Statistics_Basic_Assignment_2

Q1. What are the three measures of central tendency?

1. Three measures of Central tendency are :
   A. Mean
   B. Median
   C. Mode
2. Mean --> It is the average of all the values in a data or it is also known as sum of all values in a dataset divided by the total number of values,it is sensitive to outliers.
3. Median --> It is the middle value in a dataset when it is arranged in ascending or descending order,it is not sensitive to outliers.
4. Mode --> It is the most frequently occuring value in a dataset.
5. Mean,Median,Mode i.e., Measures of Central tendency are used to summarize data.
6. It also has its own disadvantages as the mean can be affected if the data has outliers whereas the median affects less than the mean but the outliers also affects the mode.

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

1. Measures of Central tendency :
   A. Mean
   B. Median
   C. Mode
2. Mean -> The mean is used to find the average of all the values in a dataset.
3. Median -> The median is used to find the middle value in a dataset.
4. Mode -> The mode is used to find the most frequent value in the dataset.
5. Mean,Median,Mode are the measures of Central tendancy which are the part of Descriptive Statistics used to organize and summarize the data to describe the data in depth.
6. Among Mean,Median,Mode the Mean and Mode are sensitive to outliers i.e.,they are affected if the outliers are present in the data, the Median is not sensitive to outliers i.e.,it is not affected although the outliers are present in 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]

In [5]:
```
import numpy as np
from scipy import stats
data = [178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,17
# To find Mean
print(np.mean(data))
# To find Median
print(np.median(data))
# To find Mode
print(stats.mode(data))
```

```
177.01875
177.0
ModeResult(mode=array([177.]), count=array([3]))

C:\Users\yashg\AppData\Local\Temp\ipykernel_13672\755960063.py:9: FutureWarni
ng: Unlike other reduction functions (e.g. `skew`, `kurtosis`), the default b
ehavior of `mode` typically preserves the axis it acts along. In SciPy 1.11.
0, this behavior will change: the default value of `keepdims` will become Fal
se, the `axis` over which the statistic is taken will be eliminated, and the
value None will no longer be accepted. Set `keepdims` to True or False to avo
id this warning.
  print(stats.mode(data))
```

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]

In [10]:
```
import numpy as np
data = [178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,17
print("Standard deviation : {}".format(np.std(data)))
```

```
Standard deviation : 1.7885814036548633
```

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

1. Measures of Variability :
   A. Range
   B. Variance
   C. Standard deviation
2. MEASURES OF VARIABILITY
   A. Range -> Range is used to find the difference between the highest and lowest values in a dataset.It is a simple measure of how spread out the data is.
   B. Variance -> The variance is used to find the measure of how far the values in a dataset are spread out from the mean. Basically,it is used to find out the spread of data.
   C. Standard deviation -> The standard deviation also tells about how the data is spreaded along with median,it is more interpretable measure of variability than the variance, because it is in the same units as the data.

** Example **

Range: The range is the simplest measure of dispersion. It is the difference between the largest and smallest values in a data set. For example, if the heights of a group of people are 5 feet, 6 feet, 6 feet, 7 feet, and 7 feet, then the range is 2 feet.

Variance: The variance is a more complex measure of dispersion. It is calculated by taking the average of the squared deviations from the mean for all values in a data set. For example, if the heights of the people in the previous example are all 6 feet, then the variance is 0. However, if the heights are 5 feet, 6 feet, 6 feet, 7 feet, and 8 feet, then the variance is 2.25.

Standard deviation: The standard deviation is the square root of the variance. It is a more commonly used measure of dispersion than the variance because it is easier to interpret. For example, the standard deviation for the heights of the people in the previous example is 1.5.

Q6. What is a Venn diagram?

A Venn diagram is a widely used diagram style that shows the logical relation between sets,popularized by John Venn(1834-1923) in the 1880's.

A Venn diagram is a diagram that shows all possible logical relations between a finite collection of different sets.

Venn diagrams are used to visualize the similarities an differences between two or more sets of data.

Simple Venn diagrams consist of two overlapping circles, but complex Venn diagrams may comapare up to five or more circles.

Venn diagrams are easy to understand and interpret, they can be used to visualize complex data sets, they can used to communicate complex ideas, and are versatile tool that can be used in a variety of different contexts.

Example of venn diagrams are :

1. To comapare and contrast two different types of birds.
2. To show the different types of customers that a company has.
3. To illustrate the different stages of product development process.
4. To show the relationship between different concepts in a logical argument.

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

In [13]:
```
A = {2,3,4,5,6,7}
B = {0,2,6,8,10}
# A B
# A ∪ B
```

In [14]:
```
# Union
A|B
```

Out[14]:
```
{0, 2, 3, 4, 5, 6, 7, 8, 10}
```

In [15]:
```
# Intersection
A&B
```

Out[15]:
```
{2, 6}
```

Q8. What do you understand about skewness in data?

1. Skewness is a measure of asymmetry of a distribution.
2. A distribution is asymmetrical when its right and left side are not mirror images of each other.
3. Skewness can be positive,negative or zero.
4. Positive skewness occurs when the tail of the distribution is longer on the right side. This means that there are more data points with higher values than lower values.
5. Negative skewness occurs when the tail of the distribution is longer on the left side. This means that there are more data points with lower values than higher values.
6. Zero skewness occurs when the distribution is symmetrical. That is the data points are evenly distributed.
7. Skewness can be used to understand the shape of the distribution and to identify outliers. Outliers are the data points which are very far away from the rest of the data and can skew the distribution, so, it is important to identify them.
8. Skewness can be used to determine if the variable is normally distributed,identifying outliers,comparing different distributions,improving the accuracy of statistical models.

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

> If the distribution of data is skewed to right, the mode is often less than the median,which is less than the mean.

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

1. Covariance and correlation are both statistical measures that quantify the relationship between two random variables.
2. Covariance measures the extent to which two variables vary together.It is a measure of how much the two variables change in tandem,regardless of whether they change in the same direction or in opposite directions.
3. The covariance of two variables can be positive,negative or zero.
4. Correlation measures the strength and direction of the linear relationship between two variables.
5. It is a standardized measure of covariance,which means that it is always between -1 to +1.
6. A correlation of +1 indicates perfect positive relationship,a correlation of -1 indicates perfect negative relationship and 0 indicates linear relationship.
7. There are two main types of correlation :
   A. Pearson correlation
   B. Spearman Rank correlation
8. Covariance and Correlation are used to identify relationships between variables,to measure the strength of the relationship between two variables(a larger coefficient indicates a stronger relationship),to make predictions.
9. For example -> A financial analyst might use covariance to identify stocks that tend to move in the same direction. This information could be used to construct a portfolio that is less risky than a portfolio that is randomly diversified.

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

1. Formula for calculating sample mean :

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

where,

$\bar{x}$ = sample mean

Σxi = sum of all the data

n = no. of items

In [1]:
```
import seaborn as sns
df = sns.load_dataset('iris')
```

In [2]:
```
df.head()
```

Out[2]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [8]:
```
df_sample = df['sepal_length'][:40]
```

In [9]:
```
df_sample
```

Out[9]:
```
0     5.1
1     4.9
2     4.7
3     4.6
4     5.0
5     5.4
6     4.6
7     5.0
8     4.4
9     4.9
10    5.4
11    4.8
12    4.8
13    4.3
14    5.8
15    5.7
16    5.4
17    5.1
18    5.7
19    5.1
20    5.4
21    5.1
22    4.6
23    5.1
24    4.8
25    5.0
26    5.0
27    5.2
28    5.2
29    4.7
30    4.8
31    5.4
32    5.2
33    5.5
34    4.9
35    5.0
36    5.5
37    4.9
38    4.4
39    5.1
Name: sepal_length, dtype: float64
```

In [10]:
```
len(df_sample)
```

Out[10]:
```
40
```

In [12]:
```
sum = 0
for i in range(len(df_sample)):
    sum = sum+df_sample[i]
print(sum)
```

```
201.49999999999997
```

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

1. In a normal distribution, the mean,median and mode are equal.This is because the normal distribution is symmetrical and the mean,mode and median are all located at center of the distribution.
2. In a normal distribution the mean, mode and median are equal means there are equal no. of values above and below the mean, and the distribution is not skewed in any direction.
3. The mean is the average of all the values in the distribution,the median is the value that falls in the middle of the distribution, when all the values are arranged in ascending or descending order and the mode is the value that appears most frequrently in the distribution.

Q13. How is covariance different from correlation?

1. Covariance and correlation are both statistical measures that quantify the relationship between two random variables.
2. Covariance could be positive or negative depending upon the nature of the dataset.
3. If both the parameters x and y are increasing or decreasing then the covariance is positive and if one of the parameter is positive and another is negative or vice versa then the covariance is negative.
4. With the help of covariance, we can find the relationship between two variables x and y(+ve or -ve) i.e. quantified relationship between x and y.
5. Covariance does not have a specific limit value, whereas to overcome this problem correlation is used.
6. Correlation is a standardized measure of covariance i.e. it is always between -1 to +1.
7. A correlation of +1 indicates perfect positive relationship,-1 indicates perfect negative relationship and 0 indicates linear relationship.
8. There are two types of Correlation :
    A. Pearson Correlation - In this,we focus on the values,used for linear data.
    B. Spearman Rank Correlation - In this,we focus on the rank between values,used for non-linear data.

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

1. Outliers are the data points that are very large or different from the rest of the data.
2. Mean is sensitive to outliers among the other measures of central tendency,whereas the median is not sensitive to ouliers and outliers do not affect mode at all.
3. Outliers also affects measures of dispersion as they can make the data points more spread out.

In [19]:
```
# For example ->
import numpy as np
from scipy import stats
data = [3,6,9,12,15,18,21,24,100]
# Here,in the above data '100' is an outlier
# Finding mean
print('Mean = {}'.format(np.mean(data)))
print('Median = {}'.format(np.median(data)))
print('Mode = {}'.format(stats.mode(data,keepdims=True)))
print('Standard deviation = {}'.format(np.std(data)))
print('Variance = {}'.format(np.var(data)))
```

```
Mean = 23.11111111111111
Median = 15.0
Mode = ModeResult(mode=array([3]), count=array([1]))
Standard deviation = 27.94615634252746
Variance = 780.9876543209876
```

Observation

In the above example, '100' is an outlier which affects some of the methods of measures of central tendency and dispersion.We can see that the outliers affects the value of Mean and also Variance

In [ ]: