Statistics Fundamentals

Agenda

In this lesson, we will cover the following concepts with the help of a business use case:

- Statistics and its importance in Data Science:
  - Types of Statistics
- Data Categorization:
  - Types of Data
- Levels of Measurement
- Measures of Dispersion
- Random Variables
- Sets with its Operations
- Measures of Shape(Skewness)

Importance of Statistics for Data Science

Most of the major developments in statistics occurred in the last century. Probability theory, which is the foundation of statistics, was developed between the 17th and 19th centuries by Thomas Bayes, Pierre Simon Laplace, and Carl Gauss. While probability theory is theoretical, statistics is an applied branch of science whose primary aim is to analyze data.

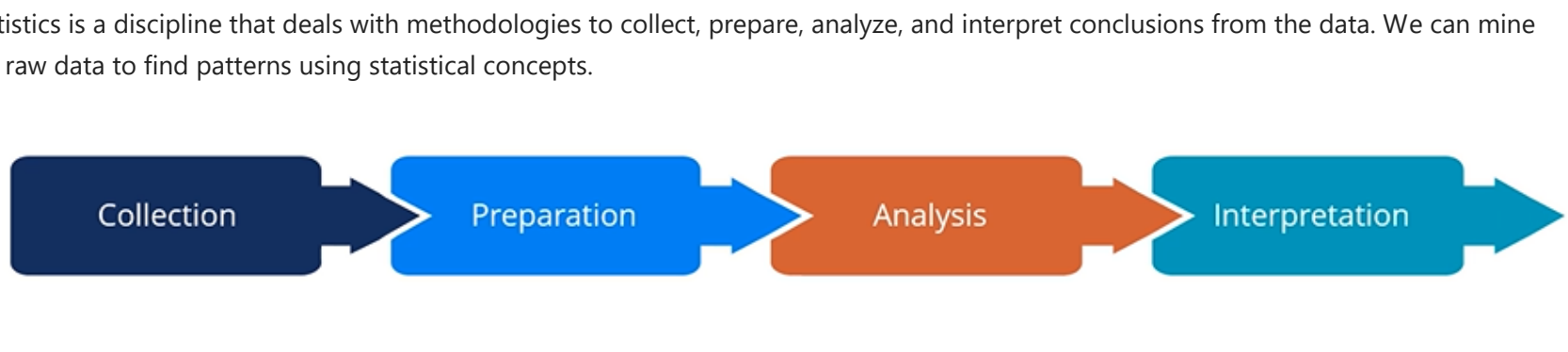
In this Lesson, we will explore the different concepts of statistics and their importance in data science.

The first step of any data science project is to explore data. Exploratory Data Analysis or EDA is a relatively new area of statistics. Classical statistics focuses exclusively on Inference, which is a set of procedures to conclude about large populations by studying small samples. With the explosion of computing power and expressive data analysis software, EDA has evolved way beyond its original scope.

The rapid development of new technology, access to bigger data, and the greater use of quantitative analysis in a variety of disciplines have driven this growth.

What Is Statistics?

Statistics is a discipline that deals with methodologies to collect, prepare, analyze, and interpret conclusions from the data. We can mine the raw data to find patterns using statistical concepts.



We combine this with the domain expertise to interpret these patterns and use the findings for decision-making in real-world situations.The end objective is to generate value for an organization

Data science is a combination of Computer Science and Statistics. A thorough understanding of statistics is vital for developing a strong intuition for machine learning algorithms.

Otherwise, the algorithms resemble black boxes that just output a certain level of accuracy.The knowledge of core concepts of statistics helps in understanding the origin of this accuracy and in feature engineering. This knowledge also helps in building ensemble methods where different machine learning algorithms can be combined to get the highest possible level of accuracy.

Common Terms Used in Statistics

1. **Population and Sample:** Population is the complete data pool from where a sample is drawn for further analysis.The sample is a subset of the population.

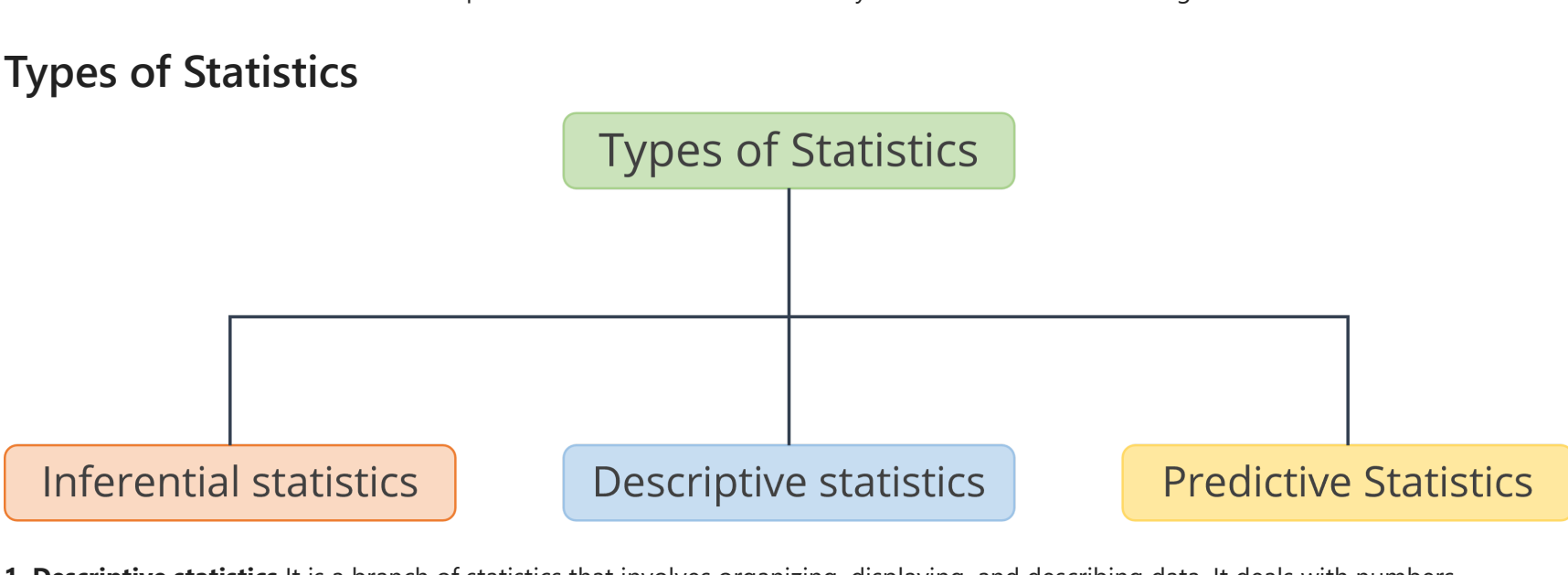
1. **Measurement and Sample Data:** A measurement is a number or attribute calculated for each member of the population or sample. The measurements of the sample members are collectively called sample data.

1. **Parameter:** A parameter is a characteristic of the population that we want to estimate or test, such as the population mean.

1. **Variable:** A variable is something that can take on different values in the data set.

1. **Distribution:** Distribution is the sample data that takes values from only a Measurement certain range.

Types of Statistics



1. Descriptive statistics It is a branch of statistics that involves organizing, displaying, and describing data. It deals with numbers (numerical facts, figures, or information) to describe any phenomena which can be referred to as descriptive statistics.

2. Inferential statistics It is a branch of statistics that involves drawing conclusions about a population based on the information obtained from a sample taken from that population.

Inferential statistics is used to make inferences from data whereas descriptive statistics simply describes what is happening with data.

Example:

Estimating the automobiles manufactured in a month, the entire output is considered as the population. The automobiles inspected for quality characteristics like mileage per gallon of gasoline constitute a sample.The average mileage of all cars is a parameter while the average life of the sample inspected is a statistic.

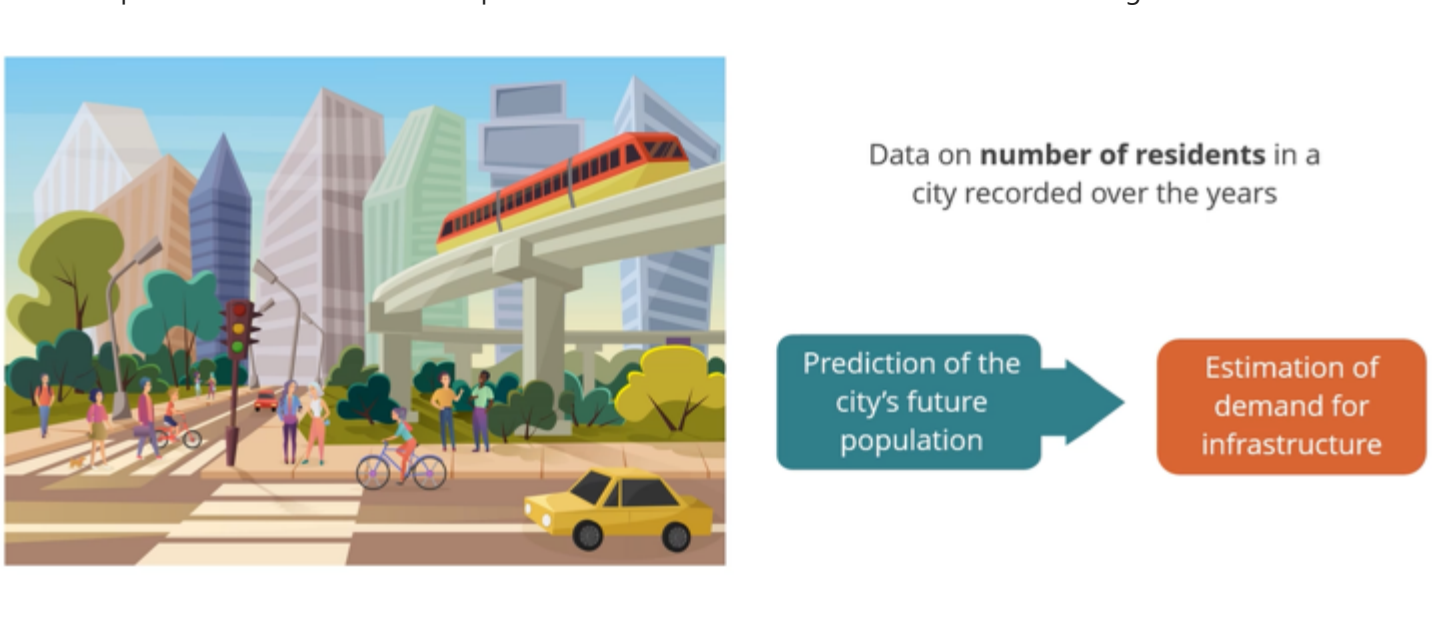
If we deal with descriptive statistics here, it will include the selection of a sample, the presentation of sample data as diagrams or tables, and the computation of the value of a statistic.

Inferential Statistics is used to generalize such as: "Is the population average at least 23 miles per gallon from the sample being studied?"

3. Predictive statistics It is defined as the science of extracting information from data and using them to predict trends, behavior patterns, or relationships between characteristics.

Example 1:

The data on the number of residents in a city recorded over the years could be used to predict the city's future population. This could be useful to estimate the demand for infrastructure for the years to come. Sometimes data on two or more characteristics tend to have relationships. Such data can be used to predict the values of one characteristic while knowing the values of the other.

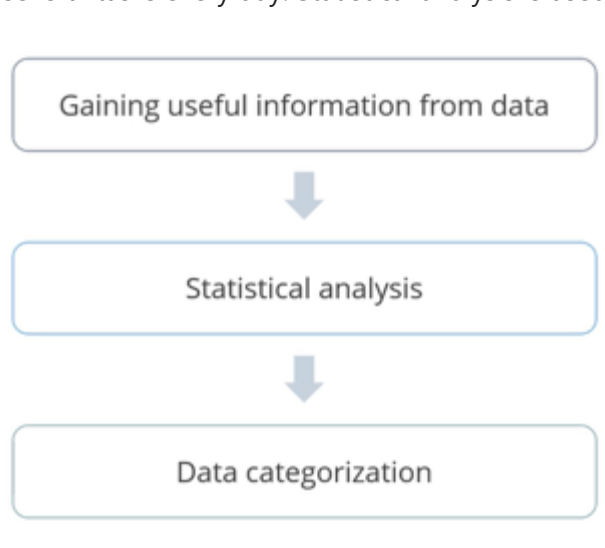


Example 2:

The demand for spares in the forthcoming years can be predicted by evaluating the sales of cars during the past few years.

Data Categorization

Data is a collection of facts and figures that can be analyzed, interpreted, and presented. We collect and use different types of data for several tasks every day. Statistical analysis is used to gain useful information from it and to categorize the data.



Example A survey is conducted to evaluate employee satisfaction. The responses can be studied by categorizing the respondents into two or more categories depending on the:

Level of seniority: Assistant managers, deputy managers, managers, senior managers, etc.

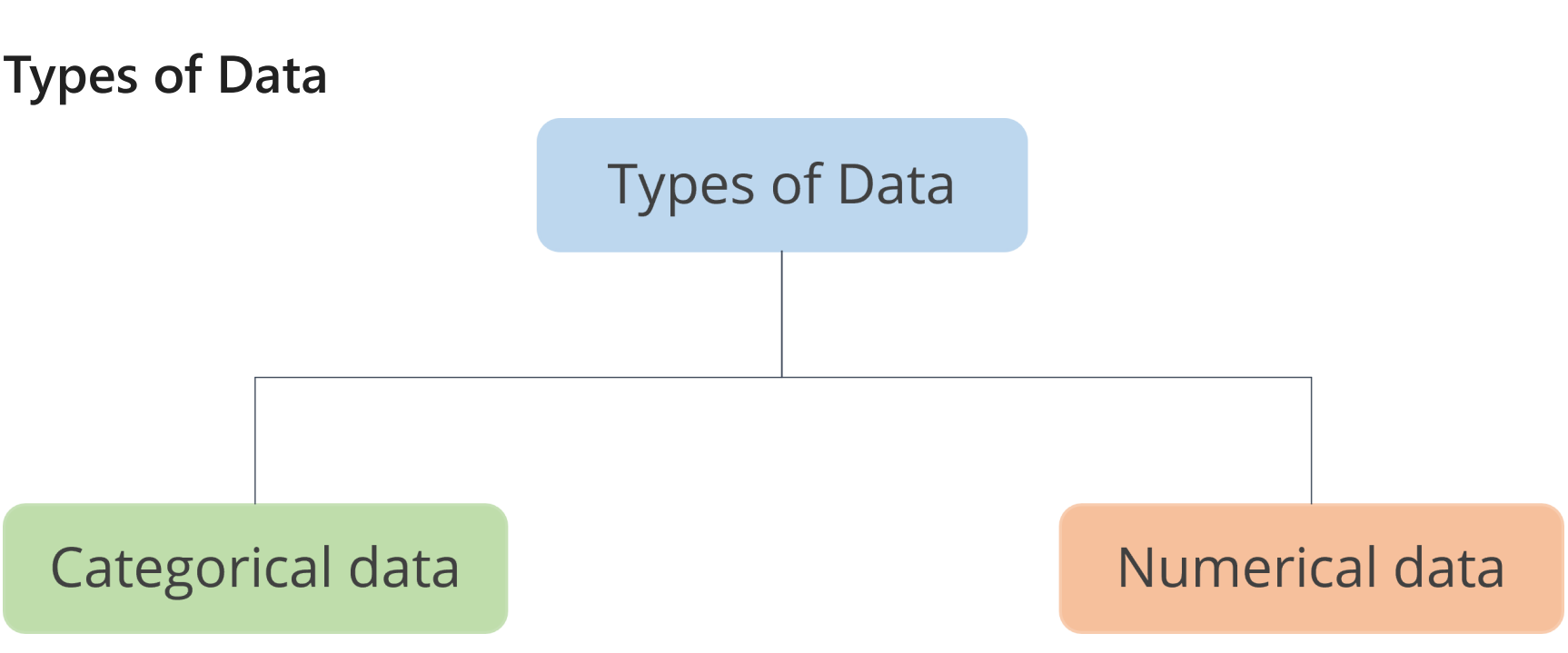
Functional affiliation: Sales and marketing, accounting, purchase, human resources, etc.

Work experience Number of years of work

Location: City-wise or branch-wise

Different categories Categorization is vital for identifying these differences

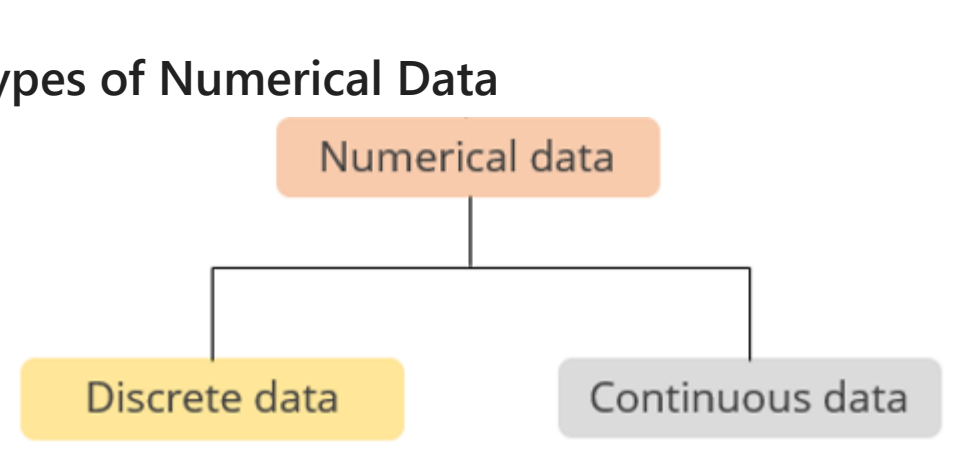
Types of Data



1.Categorical data: It represents characteristics such as a person's gender, marital status, or the types of movies they like. Categorical data can also take numerical values, such as "1" indicates male and "2" indicates female, but these numbers don't have mathematical meaning.

2. Numerical data It represents the data as a measurement, such as a person's height, weight, IQ, or blood pressure. Numerical data also represents the data that cannot be counted,such as the number of stocks owned by a person etc.

Types of Numerical Data



1. Discrete data: It represents the data that can be counted or the list of possible values.It is also called finite.

Examples:

- Number of children in a school
- Number of books in your library
- Number of cases a Lawyer has won

In all the above examples, the values can be 1, 2, or 3, but can never be 1.2, 4.6, 8.7, etc. Thus, the results are countable.

2. Continuous data: It represents the data in measurements and their values can not be counted but measured.

Examples:

- The height of a person can be described using intervals on the real number line.
- The amount of petrol purchased at the petrol pump for bikes with 20-liter tanks would be continuous data. It is represented by the interval [0, 20].

Qualitative and Quantitative data

1. Qualitative or attributed or categorical data

The data which states descriptively is referred to as qualitative data. Such data is not amenable to quantification.

The customer preferences for various brands of a product, types of faults in a tennis player such as hitting on the net and outside the baseline, etc., are examples of qualitative data.

2. Quantitative or variable or measurable data

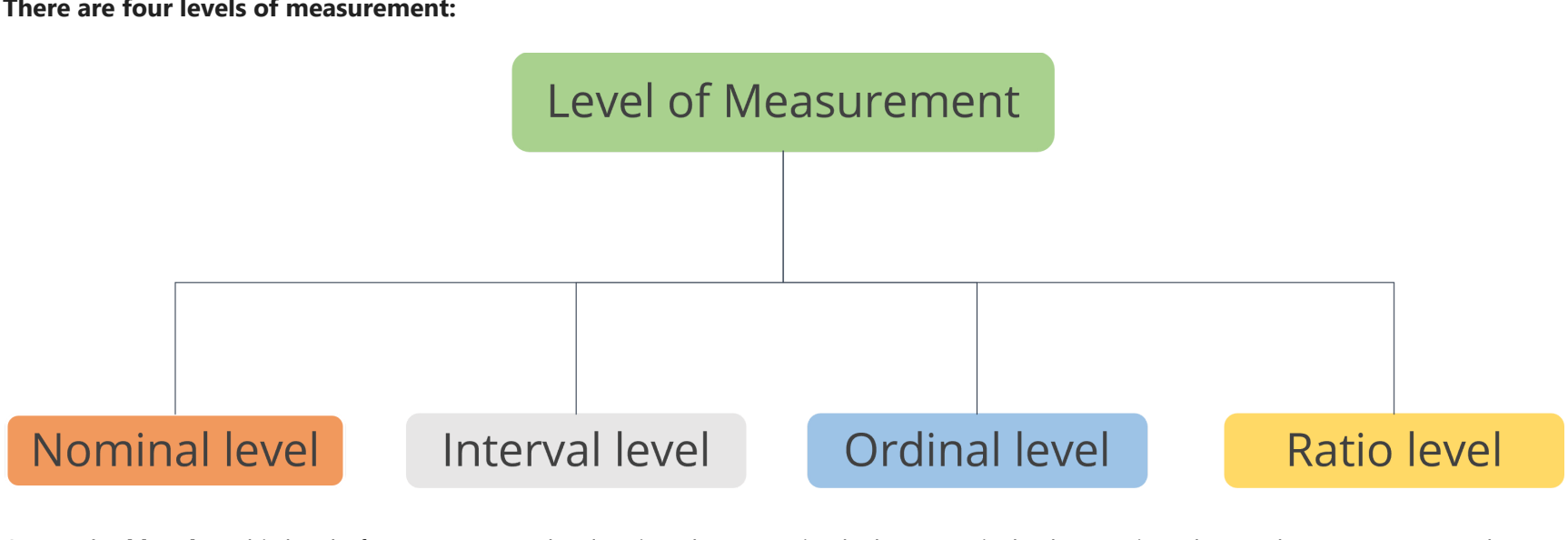
The data is specified numerically through a process of measurement or numerical count called quantitative data. Such data is amenable to arithmetic operations.

The diameter of a pole, number of seats occupied in a theater for a particular show, etc., constitute quantitative data.

Levels of Measurement

The levels of measurement demonstrate how variables are recorded. They are also called scales of measurement. It is crucial to understand the levels of measurement because they help us to determine the statistical analysis needed to address the problem.

There are four levels of measurement:



1. Nominal level: In this level of measurement, the data is only categorized. The numerical values assigned to each category cannot be construed as absolute values.

Example:

We are collecting data on cell phone brand preferences; we assign 1 to Apple, 2 to Samsung, and 3 to Nokia.These numbers are used to categorize the data.In some cases, letters or words may be used depending on the type of categorization.

2. Ordinal level: In this level of measurement, a range of values is assigned to a category, following which we rank these categories. However, we cannot distinguish between values in each category.

Example:

The students who achieve a score of 90 or above will receive grade A, those between 80 and 89 will receive grade B, and those between 70 and 79 will receive grade C. We are unable to differentiate between students of any grade category.

3. Interval level: At this level, the rank data is categorized. Additionally, the distances between each interval on the scale is equivalent along the scale from low to high intervals.However, there is no true zero point.

Example:

Celsius temperature measurement is an ideal example of this category.The difference between 100^oC and 200^oC is equivalent to the difference between 930^oC and 950^oC . But zero degree Celsius does not imply an absence of temperature.

4. Ratio level: At this level, in addition to the interval level, there exists an absolute zero value. An absolute zero means that the variable is absent.

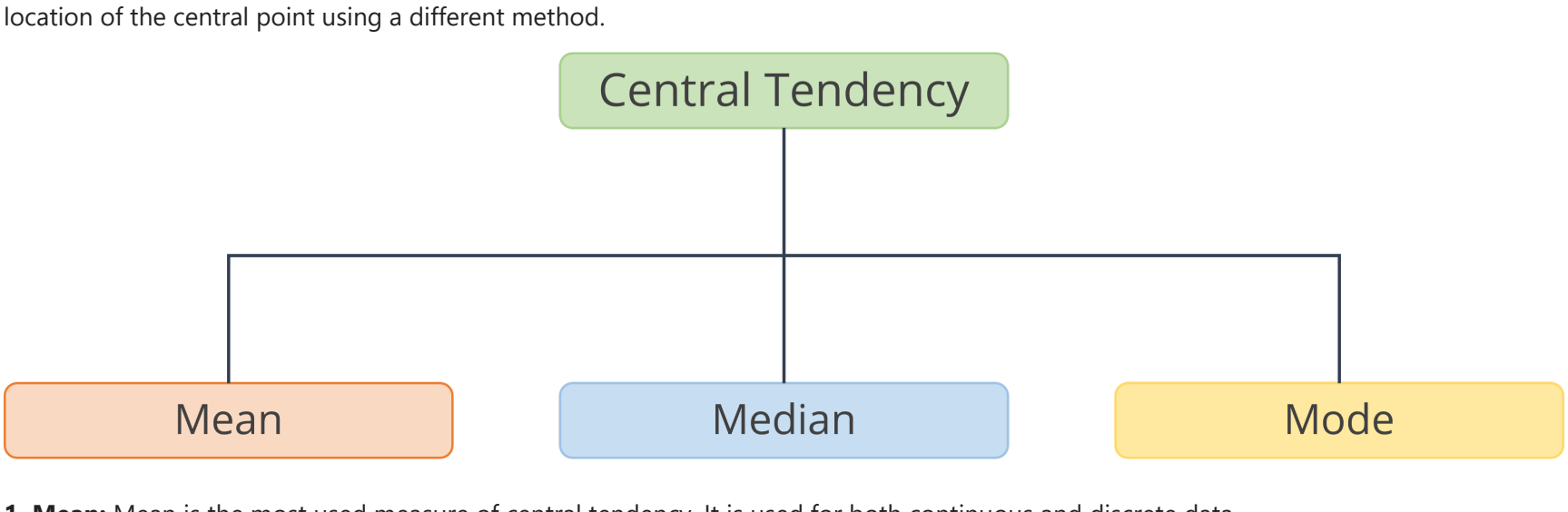
Example:

Extending the celsius temperature analogy, the Kelvin temperature scale has no negative temperatures. Zero Kelvin signifies an absolute lack of thermal energy.

Measures of Central Tendency

A measure of central tendency is a summary that describes the central position in a data set. These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution. It is also called a measure of central location.

In statistics, the three most common measures of central tendency are the mean, median,and mode. Each of these measures calculates the location of the central point using a different method.



1. Mean: Mean is the most used measure of central tendency. It is used for both continuous and discrete data.

To calculate the mean, add all the numbers in the data and divide the result by the number of data points.

Note: the Mean is sensitive to outliers and skewed data.

Suppose we have a set of n numbers given by x1,x2, ...,xn their mean is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

Example:

The mean of the 7 numbers 4, 89, 54, -7, -9, 27 and 5 is given by (4+89+54-7-9+27)/7 = 158/7 = 23.28

In [15]:

```
# Python code implementation of finding the mean
import statistics
statistics.mean([4,89,54,-7,-9,27,5])
```

Out[15]:

```
23.285714285714285
```

2. Median: The Median is the middle number obtained by arranging the data in ascending or descending order. Hence, the median is not so sensitive to outliers and skewed data.When the total number of data points is odd, we get the exact middle number.However, when the total number of data points is even, we get two middle numbers We then take the average of the two middle numbers to get the median.

Example 1:

Consider the dataset with odd numbers x = 5, 76, 98, 32, 1, -6, 34, 3, -65

Step 1: Arrange the numbers in ascending order i.e., -65, -6, 1, 3, 5, 32, 34, 76, 98

Step 2: The middle number is the fifth number as there are nine numbers in total. So, the median of the given data set is 5

Example 2:

Consider the dataset with even numbers x = 5, 76, 98, 32, 1, -6, 34, 3, -65,99

Step 1: Arrange the numbers in ascending order i.e., -65, -6, 1, 3, 5, 32, 34, 76, 98,99

Step 2: Identify the middle numbers 5 and 32

Step 3: To calculate the median, take the average of these two numbers. So, the median of the given data set is (5+32)/2 = 18.5 5.

In [17]:

```
# Python code implementation of finding the median.
x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
statistics.median(x)
```

Out[17]:

```
5.5
```

Mode: Mode is the most frequently occurring data point in the set. The advantage of mode is that it can be calculated for both numerical and categorical data. The disadvantage is that sometimes it may not reflect the centre of distribution well. A few data points at one end of the spectrum may fit the definition of mode but can be far from the true central point.

Suppose we want to infer the cell phone brand preference from categorical data. Let's say we collect data from 100 people.

Preferred brand	No. of people
—----------------------	—--------------
Nokia	12
Realme	32
Apple	10
Samsung	36
Oppo	4
Vivo	6

In the above data, the most frequently occurring preference is Samsung which indicates the value of mode.

In [18]:

```
# Python code implementation of finding the median.
x = ['Nokia','Samsung','Samsung','Apple','Oppo','Vivo']
statistics.mode(x)
```

Out[18]:

```
'Samsung'
```

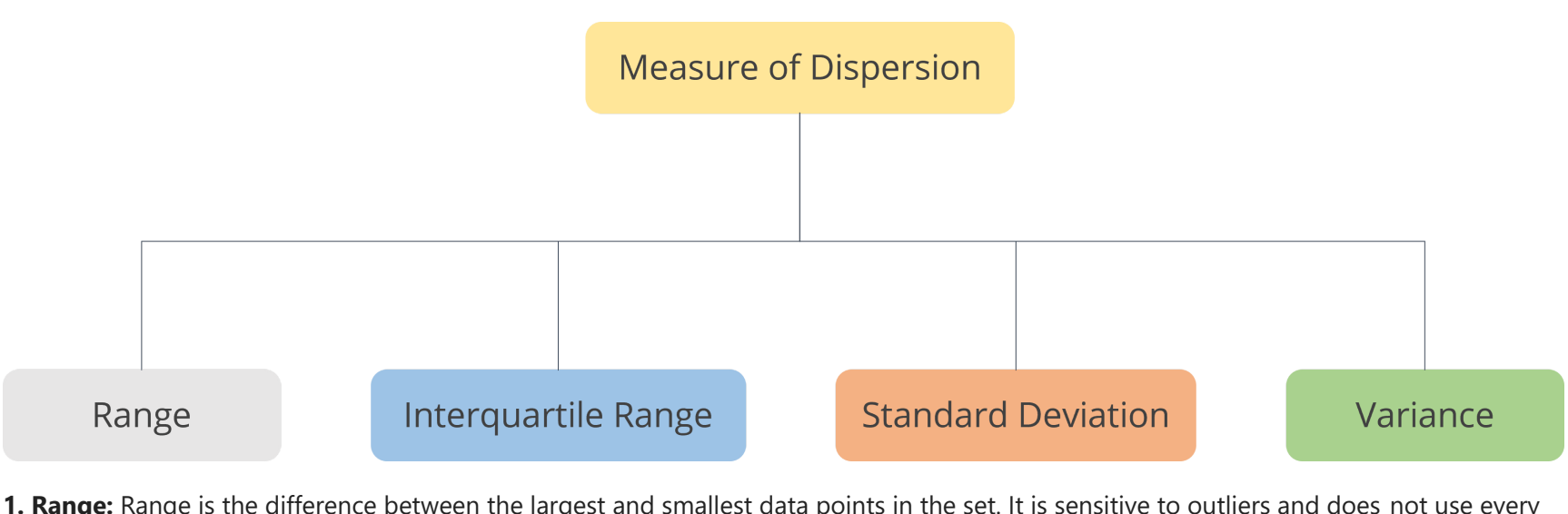
Measures of Dispersion

Measures of central tendency are insufficient to understand data distribution. While two different data sets can have the same mean or median, the variability of data around the mean can be high. Measures of dispersion give us an idea of data variability around the central point.

Uses:

1. It shows the variation in the data that provides information such as how well the average of the sample represents the entire data. Less variation gives close representation while with larger variation, the average may not closely represent all the values in the sample.
2. It enables us to compare two or more series with regard to their variations. It helps to determine consistency.
3. It also helps to control the causes behind the variations.

The commonly used measures of dispersion are:



1. Range: Range is the difference between the largest and smallest data points in the set. It is sensitive to outliers and does not use every data point in the set. It also provides maximum and minimum values in the set.

Since it is based on two extreme observations, it gets affected by fluctuations.Thus, the range is not a reliable measure of dispersion.

$$Range = X_{max} - X_{min}$$

2. Interquartile Range: Interquartile Range is the difference between the 25th and 75th percentile. It describes the middle 50% of the observations. If the middle 50% of the data points are spaced wide apart, their interquartile range will be large.

The interquartile range is useful even if the extreme values are not accurate as it is insensitive to them. However, it is not amenable to mathematical manipulation.

$$IQR = Q_3 - Q_1$$

3. Standard Deviation: Standard Deviation (SD) is the most popular measure of dispersion. It measures the spread of data around the mean. It is defined as the square root of the sum of squares of the deviation around the mean divided by the number of observations.

$$SD = [(\Sigma_i(x_i - \mu)2 \,/\, n]½$$

Note: The advantage of SD is that if the data points are from a normal distribution, 68% of the observations lie at a distance of 1 SD from the mean, 95% between 2 SDs, and 99.7% between 3 SDs. The other benefit is that it can detect skewness. However, it is not an appropriate measure of dispersion for skewed data.

In [ ]:

```
# Python code implementation for Standard Deviation


x = [1, 2, 3, 4, 5]
print(statistics.stdev(x))
```

```
1.5811388300841898
```

4. Variance Variance is defined as the average of the squared differences from the Mean.

Variance = [(Σi (xi – μ)2 ⁄ N]

In [ ]:

```
# Python code implementation for Standard Deviation


x = [1, 2, 3, 4, 5]
print(statistics.variance(x))
```

```
2.5
```

Standard Deviation and Variance for Population and Sample Data

Suppose there are N values present in a data set. To calculate the standard deviation and Variance, use:

1) Population divided by N when calculating the SD and Variance

2) Sample divided by N-1 when calculating the SD and Variance

The formula for calculating Standard Deviation and Variance changes while dealing with population and sample data.

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Standard deviation	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$	$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

How to Calculate the Variance and Standard Deviation?

Step 1: Compute the square of the difference between each value and the sample mean

Step 2: Add those values up

Step 3: Divide the sum by N-1 to obtain the Variance

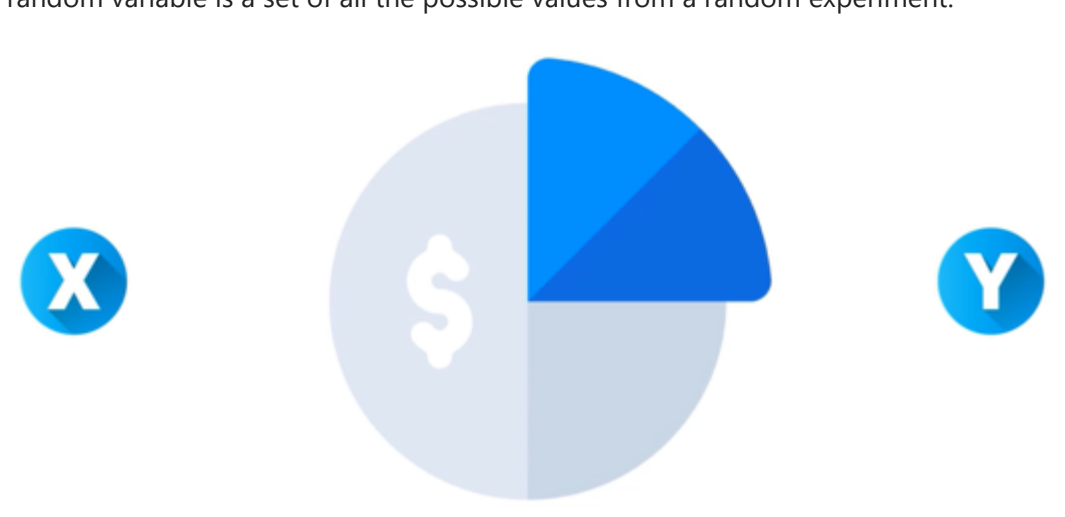
Step 4: Take the square root to obtain the Standard Deviation

Why Do We Divide Sample by (n-1) Instead of n?

In the sample variance formula, the denominator has n-1 instead of n, where n is the number of observations in the sample. The use of n-1 is the Bessel's correction method. We use this method as it corrects the bias in the estimation of the population variance.

Random Variables

A random variable is a set of all the possible values from a random experiment.



Every study has certain characteristics that are expressed numerically. Probabilities are assigned to various plausible values taken by the characteristic for further study. This is where the concepts of random variables and probability distribution come into picture.

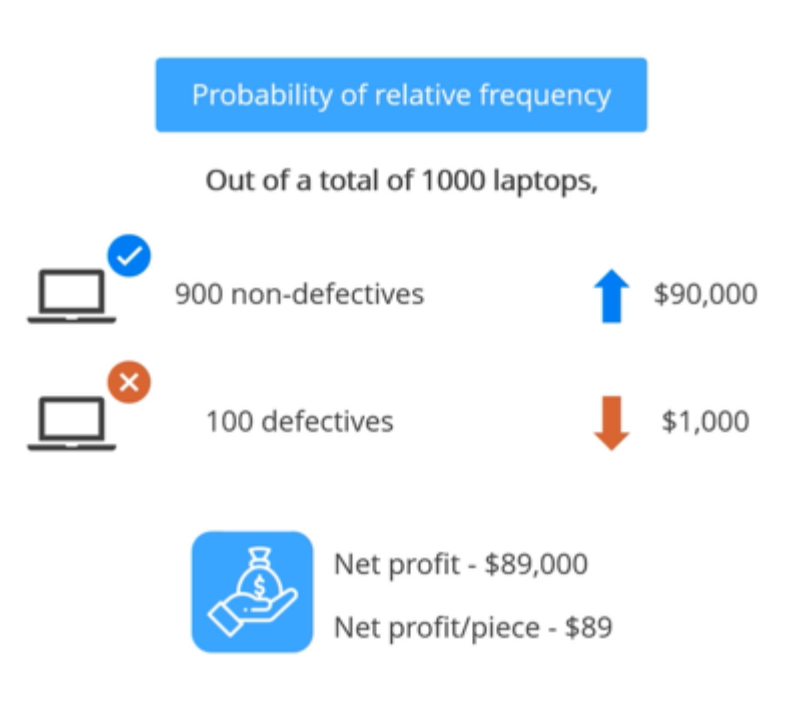
Example:

A laptop manufacturing unit produces non-defective and defective pieces with probabilities of 0.9 and 0.1 respectively. Every non-defective piece yields a profit of USD 100, while every defective piece leads to a loss of USD 10 due to wasted material, man, and machine hours.

	Non-defective pieces	Defective pieces
Probability	0.9	0.1
Value	+ $100	- $10

Let's assume that 1000 pieces are produced during a given period.

Based on the concept of probability as relative frequency:



Estimated number of non-defectives will be 900 and the estimated profit from the same will be USD 90,000. The estimated number of defective pieces will be 100 and the estimated loss from defectives will be USD 1000. Now, the net profit will be USD 89,000 and the net profit per piece will be USD 89.

We can compute the value of net profit/piece as:

{100* 900 + (-10) * 100)} /1000
= 100*0.9 + (-10)*0.1

Here, each value taken by the characteristic is multiplied by the corresponding probability and added on.

Random Variable Examples:

Sets

A set is a well-defined collection of objects. Every member of a set is called an element.Two sets are equal if they have the same elements.

A null set has no elements. If every element of set X is present in set Y, X is a subset of Y.

A set is usually denoted by capital letters like X, Y, A, B, etc.

The elements of a set are usually denoted by small letters like a, b, x, y, etc.

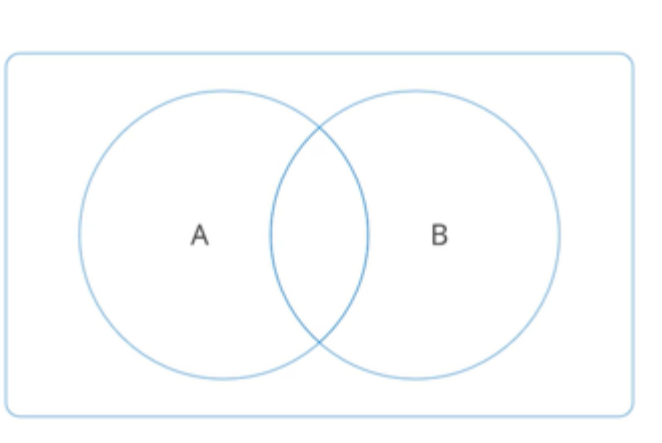
The set is represented by enclosing its elements within curly braces.

Example:

1. A set with elements 2,6,4,9,12 is denoted as shown below:

A = {2, 6, 4, 9, 12}

1. A null set is denoted by {}.



1. A set can also be represented by a rule, like a set of even numbers. In that case, it will be represented as shown below:

A = {2, 4, 6,.......}

Using Sets in Probability

There are many possible outcomes to a statistical experiment. The set of all outcomes is called sample space.

A particular outcome or a combination of outcomes is a subset of the sample space called an event.

Set Operations:

Consider a sample space S = {a, c, d, f, g, i, k} and two subsets of S, A = {a, c, d} and B = { d, f, i}

1. Union of sets: The union of two sets is the set with elements belonging to either one or both. It is denoted by U.

Example:

A U B = {a, c, d, f, i}

2. Intersection of two sets: The intersection of two sets is the set with elements common to both sets. It is denoted by ∩.

Example:

A ∩ B = {d}

3. Complement of a set: The complement of a set is the set whose elements are present in the sample space but not in the set. It is denoted by the superscript 'c' or c.

Example:

Ac = {f, g, i, k} and Bc = {a, c, g, k}

Measures of Shape (Skewness)

Measures of shape describe the distribution of data in a data set. For numerical data,the histogram is used to describe the distribution shape as low and high values on the x-axis.

The same is not possible for categorical data. Although a histogram gives the overall Shape, two quantitative measures of shape, skewness, and kurtosis are more precise.

What Is Skewness?

Skewness is the amount and direction of departure from horizontal symmetry. Any distribution is symmetric if the left and right of the center point look the same.

In many statistical inferences, we need the distribution to be normal or nearly normal.Skewness is vital as it helps us test for normality. In a normal distribution, skewness is 0.So, if the skewness is close to 0, it is nearly normal distribution.

Formula for Skewness

For a univariate data $X_1, X_2, \ldots. X_n$ skewness is given by:

$g1 = (\sum_{i=1}^{N} (X_i - \bar{x})^3 / N) / s^3$

Where $\bar{x}$ is the mean

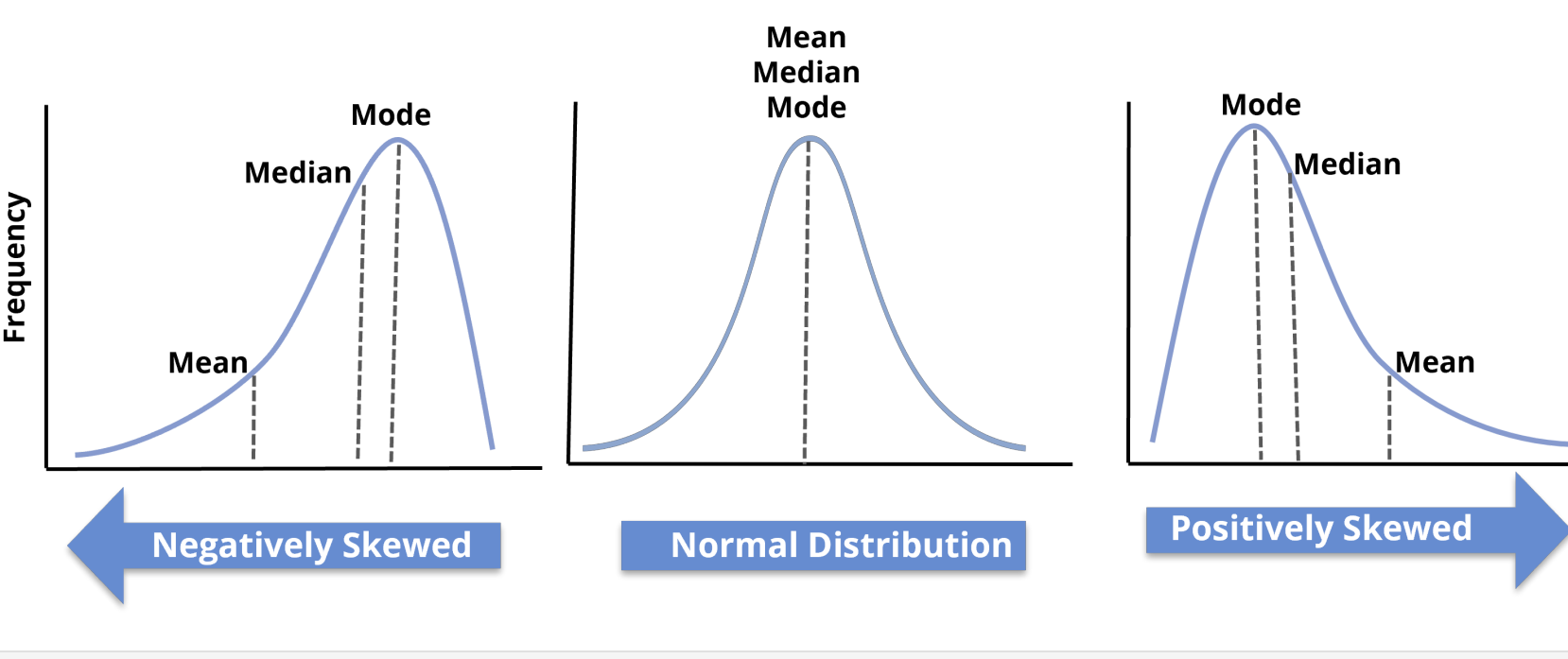
s is the standard deviation

N is the number of data points

In a normal distribution, the graph appears as a classical, symmetrical "bell-shaped curve."The mean, average, and mode or maximum point on the curve are equal if the tails on either side of the curve are exact mirror images of each other.

When a distribution is skewed to the left, the tail on the curve's lefthand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is called negative skewness.

When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is called positive skewness.



In [4]:
```python
# Importing numpy and statsmodels
import numpy as np
from statsmodels.stats.stattools import medcouple
from statsmodels.stats.stattools import robust_skewness

x = np.array([1, 2, 3, 4, 7, 8])
# Using statsmodels.robust_skewness() method
skewness = medcouple(x)

print(skewness)
```

```
0.2857142857142857
```

Measures of Shape (Kurtosis)

Kurtosis is the second measure of shape. It measures how heavy-tailed or light-tailed the distribution is relative to a normal distribution. Data with high kurtosis tend to have heavy tails or outliers. If the kurtosis is low, there will be no outliers. A uniform distribution is an extreme case of low kurtosis.

Formula for Kurtosis

For a univariate data $X_1, X_2, \ldots\ldots X_n$ kurtosis is given by:

$$\text{Kurtosis} = \left(\sum_{i=1}^{N} (X_i - \bar{x})^4 / N\right) / s^4$$

Where $\bar{x}$ is the mean

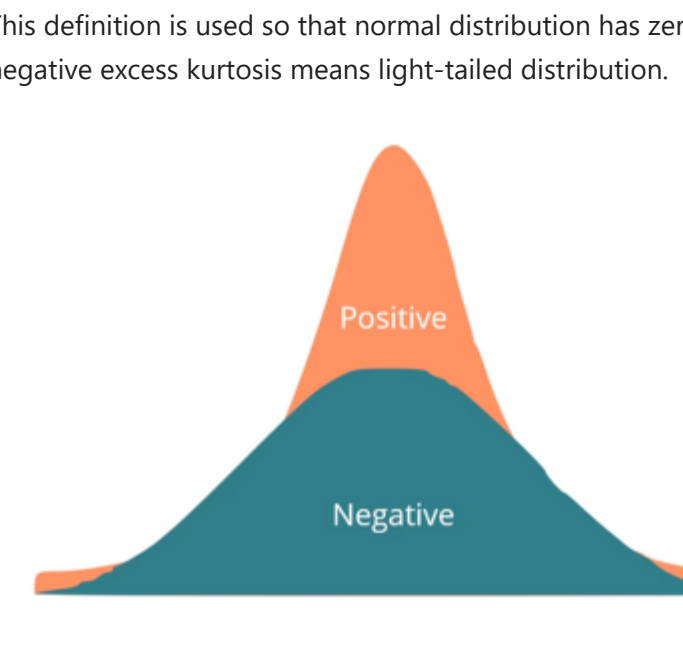
s is the standard deviation

N is the number of data points

The kurtosis for a normal distribution is 3. So sometimes the following definition of kurtosis is used. It's called excess kurtosis.

$$\text{Kurtosis} = \left(\sum_{i=1}^{N} (X_i - \bar{x})^4 / N\right) / s^4 - 3$$

This definition is used so that normal distribution has zero kurtosis. Also, positive excess kurtosis means heavy-tailed distribution and negative excess kurtosis means light-tailed distribution.



In [7]:
```python
# Python code implementation for kurtosis
import numpy as np

from statsmodels.stats.stattools import robust_kurtosis

x = np.array([2,4,5,7,8,9,11,15])

kurtosis  = robust_kurtosis(x)
kurtosis
```

Out[7]:
```
(-0.5988228527065362,
 -0.04888458916942784,
 -0.5052271228708025,
 -0.39005747798595314)
```

Covariance and Correlation

Covariance and correlation measure the relationship and dependency between two variables. While covariance gives the direction of the linear relationship, correlation gives both direction and strength. Therefore, correlation is a function of covariance. Furthermore, correlation values are standardized while covariance values are not.

Covariance

The covariance of x and y can be represented with the following equation:

$$\begin{aligned} \text{cov}(x,y) &= E[(x - \mu_x)(y - \mu_y)] \\ &= E[xy] - E[x]\,E[y] \\ &= E[xy] - \mu_x \mu_y \end{aligned}$$

Covariance Formula for Population

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Covariance Formula for Sample

$$\text{Cov}(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$

Where:

x_i is the i^{th} observation in variable x

$\bar{x}$ is the mean for variable x

y_i is the i^{th} observation in variable y

$\bar{y}$ is the mean for variable y

N is the number of observations

In [ ]:
```python
# Python implementation of the covariance of two variables
```

In [13]:
```python
import pandas as pd
df = pd.DataFrame({"X":[0, 1, 2, 3, 5, 9],
                   "Y":[11, 5, 8, 6, 7, 8],
                   "Z":[2, 5, 10, 11, 9, 8]})
print(df.cov())
```

```
           X    Y     Z
X  10.666667 -0.8   5.0
Y  -0.800000  4.3  -3.5
Z   5.000000 -3.5  11.5
```

The positive value denotes that both the variables move in the same direction.

Correlation

Correlation coefficient is also called the Pearson correlation coefficient.

The correlation coefficient of two variables by dividing their covariance with the product of their individual standard deviations. Since the standard deviation measures the absolute variability of the distribution, this division ensures that the correlation coefficient scales down to the -1 to +1 range.

Formula for Correlation Coefficient

$$\begin{aligned} \text{Corr}(x,y) &= \text{Cov}(x,y) / S_x S_y \\ &= E[(x-\mu_x)(y-\mu_y)] / S_x S_y \end{aligned}$$

Where:

S_x, S_y are standard deviations of x and y respectively.

In [15]:
```python
import pandas as pd
df = pd.DataFrame({"X":[0, 1, 2, 3, 5, 9],
                   "Y":[11, 5, 8, 6, 7, 8],
                   "Z":[2, 5, 10, 11, 9, 8]})
print(df.corr())
```

```
          X         Y         Z
X  1.000000 -0.118125  0.451447
Y -0.118125  1.000000 -0.497720
Z  0.451447 -0.497720  1.000000
```

The closer it is to -1 or 1, the higher the correlation.

A positive correlation coefficient implies that when one variable increases, the other also increases, and vice versa.

Note: Pearson, Kendall and Spearman correlation are currently computed using pairwise complete observations.