Q1. What are the three measures of central tendency?

ans.

mean

median

mode

Q2. What is the difference between the mean, median, and mode? How are they used to measure the

central tendency of a dataset?

ans.

mean=add all num and devide by total count of numbers

meadin=The middle value of a set of numbers when arranged in order from smallest to largest.

The median is the most accurate representation when the data distribution is skewed.

Mode=The value that appears most frequently in a set of numbers. The mode is preferred when there is a nominal distribution of the data

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]

mean:

177.02

Median:

177.0

Mode: 178 and

177

(bimodal distribution).

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]

ans.

1.89

Q5. How are measures of dispersion such as range, variance, and standard deviation used to describe

the spread of a dataset? Provide an example.

ans.

Measures of dispersion, such as range, variance, and standard deviation, are used to describe the spread of a dataset, indicating how much the data values differ from one another and from the central tendency (mean).

1. **Range**:

The range is the simplest measure, representing the difference between the maximum and minimum values in a dataset. It provides a quick sense of how spread out the values are.

2. **Variance**:

Variance measures the average squared deviation of each data point from the mean. A higher variance indicates more spread in the data, while a lower variance suggests the data is more clustered around the mean.

3. **Standard Deviation**:

The standard deviation is the square root of the variance. It gives a measure of spread in the same units as the original data, making it more interpretable. A larger standard deviation means more spread out data, while a smaller one indicates that the data points are closer to the mean.

These measures help to quantify the degree of variation within a dataset, allowing for a better understanding of its distribution and variability.

Q6. What is a Venn diagram?

ans.

A Venn diagram is a visual representation of sets and their relationships to each other. It uses overlapping circles (or other shapes) to illustrate how different sets intersect, differ, or share common elements.

Key Components:

Circles (or ellipses): Represent different sets.

Intersection (overlap): The area where circles overlap represents the elements common to both sets.

Non-overlapping areas: Parts of the circles that do not overlap represent elements that are unique to each set.

Purpose:

Venn diagrams are commonly used to:

Show relationships between different groups or sets.

Visualize the union, intersection, and difference between sets.

Illustrate logical relationships, such as those found in probability or set theory.

Example:

For two sets,

A and

B:

The union of the sets (written as

$A \cup B$) is the area covered by both circles, including both the overlapping and non-overlapping parts.

The intersection (written as

$A \cap B$) is the area where the circles overlap, representing the elements common to both sets.

The difference (written as

$A-B$) is the area of circle

A that does not overlap with circle

B.

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

intersection=2,3,4,5,6,7,8,10

union=2,6

Q8. What do you understand about skewness in data?

ans.

Skewness refers to the degree of asymmetry or distortion in the distribution of data. It measures whether the data is evenly distributed around the mean or if it is skewed to one side (left or right). Skewness helps identify the direction of the tail in a distribution.

Types of Skewness:

Positive Skew (Right Skew):

The right tail of the distribution is longer or fatter than the left tail.

Most of the data points are concentrated on the left side of the distribution, with fewer larger values stretching the tail to the right.

In a positively skewed distribution, the mean is typically greater than the median.

Negative Skew (Left Skew):

The left tail of the distribution is longer or fatter than the right tail.

Most of the data points are concentrated on the right side, with fewer smaller values pulling the tail to the left.

In a negatively skewed distribution, the mean is typically less than the median.

No Skew (Symmetrical Distribution):

The data is evenly distributed around the mean, with equal tails on both sides.

A normal distribution is an example of a perfectly symmetrical distribution, where the skewness is close to zero.

Measuring Skewness:

Skewness value:

A skewness of 0 indicates a perfectly symmetrical distribution.

A positive value indicates positive skew (right tail longer).

A negative value indicates negative skew (left tail longer).

Skewness is important in data analysis because it can impact the choice of statistical methods and models. For example, many statistical tests assume normality, and significant skewness may suggest the need for data transformation before applying such tests.

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

ans.

If a dataset is right-skewed (positively skewed), the mean will be greater than the median.

This happens because the long right tail of the distribution pulls the mean in the direction of the larger values, while the median remains less affected by extreme values.

Explanation:

In a right-skewed distribution, most of the data is clustered towards the lower end, with a few larger values (outliers) on the right side.

The mean, being the average of all values, is pulled in the direction of the larger values (right tail).

The median, which is the middle value of the dataset, is less influenced by extreme values, so it stays closer to the center of the data.

Q10. Explain the difference between covariance and correlation. How are these measures used in

statistical analysis?

Covariance:

Covariance is useful for understanding the general relationship between two variables. It is often used in portfolio management (for stocks and asset returns), where the covariance between returns of two stocks helps in understanding how they move together.

Correlation:

Correlation is more commonly used when comparing relationships between variables because it standardizes the relationship, making it easier to interpret and compare. It is used widely in fields like economics, social sciences, and biology to understand the strength and direction of relationships between variables.

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

dataset.

ans.

mean=total sum/num

Q13. How is covariance different from correlation?

Covariance and correlation both measure the relationship between two variables, but they differ in the following ways:

Covariance:

Measures the direction of the relationship between two variables.

Can be positive, negative, or zero:

Positive covariance: Both variables increase together.

Negative covariance: One variable increases while the other decreases.

Zero covariance: No linear relationship.

The value of covariance depends on the units of the variables, making it difficult to interpret across different datasets with different units or scales.

Correlation:

Measures both the direction and the strength of the linear relationship between two variables.

The value of correlation is always between -1 and +1:

+1 indicates a perfect positive linear relationship.

-1 indicates a perfect negative linear relationship.

0 indicates no linear relationship.

Correlation is standardized, meaning it is unit-free and easier to interpret across datasets with different units.

In summary:

Covariance shows the direction of the relationship, but its magnitude is hard to interpret due to its dependence on the scale of the data.

Correlation standardizes the covariance, making it easier to understand the strength and direction of the relationship between the variables.

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

Outliers are data points that are significantly different from the rest of the data. They can have a considerable impact on both measures of central tendency (mean, median, and mode) and dispersion (range, variance, and standard deviation).

Effect on Central Tendency:

Mean: Outliers can significantly skew the mean. Since the mean is sensitive to every value in the dataset, a very large or small outlier can pull the mean in its direction, making it unrepresentative of the majority of the data.

Median: The median is much less affected by outliers. Since the median is the middle value, it remains relatively stable even when extreme values are present.

Mode: Outliers typically have little to no effect on the mode, unless the outlier is repeated frequently.

Effect on Dispersion:

Range: Outliers can dramatically increase the range of the data, as the range is calculated by subtracting the minimum value from the maximum value.

Variance and Standard Deviation: Both variance and standard deviation are influenced by outliers because they are based on squared differences from the mean. A large outlier can lead to a disproportionately large value for these measures, indicating more variability than is actually present in the majority of the data.

Example:

Consider the following dataset:

2

,

3

,

4

,

5

,

6

,

100

2,3,4,5,6,100

Mean:

Mean

=

2

+

3

+

4

+

5

+

6

+

100

6

=

120

6

=

20

Mean=

6

2+3+4+5+6+100

=

6

120

=20

The mean is heavily influenced by the outlier (100).

Median: The median is the middle value when the data is sorted:

2

,

3

,

4

,

5

,

6

,

100

2,3,4,5,6,100

The median is the average of 4 and 5:

Median=4+52=4.5

Median=

24+5

=4.5

The median is not affected by the outlier as much as the mean.

Range:

Range=100−2=98

The range is greatly influenced by the outlier (100), giving a misleading sense of spread.

Standard Deviation:

The presence of the outlier increases the standard deviation, making the data appear more spread out than it actually is.

In this example, the outlier (100) distorts the mean, range, and standard deviation, making them appear larger than they would otherwise be. The median is a more robust measure of central tendency in this case, as it is less affected by the outlier.