and direction of the linear relationship between two variables, similar to covariance. However, correlation is standardized, ranging from -1 to +1, which makes it easier to interpret. A correlation coefficient of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. The correlation coefficient is independent of the units of the variables and allows for direct comparison between different pairs of variables.

In summary, the main differences between covariance and correlation are:
1. Scale: Covariance is not standardized and depends on the units of the variables, while correlation is standardized between -1 and +1.
2. Interpretation: Covariance alone is difficult to interpret, whereas correlation provides a clear interpretation of the strength and direction of the linear relationship.
3. Comparison: Correlation coefficients can be directly compared between different pairs of variables, while covariance cannot be easily compared due to its dependence on units.

In statistical analysis, covariance and correlation are used to understand the relationship between variables, assess the strength of association, and make predictions. Correlation is commonly used as it provides a standardized measure of association that is easier to interpret and compare. Covariance is often used in certain calculations, such as calculating the variance-covariance matrix in multivariate analysis or in specific cases where the scale and units of the variables are important.

**Q11. What is the formula for calculating the sample mean? Provide an example calculation for a dataset.**

**ANSWER:-**
The formula for calculating the sample mean, denoted by $\bar{x}$ (pronounced "x-bar"), is as follows:

$\bar{x} = (x_1 + x_2 + x_3 + ... + xn) / n$

where:
$x_1$, $x_2$, $x_3$, ..., xn are the individual data points in the dataset, and
n is the total number of data points in the dataset.

To illustrate, let's calculate the sample mean for the dataset [5, 8, 3, 9, 2]:

$\bar{x} = (5 + 8 + 3 + 9 + 2) / 5$
$= 27 / 5$
$= 5.4$

Therefore, the sample mean of the dataset [5, 8, 3, 9, 2] is 5.4.

**Q12. For a normal distribution data what is the relationship between its measure of central tendency?**
**ANSWER:-**

For a normal distribution, the three measures of central tendency (mean, median, and mode) are typically equal or very close to each other.

1. Mean: In a normal distribution, the mean is located at the center of the distribution. It represents the balance point of the data, where the sum of all data points is divided equally. The mean is equal to the median in a perfectly symmetrical normal distribution.

2. Median: The median of a normal distribution is also located at the center of the distribution. It is the middle value that divides the data into two equal halves. In a perfectly symmetrical normal distribution, the median is equal to the mean.

3. Mode: The mode of a normal distribution is the value or values that occur with the highest frequency. In a normal distribution, the mode is also equal to the mean and median. It represents the most common value(s) in the data, which coincides with the center of the distribution.

In summary, for a normal distribution, the mean, median, and mode are equal or very close to each other, as they all represent the central tendency of the data. This equality or proximity between the measures is a characteristic of the symmetric and bell-shaped nature of a normal distribution.

**Q13. How is covariance different from correlation?**
**ANSWER:-**

Covariance and correlation are both measures used to describe the relationship between two variables, but they differ in several key aspects:

1. Scale:
Covariance is not standardized and its value depends on the units of the variables being measured. Consequently, it is challenging to interpret covariance in isolation.

Correlation, on the other hand, is standardized and ranges between -1 and 1. It provides a clear measure of the strength and direction of the linear relationship between variables, independent of the units of measurement.

2. Interpretation:
Covariance indicates the direction of the linear relationship between variables (positive or negative), but its value alone does not provide a precise understanding of the strength or magnitude of the relationship.

Correlation, with its standardized scale, allows for a more meaningful interpretation. A correlation coefficient of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

3. Comparison:
Covariance is not suitable for comparing the relationships between different pairs of variables due to its dependence on the units of measurement. The magnitude of covariance can be influenced by the scale of the variables, making it difficult to make direct comparisons.

Correlation coefficients, being standardized, enable straightforward comparisons between different pairs of variables. The correlation coefficient provides a standardized measure of the strength of the linear relationship, allowing for direct comparisons.

In summary, covariance provides information about the direction of the linear relationship between variables, while correlation goes further by also quantifying the strength and direction of the relationship in a standardized manner. Correlation is often preferred over covariance due to its ease of interpretation and comparability across different datasets.

**Q14. How do outliers affect measures of central tendency and dispersion? Provide an example.**
**ANSWER:-**
Outliers can significantly affect measures of central tendency and dispersion in a dataset. Here's how outliers impact these measures:

1. Measures of Central Tendency:
- Mean: Outliers can pull the mean towards their direction. Since the mean takes into account every value in the dataset, extreme values can greatly influence its value. For

example, consider the dataset [10, 20, 30, 40, 50, 1000]. The outlier value of 1000 significantly increases the mean, pulling it away from the central values of the dataset.
- Median: Outliers have less impact on the median since it represents the middle value of the dataset. As long as the outlier is not within the middle values, the median remains relatively unaffected. In the above example, the median would remain the same as it is not influenced by the outlier.
- Mode: Outliers generally have no impact on the mode since it represents the most frequently occurring value(s) in the dataset. The mode remains the same regardless of the presence of outliers.

2. Measures of Dispersion:
- Range: Outliers affect the range by extending the spread of the dataset. The range increases as the distance between the minimum and maximum values widens due to outliers. In the previous example, the range would increase from 990 (1000 - 10) to 990 (1000 - 10).
- Variance and Standard Deviation: Outliers can substantially impact the variance and standard deviation. These measures quantify the spread of the dataset by considering the deviation of each value from the mean. Outliers increase the sum of squared differences, leading to higher variance and standard deviation. In the example, the variance and standard deviation would be much higher due to the influence of the outlier.

Overall, outliers can skew the measures of central tendency towards their direction and inflate measures of dispersion. It is crucial to identify and handle outliers appropriately to ensure accurate interpretation and analysis of the data.