# Statistics Basic-2 ¶

## Q.1 What are the three measures of central tendency?

**Ans:** The three central tendency measures are:

**Mean:** The mean is the average of a set of values. It is calculated by summing all the values in a dataset and then dividing the sum by the total number of values.

**Median:** The median is the middle value in a dataset when the values are arranged in ascending or descending order. If the dataset has an even number of values, the median is calculated by taking the average of the two middle values.

**Mode:** The mode is the value that appears most frequently in a dataset. If there are multiple values that appear with the same frequency, then the dataset is said to have multiple modes.

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

**Ans:** Calculation Usefulness use to measure central tendency

**Mean**

- Calculation
  - Sum of all values / Total number of values
- Usefulness
  - When the data is normally distributed or symmetrical
- It represents the arithmetic average of the dataset

**Median**

- Calculation
  - Middle value when data is ordered from lowest to highest or highest to lowest
- Usefulness
  - When the data has extreme values or is skewed
  - It represents the value that splits the dataset in half

**Mode**

- Calculation
  - Value that occurs most frequently in a dataset
- Usefulness
  - When we want to know which value occurs most frequently in the data.
  - It represents 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:

In [1]:

```
import numpy as np
height = [178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]
height
```

Out[1]:

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

**mean**

In [2]:

```
np.mean(height)
```

Out[2]:

```
177.01875
```

**median**

In [3]:

```
np.median(height)
```

Out[3]:

```
177.0
```

**mode**

In [4]:

```
from scipy import stats
stats.mode(height)
```

Out[4]:

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

## Q.4 : 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]

Answer :

In [5]:

```
import numpy as np
data = [178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]
np.std(data)
```

Out[5]:

```
1.7885814036548633
```

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

Answer : Measures of dispersion, such as range, variance, and standard deviation, are used to describe how spread out the data in a dataset is. Here's how each measure is used:

1. **Range:** The range is the difference between the maximum and minimum values in a dataset. It gives an idea of how much the values in the dataset vary from one another. A wider range indicates a more spread-out dataset, while a smaller range indicates a more tightly clustered dataset.

   For example, consider the following set of data: 10, 20, 30, 40, 50 , 60, 70, 80, 90, 100.

   The range is 50 - 10 = 40

   which means the data spans a range of 40 units.

   ```
   data = [10,20,30,40,50,60,70,80,90,100]

   range = max(data)-min(data)

         = 100 - 10
         = 90
   ```

2. Variance: Variance is a measure of how much the values in a dataset deviate from the mean. It is calculated by taking the sum of the squared differences between each value and the mean, divided by the total number of values. A higher variance indicates that the data is more spread out, while a lower variance indicates that the data is more tightly clustered around the mean. For example data: 10, 20, 30, 40, 50. The mean is 30. The variance is calculated as follows: variance = [(10 - 30)^2 + (20 - 30)^2 + (30 - 30)^2 + (40 - 30)^2 + (50 - 30)^2] / 5

   ```
            = 200
   ```

In [6]:

```
data = [ 10, 20, 30, 40, 50]
np.mean(data)
```

Out[6]:

```
30.0
```

In [24]:

```
np.var(data)
```

Out[24]:

```
200.0
```

3. Standard deviation: The standard deviation is the square root of the variance and is expressed in the same units as the data. It is a more intuitive measure of dispersion because it is in the same units as the data. A higher standard deviation indicates that the data is more spread out, while a lower standard deviation indicates that the data is more tightly clustered around the mean. For example, using the same set of data as above, the standard deviation is calculated as follows: standard deviation = sqrt(variance)

```
= sqrt(200)
=  14.142135623730951
```

In [7]:

```
np.std(data)
```

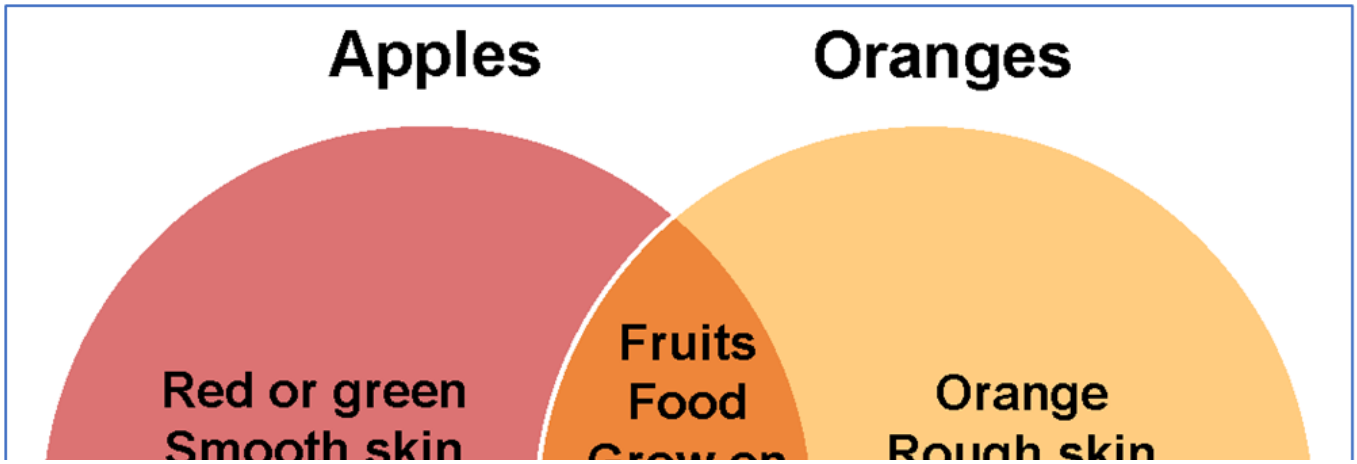
Out[7]:

```
14.142135623730951
```

## Q.6 What is a Venn diagram?

Ans: A Venn diagram is a visual representation of the relationships between sets of data. It is a diagram that shows all possible logical relations between a finite collection of sets, often represented as circles that overlap or are disjoint. Venn diagrams are commonly used to illustrate simple set relationships in logic, statistics, probability, and computer science.

In a Venn diagram, each circle represents a set, and the region where the circles overlap represents the intersection of those sets. The areas outside the circles represent the complement of those sets.




## 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:

In [8]:

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

intersection = A.intersection(B)
intersection
```

Out[8]:

```
{2, 6}
```

In [9]:

```
union = A.union(B)
union
```

Out[9]:

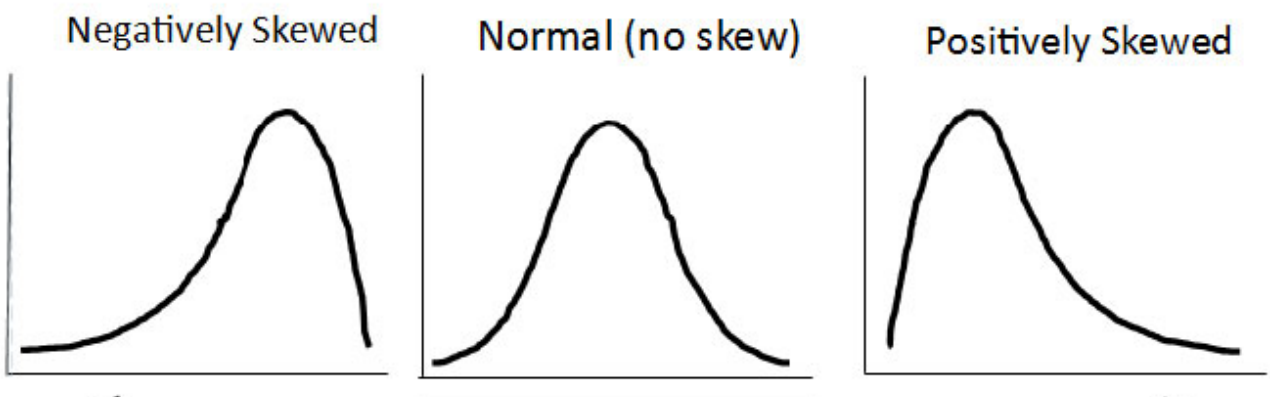
```
{0, 2, 3, 4, 5, 6, 7, 8, 10}
```

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

Ans:

In statistics, skewness is a measure of the asymmetry of a probability distribution. Skewness refers to the degree to which the data is not symmetrically distributed around the mean. It indicates whether the data is skewed to the left or to the right of the mean.

If a distribution is symmetrical, then it has zero skewness. If the tail of the distribution is longer on the right side than on the left side, then the distribution is said to be positively skewed, and it has a positive skewness. On the other hand, if the tail of the distribution is longer on the left side than on the right side, then the distribution is said to be negatively skewed, and it has a negative skewness.



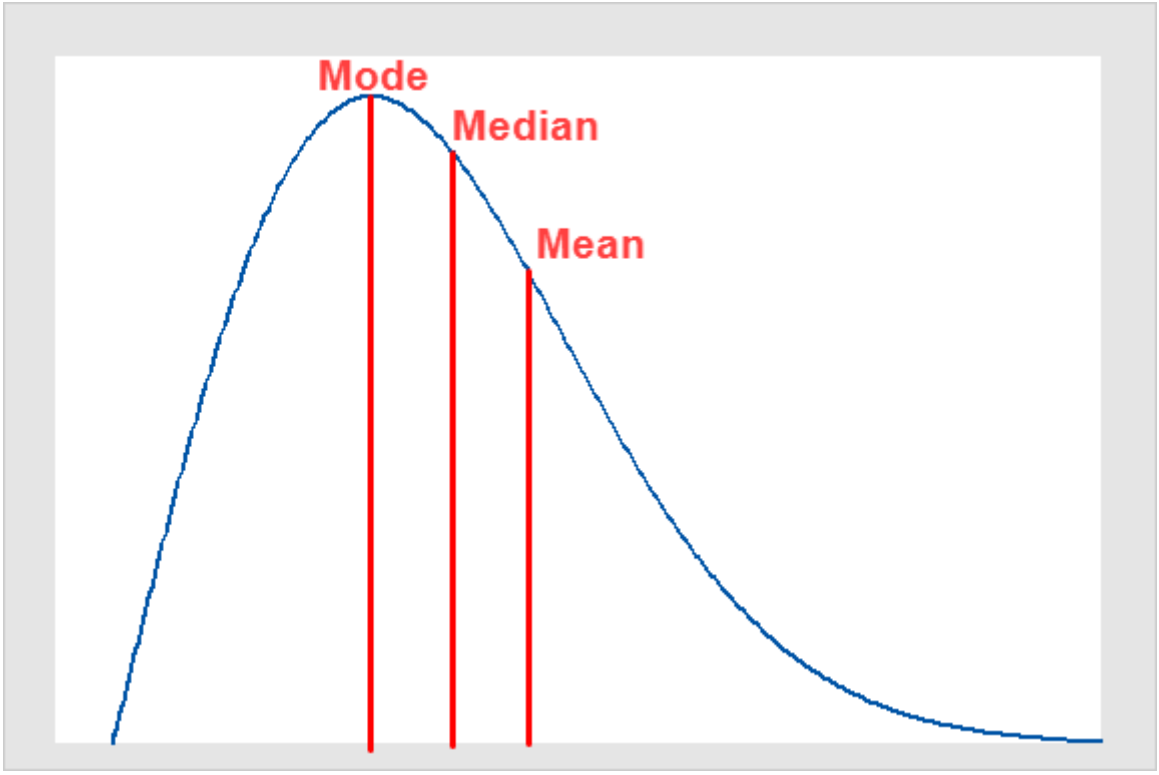

## Q9. If a data is right skewed then what will be the position of median with respect to mean?

Ans:

If a data set is right-skewed, it means that the tail of the distribution is longer on the right side, and the data is concentrated on the left side of the distribution. In this case, the mean is greater than the median.

For Example : consider a right-skewed distribution like income. In this distribution, there are a small number of individuals who earn very high incomes, which pull the mean to the right. However, most people earn lower incomes, so the median is not as affected by the high earners and remains closer to the center of the distribution.

Therefore, the median will be to the left of the mean in a right-skewed distribution.




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

Ans:

Covariance and correlation are two statistical measures that are used to quantify the relationship between two variables.

1. **Covariance**

   it is a measure of how two variables vary together. It measures the degree to which two variables are linearly related. A positive covariance indicates that the two variables move in the same direction, while a negative covariance indicates that they move in opposite directions. However, the magnitude of covariance does not have a standardized scale and it can be affected by the scale of the variables being measured. the value of the covariance can range from negative infinity to positive infinity.

2. **Correlation**

   on the other hand, is a measure of the strength and direction of the linear relationship between two variables. Correlation ranges from -1 to +1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and +1 indicates a perfect positive correlation. Correlation is standardized, which means that it is not affected by the scale of the variables being measured.
   The correlation coefficient between two variables X and Y is calculated as follows:
   corr(X,Y) = cov(X,Y) / (SD[X]*SD[Y])

In statistical analysis, covariance and correlation are used to assess the relationship between two variables. They help to identify patterns and trends in the data, and to determine the strength and direction of the relationship between the variables. They are also used to test hypotheses and to make predictions about the

## 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 observations) / (Number of observations)
In other words, the sample mean is calculated by adding up all the values in a sample and dividing by the number of observations in the sample.

$$\bar{x} = (x_1 + x_2 + ... + xn) / n$$

where $x_1$, $x_2$, ..., xn are the values in the dataset, and n is the number of values in the sample.

For Example:

Sample Mean = (4 + 7 + 9 + 12 + 15) / 5 = 47 / 5 = 9.4

In [10]:

```
data = [4,7,9,12,15]
np.mean(data)
```

Out[10]:

9.4

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

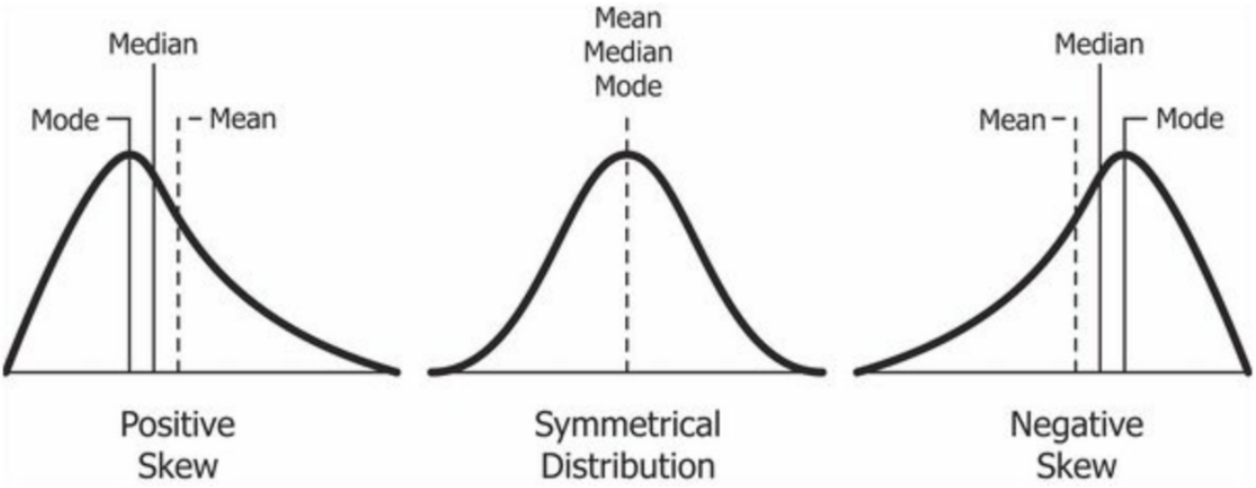
Ans:

For a normal distribution, the measures of central tendency, namely the mean, median, and mode, are all equal. In other words, they have the same value.

This is because a normal distribution is a symmetric distribution, with equal amounts of data on both sides of the mean. The median, which is the middle value of the dataset, is also the same as the mean in a normal distribution.

Additionally, the mode, which is the value that occurs most frequently in the dataset, is also equal to the mean and median in a normal distribution.

Therefore, for a normal distribution, the mean, median, and mode are all the same value, and they all represent the center or midpoint of the distribution.




## Q13. How is covariance different from correlation?

Ans:

> Covariance and correlation are both statistical measures that describe the relat ionship between two variables, but they differ in several important ways.

**Range:**

> The range of covariance is unbounded, which means it can take on any value. In c ontrast, the range of correlation is between -1 and +1.

**Standardization:**

> Covariance is not standardized, which means it is affected by the scale of the v ariables being measured. Correlation, on the other hand, is standardized, which means it is not affected by the scale of the variables.

**Interpretation:**

Covariance measures the direction and strength of the linear relationship betwee n two variables. A positive covariance indicates a positive relationship between the variables, while a negative covariance indicates a negative relationship. Ho wever, it does not provide any information about the strength of the relationshi p. Correlation, on the other hand, provides information about both the direction and strength of the linear relationship between two variables.

**Units:**

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

Ans:

Outliers are data points that are significantly different from other observation s in a dataset. Outliers can have a significant impact on measures of central te ndency and dispersion.

**Measures of central tendency**

Outliers can  affect measure of central tendency, such as the mean because they are calculated using all the data points in the dataset. If there are outliers p resent, they can significantly pull the mean in one direction or another, causin g it to be an inaccurate representation of the center of the dataset.

**measures of dispersion**

Outliers can also affect measures of dispersion, such as the range, variance, an d standard deviation. Outliers can increase the range of the dataset, which make s it difficult to interpret the spread of the data. Outliers can also increase t he variance and standard deviation, which makes the spread of the data appear la rger than it actually is.

In [12]:

```
data = [6,1,2,3,9,8,7,6,2,7,4,5,8,6,1,5,7,8,5,5,8,9,10]

# mean, mode , median without outliers

print(np.mean(data))
print(np.median(data))
print(stats.mode(data))

# varience and standard deviation

print(np.var(data))
print(np.std(data))
```

```
5.739130434782608
6.0
ModeResult(mode=array([5]), count=array([4]))
6.540642722117201
2.5574680295396073
```

In [13]:

```
data = [6,1,2,3,9,8,7,6,2,7,4,5,8,6,1,5,7,8,5,5,8,9,10,1000]

# mean, mode , median with  outliers 1000

print(np.mean(data))
print(np.median(data))
print(stats.mode(data))

print(np.var(data))
print(np.std(data))
```

```
47.166666666666664
6.0
ModeResult(mode=array([5]), count=array([4]))
39479.80555555556
198.69525800973602
```

In [ ]:

```

```