**Statistics Basic-2**

Q1. What are the three measures of central tendency?

The three measures of central tendency commonly used in statistics are:

1. Mean: The mean, also known as the average, is calculated by summing up all the values in a dataset and dividing the sum by the number of values. It represents the typical value or the balance point of the data. The mean is sensitive to outliers, meaning that extreme values can greatly influence its value.
2. Median: The median is the middle value of a dataset when it is arranged in ascending or descending order. If the dataset has an odd number of values, the median is the middle value. If the dataset has an even number of values, the median is the average of the two middle values. The median is less affected by extreme values compared to the mean and is often used when dealing with skewed data or data with outliers.
3. Mode: The mode is the value that appears most frequently in a dataset. In other words, it is the value with the highest frequency. A dataset can have one mode (unimodal) if there is one value with the highest frequency, two modes (bimodal) if two values have the same highest frequency, or more than two modes (multimodal) if multiple values share the highest frequency. The mode is useful when dealing with categorical or discrete data, but it can also be applied to numerical data.

These three measures provide different insights into the central tendency of a dataset and are used to summarize and analyze data in various fields, including statistics, economics, social sciences, and more.

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

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **Mean** | **Median** | **Mode** |
| 1. | The average taken for a set of numbers is called a mean. | The middle value in the data set is called the Median. | The number that occurs the most in a given list of numbers is called a mode. |
| 2. | Add all of the numbers together and divide the sum by the total number of values. | Place all the given numbers in an ascending order | It shows the frequency of occurrence. |
| 3. | The result is the mean or average score. | The next step is to find the middle number on the list. It is called the median. | We can have more than one mode or no mode at all. |
| 4. | Example: To find the average of the four numbers 2, 4, 6, and 8, we need to add the number first.   * 2 + 4 + 6+ 8 = 20 * Divide the sum by the total number of numbers, i. e 4. * 20/4 = 5 is the average or mean | Example: 4, 2, 8, 10, 19.   * Arrange the numbers in ascending order. i .e., 2, 4, 8, 10, 19. * As the total numbers are 5, so the middle number 8 is the median here. | Example: 3, 3, 5, 6, 7, 7, 8, 1, 1, 1, 4, 5, 6.   * Find the frequency of each number. * For number 3, it’s 2. For 5, it’s 2. For 6, it’s 2. For 7, it’s 2. For 8, it’s one. For 1, it’s 3. For 4, it’s 1. * The number with the highest frequency is the mode. Hence, the mode of the given sequence of numbers is 1. |

These measures help to describe the central tendency of a dataset in different ways. The mean provides the average value and is suitable for datasets with a symmetrical distribution. The median gives the middle value and is appropriate for skewed distributions or when outliers are present. The mode identifies the most frequently occurring value and is useful for categorical or discrete data. Researchers and analysts often use these measures in combination to gain a comprehensive understanding of the dataset's central tendency.

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]

* 1. Mean:

To calculate the mean, we'll sum up all the values and divide by the total number of values. Sum of heights = 178 + 177 + 176 + 177 + 178.2 + 178 + 175 + 179 + 180 + 175 + 178.9 + 176.2 + 177 + 172.5 + 178 + 176.5 = 2812.3 Total number of values = 16 Mean = Sum of heights / Total number of values = 2812.3 / 16 ≈ 175.77

* 1. Median:

To find the median, we'll first arrange the heights in ascending order: [172.5, 175, 175, 176, 176.2, 176.5, 177, 177, 178, 178, 178, 178.2, 178.9, 179, 180] Since we have an even number of values (16), the median will be the average of the two middle values. Median = (176.5 + 177) / 2 = 176.75

* 1. Mode:

The mode is the value that appears most frequently in the data. In this case, we can see that the height 178 appears the most number of times (4 times), making it the mode.

Therefore, the three measures of central tendency for the given height data are: Mean ≈ 175.77 Median = 176.75 Mode = 178

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]

Step 1: Find the mean

Add up all the values and divide by the total number of values: 178 + 177 + 176 + 177 + 178.2 + 178 + 175 + 179 + 180 + 175 + 178.9 + 176.2 + 177 + 172.5 + 178 + 176.5 = 2807.3 Total number of values = 16 Mean = 2807.3 / 16 = 175.45625 (rounded to 5 decimal places)

Step 2: Calculate the difference between each data point and the mean

Subtract the mean from each data point:

178 - 175.45625 = 2.54375

177 - 175.45625 = 1.54375

176 - 175.45625 = 0.54375

177 - 175.45625 = 1.54375

178.2 - 175.45625 = 2.74375

178 - 175.45625 = 2.54375

175 - 175.45625 = -0.45625

179 - 175.45625 = 3.54375

180 - 175.45625 = 4.54375

175 - 175.45625 = -0.45625

178.9 - 175.45625 = 3.44375

176.2 - 175.45625 = 0.74375

177 - 175.45625 = 1.54375

172.5 - 175.45625 = -2.95625

178 - 175.45625 = 2.54375

176.5 - 175.45625 = 1.04375

Step 3: Square each of the differences obtained in step 2

2.54375^2 = 6.470953125

1.54375^2 = 2.383486328125

0.54375^2 = 0.296044921875

1.54375^2 = 2.383486328125

2.74375^2 = 7.522185546875

2.54375^2 = 6.470953125

-0.45625^2 = 0.208203125

3.54375^2 = 12.542880859375

4.54375^2 = 20.634548828125

-0.45625^2 = 0.208203125

3.44375^2 = 11.868791015625

0.74375^2 = 0.552927734375

1.54375^2 = 2.383486328125

-2.95625^2 = 8.75850390625

2.54375^2 = 6.470953125

1.04375^2 = 1.089986328125

Step 4: Find the mean of the squared differences Add up all the squared differences and divide by the total number of values (16):

(6.470953125 + 2.383486328125 + 0.296044921875 + 2.383486328125 + 7.522185546875 + 6.470953125 + 0.208203125 + 12.542880859375 + 20.634548828125 + 0.208203125 + 11.868791015625 + 0.552927734375 + 2.383486328125 + 8.75850390625 + 6.470953125 + 1.089986328125) / 16 ≈ 6.08394921875

Step 5: Take the square root of the mean squared differences Calculate the square root of the mean obtained in step 4: √(6.08394921875) ≈ 2.46572036688

Therefore, the standard deviation for the given data is approximately 2.4657 (rounded to 4 decimal places).

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

the spread of a dataset? Provide an example.

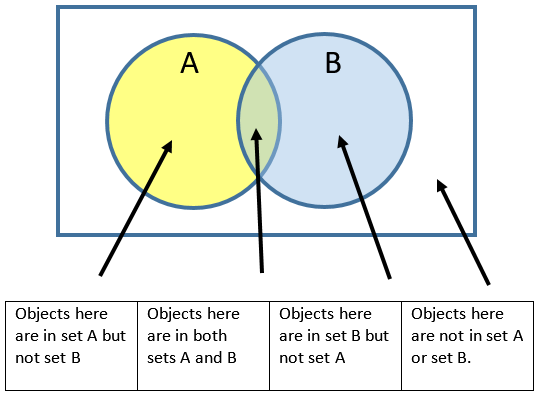
Measures of dispersion, including range, variance, and standard deviation, are used to describe the spread or variability of a dataset. They provide insights into how the values within the dataset are dispersed around the central tendency (mean or median). Here's an example to illustrate their usage:

Let's consider a dataset representing the ages of a group of individuals in a sample: {20, 22, 25, 28, 30}.

1. Range: The range is the simplest measure of dispersion and represents the difference between the maximum and minimum values in the dataset. In this case, the range would be 30 - 20 = 10. It gives an idea of the spread by indicating the extent of the age values.
2. Variance: The variance measures the average squared deviation of each data point from the mean. To calculate the variance, we follow these steps: a. Calculate the mean: (20 + 22 + 25 + 28 + 30) / 5 = 25. b. Calculate the squared deviation for each data point: (20 - 25)^2, (22 - 25)^2, (25 - 25)^2, (28 - 25)^2, (30 - 25)^2. c. Sum up the squared deviations: (5^2 + 3^2 + 0^2 + 3^2 + 5^2) = 50. d. Divide the sum by the number of data points: 50 / 5 = 10. The variance in this example is 10. It provides a measure of how the ages deviate from their mean, indicating the spread of the dataset.
3. Standard Deviation: The standard deviation is the square root of the variance and provides a measure of dispersion in the original unit of the dataset. To calculate the standard deviation, we take the square root of the variance. In this case, the standard deviation would be √10 ≈ 3.16. It gives an idea of how much the ages typically deviate from the mean and represents the spread of the dataset.

In summary, the range, variance, and standard deviation help quantify the spread of a dataset. While the range gives a simple measure of the total spread, the variance and standard deviation provide more detailed information about the dispersion and how individual data points differ from the mean.

Q6. What is a Venn diagram?



A Venn diagram is a visual representation of the relationships between different sets of items. It consists of overlapping circles or other shapes that are used to illustrate the common and distinct elements of multiple sets. The diagram was named after the English mathematician John Venn, who introduced it in the late 19th century as a way to visualize set theory concepts.

Each circle in a Venn diagram represents a set, and the overlapping regions indicate the elements that are shared between the sets. The areas outside the circles represent elements that are not part of any of the sets being compared. Venn diagrams can be used to analyze and compare various relationships, such as similarities and differences, intersections, and unions of sets.

Venn diagrams are commonly used in mathematics, logic, statistics, and other fields to illustrate concepts and solve problems related to set theory and logic. They can also be used in more general contexts to represent relationships and overlaps between different categories or groups. Venn diagrams can be hand-drawn or created using software tools for more precise and complex representations.

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

To find the set operations between sets A and B, we'll consider the given sets:

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

(i) A ∩ B (Intersection of A and B): The intersection of two sets A and B includes the elements that are common to both sets.

A ∩ B = {x | x ∈ A and x ∈ B}

In this case, the common elements between A and B are {2, 6}.

(ii) A ⋃ B (Union of A and B): The union of two sets A and B is the set that contains all the elements from both sets without any duplicates.

A ⋃ B = {x | x ∈ A or x ∈ B}

In this case, the union of A and B is {0, 2, 3, 4, 5, 6, 7, 8, 10}.

Q8. What do you understand about skewness in data?

Skewness is a statistical measure that describes the asymmetry of a probability distribution or a dataset. It indicates the degree to which the data deviates from a symmetric distribution. In other words, it measures the extent to which the data is skewed or lopsided.

Skewness can be quantified using different mathematical formulas, but the most commonly used measure is Pearson's first coefficient of skewness. It is calculated by taking the third standardized moment of the distribution, which involves subtracting the mean from each data point, dividing by the standard deviation, and then cubing the result.

The resulting value of skewness can be positive, negative, or zero:

1. Positive skewness (right-skewed or positively skewed): If the skewness value is greater than zero, it indicates that the tail of the distribution extends more towards the right side (higher values) than the left side (lower values). In other words, the distribution has a long tail on the right.
2. Negative skewness (left-skewed or negatively skewed): If the skewness value is less than zero, it means that the tail of the distribution extends more towards the left side (lower values) than the right side (higher values). In this case, the distribution has a long tail on the left.
3. Zero skewness: If the skewness value is close to zero, it suggests that the distribution is approximately symmetric, with the tails being roughly equal in length.

Skewness is an important measure because it provides insights into the shape and characteristics of a dataset. It helps us understand the presence and direction of asymmetry, which can impact various statistical analyses and modeling techniques. Additionally, skewness is often considered along with measures of central tendency (such as mean and median) and dispersion (such as standard deviation) to gain a more comprehensive understanding of the data's distribution.

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

If a data set is right-skewed, it means that the distribution of the data is skewed to the right, with a longer tail on the right side. In such a distribution, the mean is generally pulled towards the right by the presence of extreme values in that direction.

In terms of the position of the median with respect to the mean, the median is typically located closer to the left side (lower values) compared to the mean. This is because the median is less influenced by extreme values in the right tail, which pull the mean towards them. Therefore, in a right-skewed distribution, the median is generally smaller than the mean.

To summarize:

* Right-skewed distribution
* Median is typically smaller than the mean

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

Covariance and correlation are both statistical measures that describe the relationship between two variables. While they are related, they have distinct interpretations and uses in statistical analysis.

Covariance: Covariance measures how two variables vary together. It indicates the direction (positive or negative) and the magnitude of the linear relationship between two variables. A positive covariance suggests that when one variable increases, the other tends to increase as well, while a negative covariance suggests that as one variable increases, the other tends to decrease. However, covariance alone does not provide a standardized measure of the strength of the relationship.

Covariance is calculated using the following formula:

Cov(X, Y) = Σ[(X - μX) \* (Y - μY)] / (n - 1)

where X and Y are the variables, μX and μY are their respective means, and n is the number of data points.

Correlation: Correlation, on the other hand, provides a standardized measure of the linear relationship between two variables. It quantifies the strength and direction of the relationship between two variables, regardless of the scale of the variables. Correlation values range between -1 and +1, where -1 indicates a perfect negative linear relationship, +1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.

Correlation is calculated using the following formula:

Corr(X, Y) = Cov(X, Y) / (σX \* σY)

where Cov(X, Y) is the covariance between X and Y, and σX and σY are the standard deviations of X and Y, respectively.

Usage in Statistical Analysis: Covariance and correlation are both used in statistical analysis to understand the relationship between variables and to assess their association. Here's how they are used:

* 1. Covariance:
* Covariance is used to determine the direction (positive or negative) and magnitude of the relationship between two variables.
* It helps in identifying the degree of linear dependence between variables.
* Covariance is used in some statistical models, such as linear regression, to estimate the relationship between independent and dependent variables.
  1. Correlation:
* Correlation is widely used to measure the strength and direction of the linear relationship between two variables.
* It helps in determining the degree of association between variables, regardless of their scales.
* Correlation coefficients are used to identify patterns, trends, and dependencies in datasets.
* Correlation is used in various fields, including finance, economics, social sciences, and data analysis, to assess relationships and make predictions.

In summary, while covariance describes the direction and magnitude of the linear relationship between two variables, correlation provides a standardized measure of the strength and direction of the relationship. Correlation is often preferred because it allows for comparisons between different pairs of variables and is not influenced by the scale of the variables.

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

The formula for calculating the sample mean is:

Sample Mean = (Sum of all values in the dataset) / (Number of values in the dataset)

Here's an example calculation:

Let's consider the dataset: 5, 8, 2, 9, 4

To calculate the sample mean, we sum up all the values in the dataset: 5 + 8 + 2 + 9 + 4 = 28

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

Sample Mean = 28 / 5 = 5.6

Therefore, the sample mean for this dataset is 5.6.

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

For a normal distribution, the three measures of central tendency -- the mean, median, and mode -- are all equal to each other.

1. Mean: The mean of a normal distribution is the arithmetic average of all the data points. It represents the balancing point of the distribution, and it is the most commonly used measure of central tendency.
2. Median: The median is the middle value in a sorted list of data points. In a normal distribution, the median is equal to the mean. This implies that half of the data points are below the median, and half are above it.
3. Mode: The mode is the most frequently occurring value in a dataset. In a normal distribution, there is a single mode, and it is equal to both the mean and the median. This indicates that the distribution is symmetric around the mode, with the same number of values on both sides.

In summary, for a normal distribution, the mean, median, and mode are all equal, representing the center of the distribution.

Q13. How is covariance different from correlation?

Covariance and correlation are both statistical measures that describe the relationship between two variables, but they have some key differences.

Covariance is a measure of how much two variables vary together. It quantifies the degree to which changes in one variable are associated with changes in another variable. Specifically, it measures the average of the products of the differences between corresponding values of the two variables and their respective means. Covariance can take any value, positive or negative, and its magnitude indicates the strength of the relationship between the variables. However, it does not provide a standardized measure of the relationship.

Correlation, on the other hand, measures the strength and direction of the linear relationship between two variables. It is a standardized measure, ranging between -1 and +1, that is calculated by dividing the covariance by the product of the standard deviations of the two variables. Correlation allows for a more meaningful comparison of the relationship between variables, as it provides a standardized scale. A correlation of +1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

In summary, covariance indicates the direction of the relationship between two variables (positive or negative), while correlation measures the strength and direction of the linear relationship in a standardized manner.

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

Outliers can have a significant impact on measures of central tendency and dispersion. Central tendency measures, such as the mean, median, and mode, are used to represent the typical or average value of a dataset. Outliers are extreme values that deviate significantly from the majority of the data points and can distort these measures.

When outliers are present, the mean, which is the arithmetic average, can be heavily influenced by their extreme values. Since the mean takes into account every data point, even a single outlier can cause a significant shift in its value. For example, consider a dataset of exam scores: {85, 90, 92, 88, 95, 100, 52}. The outlier value of 52 is much lower than the other scores and will drag down the mean, making it an inaccurate representation of the overall performance.

The median, which is the middle value when the data is sorted, is less affected by outliers compared to the mean. However, if the dataset has a substantial number of outliers, the median may still be influenced and deviate from the typical value of the majority of the data points. In the previous example, the median would be less affected by the outlier since it disregards the extreme values and would be closer to the central tendency of the majority of the scores.

Outliers also impact measures of dispersion, such as the range, variance, and standard deviation, which quantify the spread or variability of the data. These measures are sensitive to extreme values because they consider the differences between individual data points.

For example, consider a dataset of income levels in a small town: {20,000, 25,000, 30,000, 40,000, 300,000}. The presence of the outlier value of 300,000, which is much higher than the other incomes, will result in a larger range, variance, and standard deviation. These measures will overstate the variability in income levels in the town because of the extreme outlier value.

In summary, outliers can have a substantial impact on measures of central tendency and dispersion. They can distort the mean, while the median is more robust but can still be influenced. Measures of dispersion are particularly sensitive to outliers, leading to an overestimation of variability. It is crucial to identify and handle outliers appropriately to ensure accurate statistical analysis and interpretation.