7th Mar Assignment

March 15, 2023

1 Assignment 31

Q1. What are the three measures of central tendency?

Ans.The three measures of central tendency are:

1. Mean: The mean is the arithmetic average of a set of numbers. It is calculated by adding all the values in the set and then dividing the sum by the total number of values.
2. Median: The median is the middle value of a set of numbers. It is the value that separates the higher half of the values from the lower half. If there are an even number of values in the set, the median is the average of the two middle values.
3. Mode: The mode is the value that occurs most frequently in a set of numbers. In some cases, there may be multiple modes if there are several values that occur with the same frequency.

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

Ans.The mean is the sum of all the values in a dataset divided by the total number of values. It is sensitive to outliers, meaning that extreme values in the dataset can significantly affect the value of the mean. The mean is commonly used to summarize numerical data.

The median is the middle value of a dataset when the values are arranged in numerical order. It is less sensitive to outliers than the mean, and it is commonly used to summarize skewed data or data with extreme values.

The mode is the most frequently occurring value in a dataset. It is commonly used to summarize categorical or discrete data, such as the number of students in a classroom with different eye colors.

All three measures of central tendency provide different information about the center of a dataset and are used depending on the type of data and the research question. For instance, if a dataset has a normal distribution, the mean and median are similar, and both provide useful information about the center of the dataset. If a dataset has a skewed distribution, the median may be a better representation of the center, while the mode may be more useful for identifying the most common value in the dataset.

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]

Ans.

[1]:
```
data=[178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]
```

[2]:
```
import pandas as pd
```

[3]:
```
s1=pd.Series(data)
```

[4]:
```
s1
```

[4]:
```
0     178.0
1     177.0
2     176.0
3     177.0
4     178.2
5     178.0
6     175.0
7     179.0
8     180.0
9     175.0
10    178.9
11    176.2
12    177.0
13    172.5
14    178.0
15    176.5
dtype: float64
```

[5]:
```
s1.mean()
```

[5]:
```
177.01875
```

[8]:
```
s1.mode()[0]
```

[8]:
```
177.0
```

[9]:
```
s1.median()
```

[9]:
```
177.0
```

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.

[10]:
```
s2=pd.Series([178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.
↪5,178,176.5])
```

[11]:
```
s2.std()
```

[11]: 1.847238930584419

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.Range: The range is the difference between the highest and lowest values in a dataset. It provides a rough measure of the spread of the dataset, but it is sensitive to outliers.

Variance: The variance is the average of the squared differences of each value from the mean of the dataset. It measures how spread out the values are around the mean. A higher variance indicates a greater spread of the values from the mean.

Standard deviation: The standard deviation is the square root of the variance. It measures the typical distance of the values from the mean. A higher standard deviation indicates a greater spread of the values from the mean.

For example, consider a dataset of the ages of 10 people: [20, 22, 23, 25, 26, 27, 28, 29, 30, 35]. The mean age of this dataset is 27.5. We can calculate the range, variance, and standard deviation of the dataset using Python pandas:

[12]:
```
s3=pd.Series([20, 22, 23, 25, 26, 27, 28, 29, 30, 35])
```

[16]:
```
Range=s3.max()-s3.min()
Variance=s3.var()
standard_deviation=s3.std()
```

[17]:
```
Range
```

[17]: 15

[18]:
```
Variance
```

[18]: 18.944444444444443

[19]:
```
standard_deviation
```

[19]: 4.352521619066865

Q6. What is a Venn diagram?

Ans.A Venn diagram is a visual representation of the relationships between different sets of data. It consists of overlapping circles, where each circle represents a set, and the overlapping area represents the intersection of the sets. Venn diagrams are often used in mathematics, logic, statistics, and other fields to illustrate concepts such as set theory, logic, and probability.

The circles in a Venn diagram can be used to represent any category or group of data, and the intersection of the circles can represent any overlap between those categories or groups. For example, a Venn diagram can be used to show the relationships between different groups of animals, such as mammals, birds, and reptiles.

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

Ans.

[21]:
```
A={2,3,4,5,6,7}
B={0,2,6,8,10}
```

[23]:
```
A.intersection(B)
```

[23]:
```
{2, 6}
```

[24]:
```
A.union(B)
```

[24]:
```
{0, 2, 3, 4, 5, 6, 7, 8, 10}
```

Q8. What do you understand about skewness in data?

Ans.Skewness is a measure of the asymmetry of a probability distribution, or how much a dataset deviates from a symmetric distribution. In other words, it measures the degree to which the values in a dataset are concentrated on one side of the mean compared to the other side.

A dataset is considered skewed if it has a non-zero skewness value. There are two types of skewness: positive skewness and negative skewness.

- Positive skewness: A dataset is positively skewed if the tail of the distribution is longer on the right side than on the left side, and the mean is greater than the median. This means that there are more low values in the dataset, and a few high values that pull the mean to the right.
- Negative skewness: A dataset is negatively skewed if the tail of the distribution is longer on the left side than on the right side, and the mean is less than the median. This means that there are more high values in the dataset, and a few low values that pull the mean to the left.

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

Ans.If a data is right skewed, then the mean will be greater than the median. This is because the tail of the distribution is longer on the right side, which means there are more high values that pull the mean to the right.

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

Ans.Covariance measures how much two variables change together. It is a measure of the joint variability of two random variables. A positive covariance indicates that when one variable increases, the other variable tends to increase as well, and vice versa for a negative covariance. However, the magnitude of the covariance is difficult to interpret as it depends on the scale of the variables.

Correlation measures the strength and direction of the linear relationship between two variables, regardless of the scale of the variables. It is a standardized version of covariance, and it ranges between -1 and 1, with -1 indicating a perfect negative correlation, 0 indicating no correlation, and 1 indicating a perfect positive correlation. Correlation is more widely used than covariance because it provides a more interpretable measure of the strength of the relationship between variables.

In statistical analysis, both covariance and correlation are used to identify patterns and relationships between variables. Correlation is particularly useful in identifying the strength and direction of a relationship between two variables, and it is commonly used in regression analysis to model the relationship between a dependent variable and one or more independent variables. Covariance, on the other hand, is less commonly used due to its scale-dependent nature, but it is used in some statistical tests and can provide additional information about the joint variability of two variables.

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

Ans.The formula for calculating the sample mean is:

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

In other words, you add up all the values in the sample, and then divide by the number of values to get the average value.

For example, consider the following dataset: [3, 6, 7, 10, 12] #### To calculate the sample mean, we add up all the values in the dataset, then divide by the number of values: #### sample mean = (3 + 6 + 7 + 10 + 12) / 5 #### sample mean = 38 / 5 #### sample mean = 7.6 #### Therefore, the sample mean of this dataset is 7.6.

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

Ans.For a normal distribution, the mean, median, and mode are all equal. This means that the center of the distribution is at the same point regardless of which measure of central tendency is used. This property of normal distribution is often referred to as symmetry.

The normal distribution is a bell-shaped curve, with the highest point at the center of the distribution. The mean, median, and mode are all located at this highest point, which is also the point of maximum density. Because the normal distribution is symmetric, the mean is located at the center of the distribution, with half the data values falling to the left of the mean and half falling to the right. The median, which is the middle value of the dataset, is also located at the center of the distribution, as is the mode, which is the value that occurs most frequently in the dataset.

Q13. How is covariance different from correlation?

Ans.Covariance and correlation are both measures used in statistical analysis to describe the relationship between two variables, but they differ in a few key ways:

- Scale dependency: Covariance is scale-dependent, meaning it is influenced by the scale of the variables being measured. In contrast, correlation is scale-independent, and is not affected by the units of measurement.
- Magnitude: Covariance can take on any value between negative infinity and infinity, depending on the magnitude of the joint variability of the two variables. In contrast, correlation always ranges between -1 and 1, with 1 indicating a perfect positive correlation, -1 indicating a perfect negative correlation, and 0 indicating no correlation.
- Interpretability: Correlation is more interpretable than covariance because it is scale-independent and ranges between -1 and 1. The magnitude of the correlation coefficient provides a measure of the strength of the linear relationship between two variables, whereas the magnitude of the covariance is difficult to interpret as it depends on the scale of the variables.

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

Ans.Outliers are extreme values that lie far away from the other values in a dataset. They can have a significant impact on measures of central tendency and dispersion.

Outliers can affect measures of central tendency, such as the mean and median, by pulling the center of the distribution towards themselves. If an outlier is much larger or smaller than the other values in the dataset, it can greatly increase or decrease the mean. The median is less affected by outliers, but it can still be shifted if the outlier is far enough from the other values.

[25]:
```
sample_values1 = pd.Series([25, 27, 28, 30, 32, 33, 34, 35, 36, 40, 100])
sample_values2 = pd.Series([25, 27, 28, 30, 32, 33, 34, 35, 36, 40])
```

[26]:
```
sample_values1.mean()
```

[26]: 38.18181818181818

[27]:
```
sample_values2.mean()
```

[27]: 32.0

[34]:
```
sample_values1.median()
```

[34]: 33.0

[35]:
```
sample_values2.median()
```

[35]: 32.5

Mean is highly affected by outliers but median is slightly affected