

# Intermediate

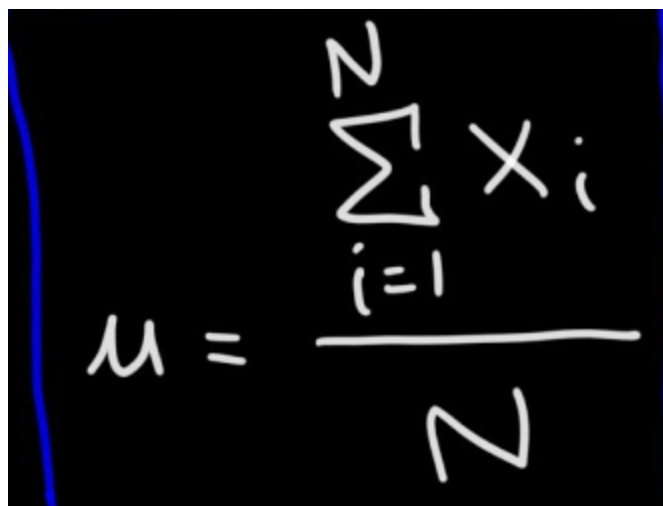
## Central Tendency

- Mean, median, mode
- Refers to the measure used to determine the centre of the distribution of the data.

## Arithmetic Mean

- Mean = Avg
- Population (N), Sample (n)
- Population mean =  $\mu$
- Sample Mean =  $\bar{x}$

**Population mean:**

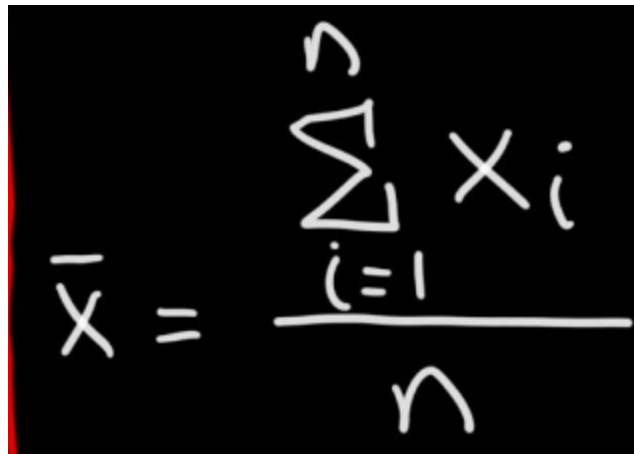
A handwritten formula for the population mean,  $\mu$ , is shown on a black background with blue borders. The formula is  $\mu = \frac{\sum_{i=1}^N x_i}{N}$ . The summation symbol  $\sum$  is positioned between the upper index  $N$  and the lower index  $i=1$ . The variable  $x_i$  is to the right of the summation, and  $N$  is below the horizontal line of the fraction.

where:

- $\Sigma$  denotes the sum of the values,
- $x_i$  represents each individual value in the set,
- $N$  is the total number of values.

**Use:** It is used when you have access to the entire population and want to calculate the true average.

### Sample Mean:


$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

**Use:** It is used when you only have access to a part of the population and want to estimate the population mean.

### Weighted Mean

- Each data point contributes a different amount to the overall mean based on its **weight**. This is useful when some values are more important than others.

## Formula for Weighted Mean:

$$\text{Weighted Mean} = \frac{\sum (x_i \cdot w_i)}{\sum w_i}$$

Where:

- $x_i$  = the value of each data point.
- $w_i$  = the weight corresponding to each data point.
- $\sum (x_i \cdot w_i)$  = sum of the product of each value and its corresponding weight.
- $\sum w_i$  = sum of all the weights.

## Trimmed Mean

- specified percentage of the highest and lowest values are removed from the dataset before calculating the mean.
- Mean without outliers
- The percentage of values removed is called the trimming percentage.

## Median

- Represents the middle value of a dataset when it is ordered from **least to greatest**.
- It is particularly useful for **understanding the center of a dataset**, especially when there are **outliers** that can skew the mean.

## Example

### Odd Number of Observations:

Consider the dataset: 3, 1, 4, 2, 5.

1. Order the data (sort): 1, 2, **3**, 4, 5

2. Since there are 5 numbers (odd), the median is the middle number:  
Median=3

### Even Number of Observations:

Consider the dataset: 7, 1, 3, 2.

1. Order the data: 1, 2, 3, 7
2. Since there are 4 numbers (even), the median is the average of the two middle numbers:  
Median=(2+3)/2=2.5  
$$\text{Median} = \frac{2 + 3}{2} = 2.5$$

## Mode

- Represents the value or values that occur most frequently in a dataset.
- useful for understanding the most common items in your data.

### How to Calculate the Mode?

1. **Identify Frequency:** Count how many times each value appears in the dataset.
2. **Determine the Most Frequent Value:**
  - If one value appears most frequently, that value is the mode.
  - If two or more values have the same highest frequency, the dataset is multimodal (having multiple modes).
  - If no number repeats, the dataset has no mode.

### Single Mode:

Consider the dataset: 4, 1,  
2, 2, 3, 5

Here, the mode is  
**2**, as it appears most frequently.

### **Multimodal:**

Consider the dataset: 1,  
**2, 2, 3, 3, 4.**

This dataset is bimodal, with modes  
**2** and **3**.

### **No Mode:**

Consider the dataset: 1, 2, 3, 4.

- All values appear only once, so there is no mode.

## Measure of Dispersion

1. Variance
2. Standard Deviation
3. Range

- It tells you the Spread of your data.

**Dispersion= Spread**

## Range

- The range is the difference between the maximum and minimum values in the dataset.
- Can be affected by outliers.
  - Therefore, it's not generally used to calculate spread.

## Variance (Imp)

- concept of measure of dispersion
- It tells us how the data is dispersed in the given data value.
- **A higher variance indicates greater variability means the data is spreaded, while a lower variance suggests the data points are closer to the mean.**

### 2 Types:

- i. Population Variance ( $\sigma^2$ )
- ii. Sample Variance ( $s^2$ )

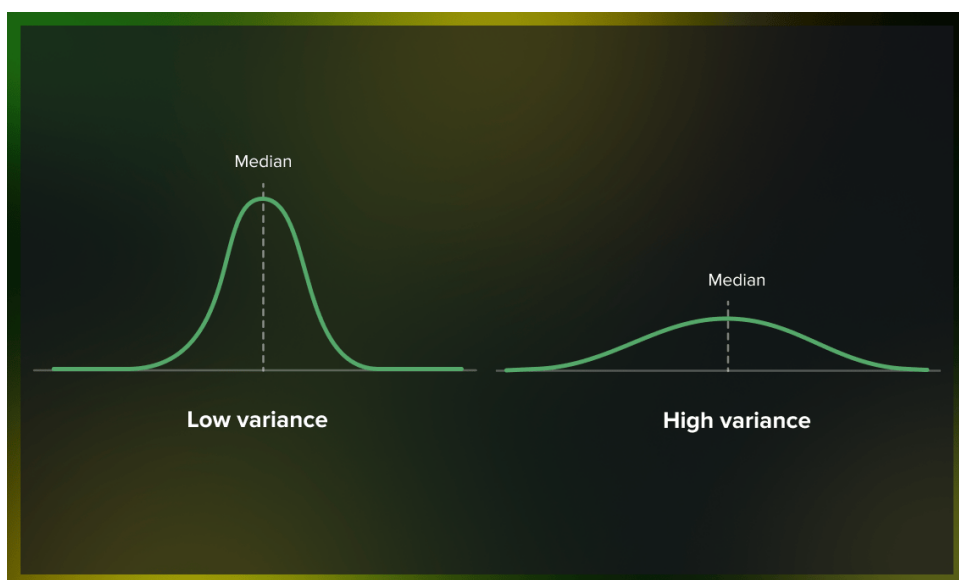
Population	Sample
$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$ <p><math>\mu</math> - Population Average <math>x_i</math> - Individual Population Value <math>n</math> - Total Number of Population <math>\sigma^2</math> - Variance of Population</p>	$S^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$ <p><math>\bar{x}</math> - Sample Average <math>x_i</math> - Individual Population Value <math>n</math> - Total Number of Sample <math>S^2</math> - Variance of Sample</p>

How do we calculate it?

18

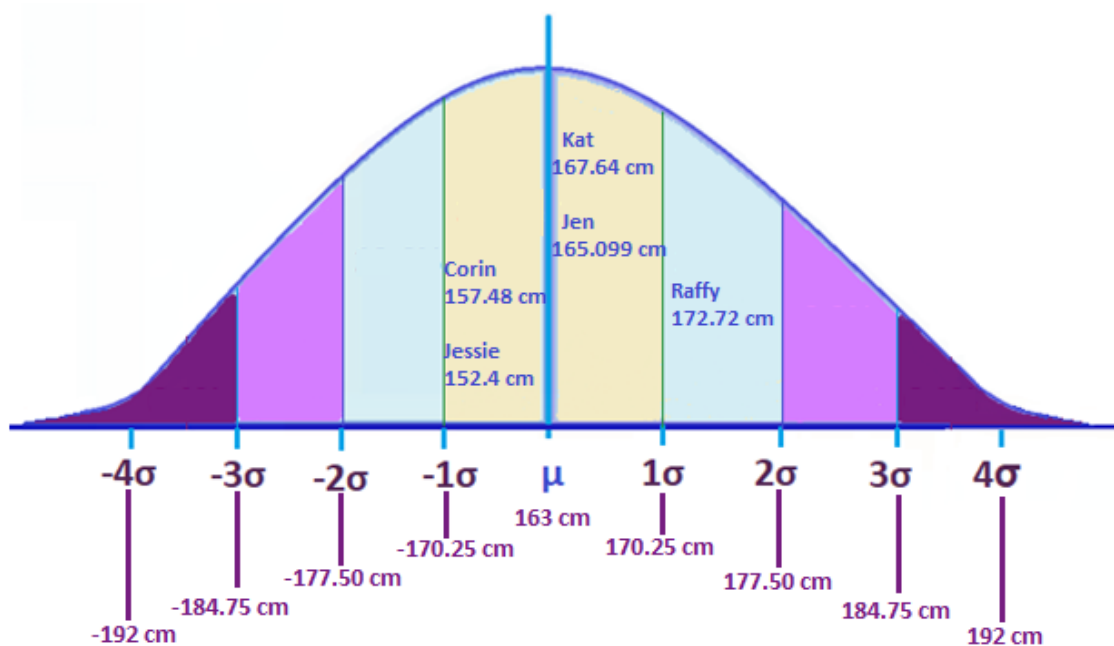
$x$	$\mu$	$x - \mu$	$(x - \mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	+0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71
<u>5</u>	<u>2.83</u>		<u>10.84</u>
$\mu = 2.83$			<u><u>10.84</u></u>

- We will divide 10.84 with 6 to get the **Variance**.
- We get 1.81
- High variance= data spread high



## Standard Deviation

- SD= Square root of variance



- Here, mean is 163.
- Variance is 52.64
- SD is 7.25 (Sq. root of Variance)
- So, if we go 7.25 to left or right from the mean, it is 1 SD.



Population	Sample
$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$ <p> <math>\mu</math> - Population Average  <math>x_i</math> - Individual Population Value  <math>n</math> - Total Number of Population </p>	$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$ <p> <math>\bar{x}</math> - Sample Average  <math>x_i</math> - Individual Population Value  <math>n</math> - Total Number of Sample </p>

## Why use SD when there's Variance?

- The unit of SD is same as of the data.

## Coefficient of Variation

- standard deviation as a percentage of the mean
- $CV = (\text{standard deviation} / \text{mean}) \times 100\%$**

$$CV = \frac{\sigma}{\mu} \times 100$$

Where:

- $\sigma$  = standard deviation of the dataset.
- $\mu$  = mean of the dataset.
- The result is expressed as a percentage.

- Unit-independent:** Since CV is a ratio of the standard deviation to the mean, it is **unitless**

- A **lower CV** indicates that the data is more consistent (less spread relative to the mean), while a **higher CV** indicates more variability.

## Percentiles & Quartiles

### Quantile

- Quantiles are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.
- They are used to understand the distribution and spread of data.
- **Common Types:**
  - **Quartiles:** Divide data into four equal parts (25th, 50th, 75th percentiles).
  - **Deciles:** Divide data into ten equal parts (10th, 20th, ..., 90th percentiles).
  - **Percentiles:** Divide data into one hundred equal parts (1st, 2nd, ..., 99th percentiles).
  - **Quintiles:** Divides the data into 5 equal parts
  - **Deciles:** Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile), ..., D9 (90th percentile).

### Things to remember while calculating these measures:

1. Data should be sorted from low to high
2. You are basically finding the location of an observation
3. They are not actual values in the data
4. All other tiles can be easily derived from Percentiles

# Percentiles

- Used to find outliers
- Def- Percentile is a value below which is a certain percentage of observations lie.
  - 25 Percentile = 25% of entire distribution is less than that particular value.
  - For example, if you are in the 70th percentile for height, it means you are taller than 70% of the people in the dataset.

***Formula to calculate the percentile rank of a specific value in a dataset:***

$$P = n/N * 100\%$$

Where P =percentile,

n= is the number of data points below the data point of interest

N= number of data points in the data set.

Ex.

**Scores:** 56, 67, 70, 72, 75, 78, 80, 85, 90, 95

Suppose you want to find the percentile rank of the score **78**.

**Sort in ascending order.**

1. **Identify the Value:** The value of interest is **78**.

2. **Count the Number of Values Below It (n):**

- The scores below 78 are: 56, 67, 70, 72, 75.
- So, **n=5**

**Determine the Total Number of Values (N)** :There are 10 scores in total, so  
**N=10**

**Calculate the Percentile Rank (P):**

$$P = \frac{n}{N} \times 100\% = \frac{5}{10} \times 100\% = 50\%$$

**For multiple occurrences:**

$$\text{Percentile Rank} = \frac{\text{Number of values below } X + 0.5 \cdot \text{Number of values equal to } X}{\text{Total number of values}} \cdot 100$$

Suppose you have a dataset: [10, 20, 30, 40, 50], and you want the percentile rank of  $X = 30$ .

1. Number of values below 30: 2 (10 and 20).
2. Number of values equal to 30: 1 (30).
3. Total number of values: 5.

$$\text{Percentile Rank} = \frac{2 + 0.5 \cdot 1}{5} \cdot 100 = \frac{2.5}{5} \cdot 100 = 50$$

So, the percentile rank of 30 is 50%.

- What value exists at percentile rank x?

## Using Python

```
from scipy import stats
```

```
stats.percentileofscore(data, value, kind='rank')
```

## We can do the reverse.

Ex.

**Scores:** 56, 67, 70, 72, 75, 78, 80, 85, 90, 95

**Sort in ascending order.**

**25th Percentile (P25):** This is the score below which 25% of the data falls.

To find P25, we can use the formula

To find P25, we can use the formula:

$$P_k = \frac{k}{100} \times (N + 1)$$

where  $P_k$  is the k-th percentile,  $k$  is the desired percentile, and  $N$  is the number of observations.

For P25:

$$P_{25} = \frac{25}{100} \times (10 + 1) = 2.75$$

Since this is not a whole number, we round up to 3. The 3rd score in the ordered list is 70.

Meaning- **P25 (25th Percentile):** 70 (25% of students scored below this)

## Calculate percentile in Python:

**Using NumPy:**

```
import numpy as np
np.percentile(data, 90) py
```

**Using Pandas:**

```
import pandas as pd
pd.Series(data).quantile(0.9) # 90th percentile
```

## FIVE NUMBER SUMMARY

1. Minimum
2. First Quartile (Q1)
3. Median
4. Third Quartile (Q3)
5. Maximum

- **We remove an outlier with the help of this.**
- Visually represented using **box plot**.

### Removing an Outlier

- We have to first define a **lower fence** & **higher fence**.
- In short, we need a demarcation line.

**Lower fence=  $Q1 - 1.5 (IQR)$**

**Upper fence=  $Q3 + 1.5 (IQR)$**

***IQR= Interquartile range=  $Q3 - Q1$***

Q3= 75 Percentile

Q1= 25 Percentile

Ex.

### Example 1:

Consider the following dataset:

2, 4, 6, 8, 10, 12, 14, 16, 18, 20

#### 1. Find Q1:

- There are 10 data points.
- Q1 is the value at the  $(n+1)/4$  position =  $(10+1)/4 = 2.75$ th position.
- This means Q1 lies between the 2nd and 3rd data points.
- $Q1 = (4 + 6) / 2 = 5$

#### 2. Find Q3:

- Q3 is the value at the  $3(n+1)/4$  position =  $3(10+1)/4 = 8.25$ th position.
- Q3 lies between the 8th and 9th data points.
- $Q3 = (16 + 18) / 2 = 17$

#### 3. Calculate IQR:

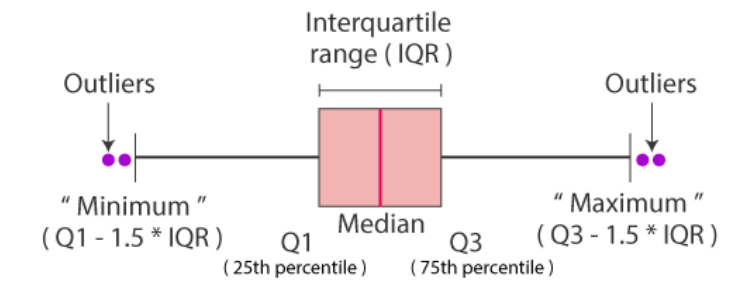
- $IQR = Q3 - Q1 = 17 - 5 = 12$

Lower fence  $Q1 - 1.5(IQR) = -13.0$

Upper fence  $Q3 + 1.5(IQR) = 35.0$

## Box Plot

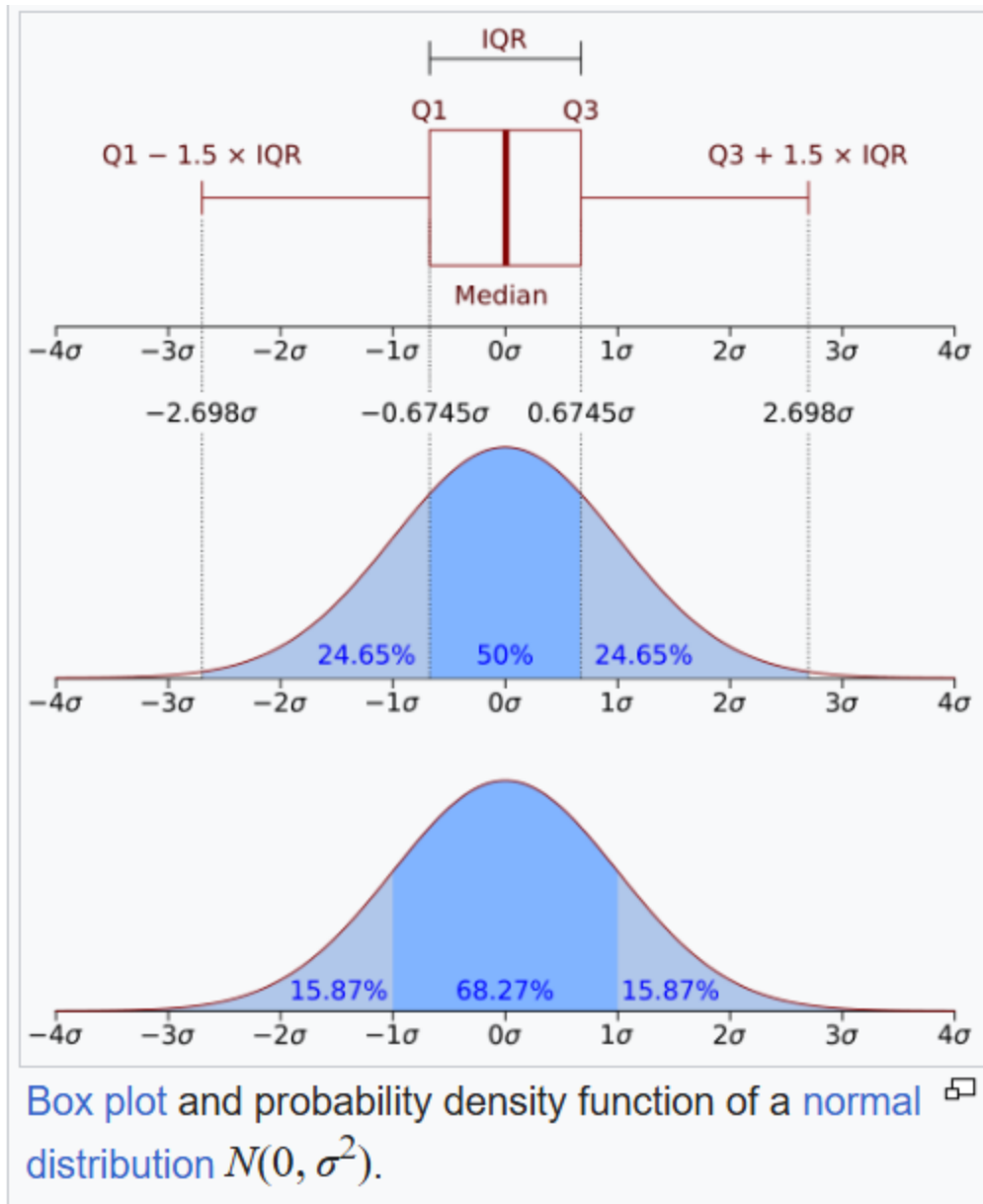
Refer →



**Different parts of boxplot**

© Byjus.com



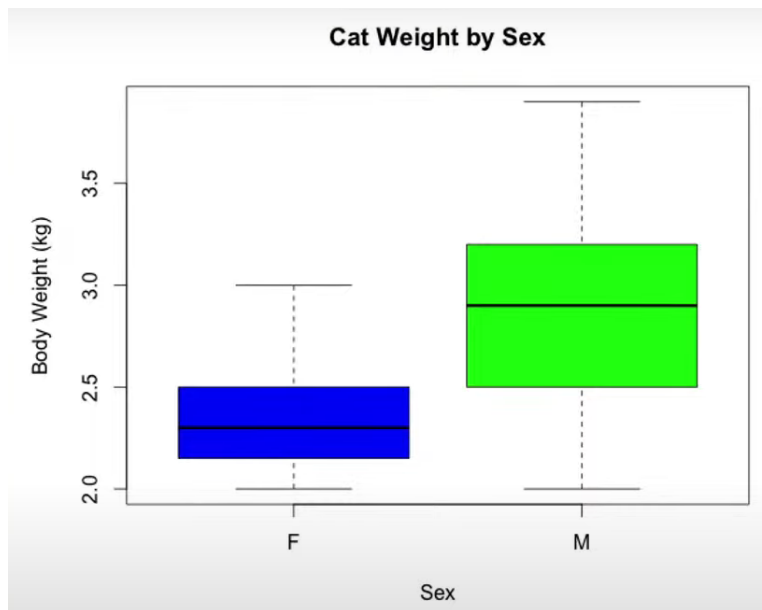


- Box plots are used to determine outliers.

### Benefits of a Boxplot:

- Easy way to see the distribution of data
- Tells about skewness of data

- Can identify outliers
- Compare 2 categories of data



## Variance

**The Sample Variance Formula**

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

$x_i$  are the individual data values  
 $\bar{x}$  is the sample mean  
 $n$  is the number of data values

© Maths at Home [www.mathsat-home.com](http://www.mathsat-home.com)

**n-1** → Bessel's correction, degree of freedom.

### Why Sample Variance is Divided by n-1?

- After trying n-1, n-2, n-3, etc, it was found that when we use n-1, the population variance and sample variance are approximately equal.

- This correction, known as **Bessel's correction**.
- It is necessary because when we calculate the sample mean, it uses one degree of freedom, meaning we lose one piece of independent information from our sample.
- Using  $n-1$  provides an **Unbiased estimation**.

## Mean Absolute Deviation

- It doesn't use square
- It used MOD i.e.  $|x_i - \bar{x}|$  will always give a positive value

### Formula for Mean Absolute Deviation (MAD):

$$\text{MAD} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

Where:

- $n$  = number of data points in the dataset.
  - $x_i$  = individual data points.
  - $\bar{x}$  = mean of the dataset.
  - $|x_i - \bar{x}|$  = absolute difference between each data point and the mean.
- It is **less sensitive to extreme outliers** because it uses absolute values rather than squaring differences (as in variance).
  - This makes it a good measure of spread when you don't want the outliers to overly influence the result.

## Graphs

## Univariate Analysis

- Single column

### Categorical Columns- Frequency Distribution Table & Cumulative Frequency

- A **frequency distribution table** is a table that summarizes the number of times (or frequency) that each value occurs in a dataset.

Type of Vacation	Frequency
Beach →	60
City	40
Adventure	30
Nature	35
Cruise	20
Other	15

- You can plot- Histogram, Bar chart

**Relative frequency** is the proportion or percentage of a category in a dataset or sample.

- It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample.

Type of Vacation	Frequency	Relative Frequency
Beach	60	0.3
City	40	0.2
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
Other	15	0.075

- You can plot - Pie chart

### Example:

Suppose you have a frequency distribution of fruits like this:

Fruit	Frequency
Apple	40
Banana	30
Orange	20
Grapes	10

The total number of fruits is:

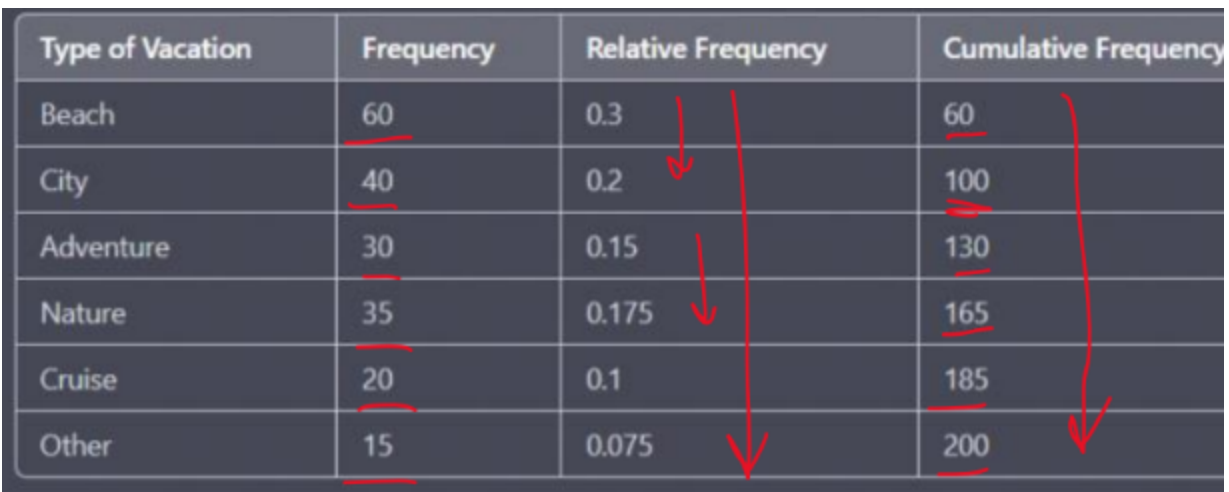
$$40 + 30 + 20 + 10 = 100$$

Now, we calculate the **relative frequency** for each fruit:

- **Apple:**  $\frac{40}{100} = 0.4$  or **40%**
- **Banana:**  $\frac{30}{100} = 0.3$  or **30%**
- **Orange:**  $\frac{20}{100} = 0.2$  or **20%**
- **Grapes:**  $\frac{10}{100} = 0.1$  or **10%**

**Cumulative frequency** is the running total of frequencies of a variable or category in a dataset or sample. It is calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample.

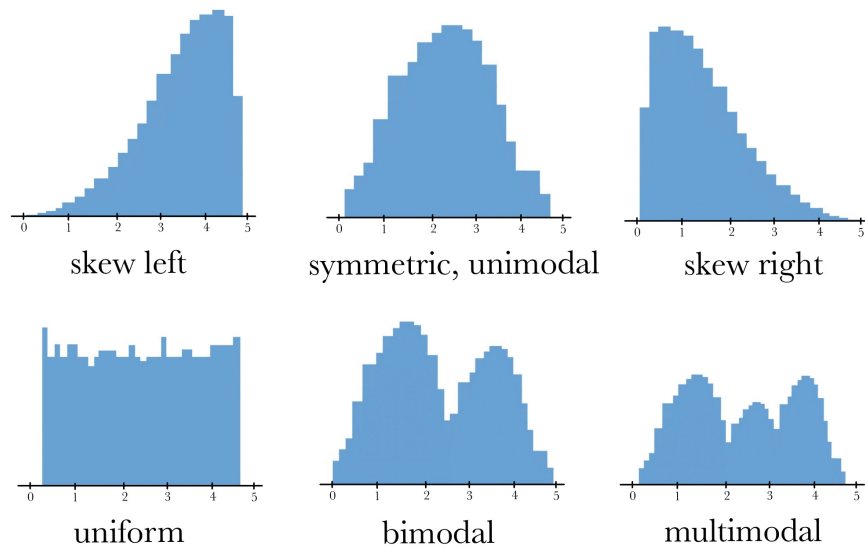
Type of Vacation	Frequency	Relative Frequency	Cumulative Frequency
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200



- You can plot a line chart for this.

## Numerical Columns- Frequency Distribution Table & Histogram

- You can still create a frequency distribution table
- Categories are called bins/bucket.
- You plot → **Histogram**



## Graphs for Bivariate Analysis

### 1. Categorical - Categorical

- Contingency table, also known as a cross-tabulation or **crosstab**
- It displays the frequencies or relative frequencies of the observed values of the two variables, organized into rows and columns.

Count of Survived	Column Label				
Row Labels	1	2	3	(blank)	Grand Total
0	80	97	372		549
1	136	87	119		342
(blank)					
Grand Total	216	184	491		891

### 2. Numerical - Numerical

- Scatter Plot

### 3. Categorical - Numerical

- Bar chart

<b>Univariate</b>		
<b>Categorical</b>	Frequency Distribution Table,Cumulative Frequency	Histogram, Bar chart
	Relative frequency	Pie chart
	Cumulative frequency	Line Chart
<b>Numerical</b>	Frequency Distribution Table	Histogram
<b>Bivariate</b>		
<b>Cat-Cat</b>	Contingency Table/Crosstab	Stacked bar chart
<b>Num-Num</b>	-	Scatter Plot
<b>Cat-Num</b>	Contingency Table/Crosstab	Bar chart