Statistics Assignment

January 12, 2024

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

1.0.1 ANS:

- The three measures of central tendency are mean, median, and mode. The mean is the average value of a dataset, calculated by adding up all the values in the dataset and dividing by the number of values. The median is the middle value of a dataset when it is arranged in ascending or descending order. The mode is the value that appears most frequently in a dataset. These measures are used to describe the central location of a dataset and provide insight into its distribution

[ ]:

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

2.0.1 ANS:

- The mean, median, and mode are the three measures of central tendency used to describe the central location of a dataset and provide insight into its distribution. The mean is the average value of a dataset, calculated by adding up all the values in the dataset and dividing by the number of values. The median is the middle value of a dataset when it is arranged in ascending or descending order. The mode is the value that appears most frequently in a dataset.

- The mean is the most commonly used measure of central tendency, but it can be heavily influenced by extreme values or outliers. In contrast, the median and mode are less sensitive to outliers and are better suited for skewed datasets . The median is often used when the dataset has extreme values or outliers, while the mode is used when the dataset has multiple peaks or is bimodal .

[ ]:

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]

[6]:

```python
import numpy as np
import pandas as pd

data = np.array([178, 177, 176, 177, 178.2, 178, 175, 179, 180, 175, 178.9, 176.
  ↪2, 177, 172.5, 178, 176.5])

mean_value = np.mean(data)
median_value = np.median(data)
mode_value = pd.Series(data).mode()

print(f"The mean of the dataset is {mean_value:.2f}")
print(f"The median of the dataset is {median_value:.2f}")
if len(mode_value) > 1:
    print(f"The modes of the dataset are {', '.join(map(str, mode_value))}")
else:
    print(f"The mode of the dataset is {mode_value[0]}")
```

```
The mean of the dataset is 177.02
The median of the dataset is 177.00
The modes of the dataset are 177.0, 178.0
```

[ ]:

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]

[11]:

```python
import numpy as np
data = np.std([178, 177, 176, 177, 178.2, 178, 175, 179, 180, 175, 178.9, 176.
  ↪2, 177, 172.5, 178, 176.5])
print(f"The standard deviation of the dataset is {data:.2f}")
```

```
The standard deviation of the dataset is 1.79
```

[ ]:

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.

5.0.1 ANS:

- Measures of dispersion such as range, variance, and standard deviation are used to describe the spread of a dataset . The range is the difference between the maximum and minimum values in a dataset and provides a rough estimate of the spread of the data . The variance and standard deviation are more precise measures of dispersion that describe how far the data is from the mean . The variance is calculated by taking the average of the squared differences between each data point and the mean of the dataset, while the standard deviation is the square root of the variance .
- For example, consider the following dataset of exam scores: 80, 85, 90, 95, 100. The mean of this dataset is 90. The range is 20 (100 - 80), which tells us that the spread of the data is 20 points. The variance is 62.5, and the standard deviation is 7.91. These measures tell us that the data is tightly clustered around the mean, with most of the scores falling within a few points of 90 .

[ ]:

6 Q6. What is a Venn diagram?

6.0.1 ANS:

- A Venn diagram is a diagram that shows all possible logical relations between a finite collection of different sets . It is used to visually represent the differences and similarities between two or more concepts, and is widely used in set theory, logic, mathematics, businesses, teaching, computer science, and statistics . Venn diagrams are also called logic or set diagrams.

[ ]:

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

7.0.1 ANS:

- (i) A B: The intersection of sets A and B is the set of all elements which are common to both A and B 12. Given A = {2, 3, 4, 5, 6, 7} and B = {0, 2, 6, 8, 10}, the intersection of A and B is {2, 6}.

Therefore, A B = {2, 6}.

- (ii) A B: The union of sets A and B is the set of all distinct elements that are in any of these sets 12. Given A = {2, 3, 4, 5, 6, 7} and B = {0, 2, 6, 8, 10}, the union of A and B 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?

8.0.1 ANS:

- Skewness is a measure of the asymmetry of a distribution . A distribution is said to be skewed if its left and right sides are not mirror images of each other . Skewness can be positive, negative, or zero. A positive skew indicates that the distribution has a long tail on the right side and most of the observations are concentrated on the left side of the distribution. A negative skew indicates that the distribution has a long tail on the left side and most of the observations are concentrated on the right side of the distribution. A zero skew indicates that the distribution is symmetric .

[ ]:

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

9.0.1 ANS:

- If a dataset is right-skewed, then the mean will be greater than the median . This is because the right-skewed distribution has a long tail on the right side, which pulls the mean towards the right. The median, on the other hand, is less affected by extreme values and is more representative of the central tendency of the dataset . Therefore, the median will be to the left of the mean in a right-skewed distribution.

[ ]:

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

10.0.1 ANS:

- Covariance and correlation are two statistical measures used to describe the relationship between two variables . The main difference between covariance and correlation is that covariance measures the direction of the linear relationship between two variables, while correlation measures both the strength and direction of the linear relationship between two variables.

- Covariance is a measure of how much two variables change together. It is calculated by taking the average of the product of the deviations of each variable from its mean . A positive covariance indicates that the two variables are positively related, while a negative covariance indicates that the two variables are negatively related .

- Correlation is a standardized measure of the linear relationship between two variables. It is calculated by dividing the covariance of the two variables by the product of their standard deviations . Correlation ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation .
- Covariance and correlation are used in statistical analysis to determine the relationship between two variables. They are used to identify patterns and trends in data, and to make predictions about future behavior . For example, in finance, covariance and correlation are used to measure the relationship between different stocks and to construct portfolios that minimize risk and maximize returns . In medicine, covariance and correlation are used to identify risk factors for diseases and to develop treatment plans .

[ ]:

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

11.0.1 ANS:

The formula for calculating the sample mean is:

$\bar{x} = (\Sigma xi) / n$

where $\bar{x}$ is the sample mean, Σ xi is the sum of all the values in the sample, and n is the number of values in the sample 1.

For example, consider the following dataset of 10 numbers: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20. To calculate the sample mean, we first add up all the values in the dataset:

$2 + 4 + 6 + 8 + 10 + 12 + 14 + 16 + 18 + 20 = 110$

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

$110 / 10 = 11$

Therefore, the sample mean of the dataset is 11.

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

12.0.1 ANS:

- For a normal distribution, all measures of central tendency such as mean, median, and mode are equal . This is because the normal distribution is symmetric and has a bell-shaped curve, with the mean, median, and mode all located at the center of the distribution . Therefore, the relationship between the measures of central tendency in a normal distribution is that they are all equal.

[ ]:

13 Q13. How is covariance different from correlation?

13.0.1 ANS:

- Covariance and correlation are two statistical measures used to describe the relationship between two variables . The main difference between covariance and correlation is that covariance measures the direction of the linear relationship between two variables, while correlation measures both the strength and direction of the linear relationship between two variables .
- Covariance is a measure of how much two variables change together. It is calculated by taking the average of the product of the deviations of each variable from its mean . A positive covariance indicates that the two variables are positively related, while a negative covariance indicates that the two variables are negatively related .
- Correlation is a standardized measure of the linear relationship between two variables. It is calculated by dividing the covariance of the two variables by the product of their standard deviations . Correlation ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation .
- Covariance and correlation are used in statistical analysis to determine the relationship between two variables. They are used to identify patterns and trends in data, and to make predictions about future behavior . For example, in finance, covariance and correlation are used to measure the relationship between different stocks and to construct portfolios that minimize risk and maximize returns. In medicine, covariance and correlation are used to identify risk factors for diseases and to develop treatment plans .

[ ]:

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

14.0.1 ANS:

- Outliers are extreme values that differ from most other data points in a dataset . Outliers can significantly affect measures of central tendency and dispersion, such as the mean, median, mode, range, variance, and standard deviation .
- For example, consider the following dataset of 10 numbers: 2, 4, 6, 8, 10, 12, 14, 16, 18, 200. The mean of this dataset is 26, while the median is 11. The presence of the outlier (200) has significantly affected the mean, pulling it towards the right side of the distribution. The median, on the other hand, is less affected by extreme values and is more representative of the central tendency of the dataset .
- Similarly, outliers can also affect measures of dispersion such as the range, variance, and standard deviation. Outliers can increase the range of the dataset, making it appear more spread out than it actually is. Outliers can also increase the variance and standard deviation of the dataset, making it more difficult to draw conclusions about the data .

[ ]: