assignment=28

July 28, 2023

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

The three measures of central tendency are:

Mean: The mean is the arithmetic average of a set of numbers. It is calculated by adding up all the numbers in the set and then dividing the total by the number of items in the set.

Median: The median is the middle value in a set of numbers that are arranged in order. To find the median, you need to sort the numbers in ascending or descending order and then find the middle value. If there are an even number of values, the median is the average of the two middle values.

Mode: The mode is the value that occurs most frequently in a set of numbers. A set of numbers can have one mode, more than one mode (if two or more values occur with the same highest frequency), or no mode (if no value occurs more frequently than any other value in the set).

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

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 the way they are calculated and in their interpretation.

The mean is the sum of all the values in the dataset divided by the total number of values. It is affected by extreme values, or outliers, in the dataset and may not be a good representation of the central tendency if the dataset has a skewed distribution.

The median is the middle value in the dataset, such that half of the values are greater than or equal to the median, and the other half are less than or equal to the median. The median is less sensitive to outliers than the mean and is a better measure of central tendency for skewed distributions.

The mode is the value that occurs most frequently in the dataset. It is not affected by extreme values, but it may not exist or may not be unique if there are multiple values with the same frequency.

In practice, the choice of the measure of central tendency depends on the characteristics of the dataset and the research question. For example, the mean is often used when the dataset has a normal distribution, while the median is preferred when the dataset is skewed. The mode is often used when describing categorical data or when looking for the most typical value in a dataset.

Overall, the mean, median, and mode are all useful measures of central tendency and should be used in conjunction with other statistical methods to fully describe a dataset.

3 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]

To find the three measures of central tendency for the given height data, we can use the following formulas:

Mean = (sum of all values) / (number of values)

Median = middle value in the sorted dataset

Mode = value that appears most frequently in the dataset

First, we need to sort the dataset in ascending order:

[172.5, 175, 175, 176, 176.2, 176.5, 177, 177, 178, 178, 178, 178.2, 178.9, 179, 180]

Next, we can use the formulas to calculate the measures of central tendency:

Mean = (172.5 + 175 + 175 + 176 + 176.2 + 176.5 + 177 + 177 + 178 + 178 + 178 + 178.2 + 178.9 + 179 + 180) / 15

Mean = 176.97

Median = 177

Mode = 178

Therefore, the mean height of the dataset is approximately 176.97 cm, the median height is 177 cm, and the mode is 178 cm.

4 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]

To find the standard deviation for the given data, we can use the following formula:

s = sqrt([(x - $\bar{x}$)^2] / (n - 1))

where s is the sample standard deviation, x is each value in the dataset, $\bar{x}$ is the sample mean, and n is the sample size.

First, we need to calculate the sample mean:

$\bar{x}$ = (178 + 177 + 176 + 177 + 178.2 + 178 + 175 + 179 + 180 + 175 + 178.9 + 176.2 + 177 + 172.5 + 178 + 176.5) / 16

$\bar{x}$ = 177.31

Next, we can calculate the sum of squared deviations from the mean:

(x - $\bar{x}$)^2 = (178 - 177.31)^2 + (177 - 177.31)^2 + (176 - 177.31)^2 + (177 - 177.31)^2 + (178.2 - 177.31)^2 + (178 - 177.31)^2 + (175 - 177.31)^2 + (179 - 177.31)^2 + (180 - 177.31)^2 + (175

- 177.31)^2 + (178.9 - 177.31)^2 + (176.2 - 177.31)^2 + (177 - 177.31)^2 + (172.5 - 177.31)^2 + (178 - 177.31)^2 + (176.5 - 177.31)^2

(x - $\bar{x}$)^2 = 314.46

Finally, we can calculate the sample standard deviation:

s = sqrt([(x - $\bar{x}$)^2] / (n - 1))

s = sqrt(314.46 / 15)

s = 1.978

Therefore, the sample standard deviation for the given data is approximately 1.978 cm.

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

Measures of dispersion such as range, variance, and standard deviation are used to describe the spread of a dataset by indicating how spread out or tightly clustered the data points are around the central tendency (mean, median, or mode).

The range is the simplest measure of dispersion and is defined as the difference between the largest and smallest values in the dataset. It provides a rough estimate of how much the data points vary from one another.

The variance and standard deviation are more precise measures of dispersion that take into account the distance of each data point from the mean. The variance is calculated as the average of the squared deviations from the mean, while the standard deviation is the square root of the variance. A higher variance or standard deviation indicates that the data points are more spread out from the mean, while a lower variance or standard deviation indicates that the data points are more tightly clustered around the mean.

For example, consider two datasets of student scores on a test:

Dataset 1: [70, 75, 80, 85, 90] Dataset 2: [60, 70, 80, 90, 100]

Both datasets have the same mean score of 80, but they have different spreads of scores. The range of Dataset 1 is 20 (90-70), while the range of Dataset 2 is 40 (100-60), indicating that the scores in Dataset 2 are more spread out. The variance of Dataset 1 is 62.5, and the standard deviation is 7.91, while the variance of Dataset 2 is 200 and the standard deviation is 14.14, again indicating that the scores in Dataset 2 are more spread out. Therefore, measures of dispersion such as range, variance, and standard deviation can provide valuable information about the spread of a dataset and help us better understand the distribution of data points.

6 Q6. What is a Venn diagram?

Measures of dispersion such as range, variance, and standard deviation are used to describe the spread of a dataset by indicating how spread out or tightly clustered the data points are around the central tendency (mean, median, or mode).

The range is the simplest measure of dispersion and is defined as the difference between the largest and smallest values in the dataset. It provides a rough estimate of how much the data points vary from one another.

The variance and standard deviation are more precise measures of dispersion that take into account the distance of each data point from the mean. The variance is calculated as the average of the squared deviations from the mean, while the standard deviation is the square root of the variance. A higher variance or standard deviation indicates that the data points are more spread out from the mean, while a lower variance or standard deviation indicates that the data points are more tightly clustered around the mean.

For example, consider two datasets of student scores on a test:

Dataset 1: [70, 75, 80, 85, 90] Dataset 2: [60, 70, 80, 90, 100]

Both datasets have the same mean score of 80, but they have different spreads of scores. The range of Dataset 1 is 20 (90-70), while the range of Dataset 2 is 40 (100-60), indicating that the scores in Dataset 2 are more spread out. The variance of Dataset 1 is 62.5, and the standard deviation is 7.91, while the variance of Dataset 2 is 200 and the standard deviation is 14.14, again indicating that the scores in Dataset 2 are more spread out. Therefore, measures of dispersion such as range, variance, and standard deviation can provide valuable information about the spread of a dataset and help us better understand the distribution of data points.

Vishvesh Jain Q6. What is a Venn diagram? A Venn diagram is a graphical representation of the relationships between different sets or groups of data. It consists of overlapping circles or ellipses, each representing a set or group, with the overlap representing the elements that are common to both sets.

Venn diagrams are used to visually represent the similarities and differences between two or more sets or groups of data. They can be used to identify and compare the elements that are unique to each set or common to multiple sets, making them a useful tool in various fields such as mathematics, logic, statistics, and data science.

Venn diagrams can be used to represent different types of relationships between sets or groups, including the following:

Disjoint sets: Two sets that have no elements in common.

Overlapping sets: Two or more sets that share some common elements.

Nested sets: Sets that are contained within other sets.

Venn diagrams can be drawn by hand or using software such as Microsoft Excel, Google Sheets, or specialized Venn diagram software. They are a powerful tool for visualizing data and communicating complex ideas in a clear and concise manner.

7 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

(ii) A B represents the set of elements that are common to both sets A and B. To find A B, we need to identify the elements that appear in both sets. From the given sets, we see that the

common element between A and B is 6. Therefore, A B = {6}.

(iii) A B represents the union of sets A and B, which includes all the elements that are in either set A or set B, or both. To find A B, we need to combine the elements from both sets and remove any duplicates. The combined set is:

{0, 2, 3, 4, 5, 6, 7, 8, 10}

Therefore, A B = {0, 2, 3, 4, 5, 6, 7, 8, 10}.

8 Q8. What do you understand about skewness in data?

Skewness is a measure of the asymmetry or lack of symmetry in a dataset. It indicates the degree to which the data is skewed or distorted from a normal distribution.

In a normal distribution, the data is symmetrical, with an equal number of data points on either side of the mean. However, in a skewed distribution, the data is not evenly distributed, with one tail of the distribution being longer or stretched out than the other.

There are two types of skewness:

Positive skewness: In a positively skewed distribution, the tail of the distribution extends to the right, with most of the data concentrated on the left side of the distribution. This is also called right-skewed or right-tailed distribution.

Negative skewness: In a negatively skewed distribution, the tail of the distribution extends to the left, with most of the data concentrated on the right side of the distribution. This is also called left-skewed or left-tailed distribution.

Skewness can have an impact on statistical analysis, as it can affect the central tendency of the data (mean, median, and mode) and the accuracy of some statistical tests. It is important to identify the presence and degree of skewness in a dataset before making any conclusions or decisions based on the data.

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

If a data is right skewed, the median will be less than the mean. This is because in a right-skewed distribution, the tail of the distribution is on the right side, which means there are some extreme values on the right side that pull the mean towards the right, making it higher than the median. The median, on the other hand, is not affected by extreme values and is only influenced by the values in the middle of the distribution. As a result, the median is typically less than the mean in a right-skewed distribution.

10 Q10. Explain the difference between covariance and correlation. How are these measures used in statistical analysis?

Covariance and correlation are both measures of the relationship between two variables in a dataset. However, there are some key differences between these two measures:

1. Definition: Covariance measures the degree to which two variables in a dataset vary toge

2. Range of values: Covariance can take on any value between negative infinity and positive

3. Interpretation: A positive covariance indicates that two variables tend to increase or de

4. Units: Covariance is expressed in the units of the variables being measured, while correl

In statistical analysis, covariance and correlation are both used to determine the relationship between two variables. Covariance is used to measure the strength of the linear relationship between two variables, while correlation is used to quantify both the strength and direction of the relationship. Correlation is considered to be a more useful measure than covariance because it is standardized and provides more meaningful results. Correlation is also used to identify relationships between variables in regression analysis, while covariance is used in portfolio analysis to calculate the risk and return of multiple assets

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

The formula for calculating the sample mean (also known as the arithmetic mean) is:

sample mean = (sum of all values in the dataset) / (number of values in the dataset)

For example, consider the following dataset:

4, 8, 6, 10, 12

To find the sample mean of this dataset, we first need to add up all the values:

4 + 8 + 6 + 10 + 12 = 40

Next, we divide the sum by the number of values in the dataset, which is 5:

sample mean = 40 / 5 = 8

Therefore, the sample mean of this dataset is 8.

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

For a normal distribution, the mean, median, and mode are all equal. In other words, the measure of central tendency is the same, and they are located at the center of the distribution. This is because a normal distribution is symmetric around its mean, so the middle value (median) is the same as the average value (mean), which is also the most frequent value (mode).

This relationship between the mean, median, and mode is a key characteristic of a normal distribution and can be used to identify if a dataset is approximately normally distributed.

13 Q13. How is covariance different from correlation?

Covariance and correlation are two different measures of the relationship between two variables in a dataset:

1. Definition: Covariance measures the degree to which two variables in a dataset vary togetl

2. Range of values: Covariance can take on any value between negative infinity and positive

3. Interpretation: A positive covariance indicates that two variables tend to increase or dec

4. Units: Covariance is expressed in the units of the variables being measured, while correla

In summary, covariance and correlation are both measures of the relationship between two variables, but correlation is considered a more useful measure than covariance because it is standardized, provides more meaningful results, and is easier to interpret.

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

Outliers are extreme values that differ significantly from the other values in a dataset. They can have a significant impact on measures of central tendency and dispersion:

1. Measures of central tendency: Outliers can greatly affect the mean, but have little effect

   30,000, 35,000, 40,000, 45,000, 50,000, 55,000, 60,000, 75,000, 100,000

   The mean salary is (30,000+35,000+40,000+45,000+50,000+55,000+60,000+75,000+100,000) / 9

2. Measures of dispersion: Outliers can greatly affect measures of dispersion such as range,

In summary, outliers can greatly affect measures of central tendency and dispersion, and it is important to identify and handle them appropriately in statistical analysis.

[ ]:

[ ]:

[ ]: