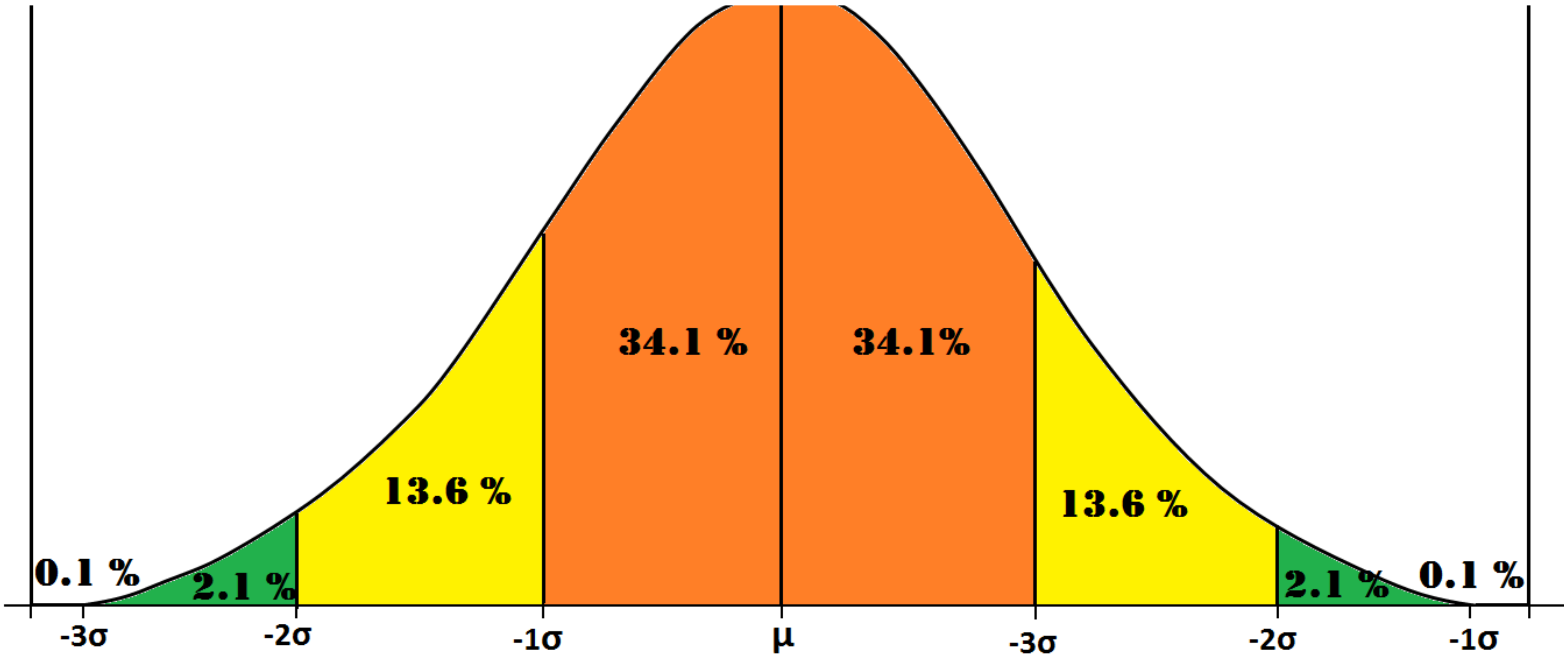
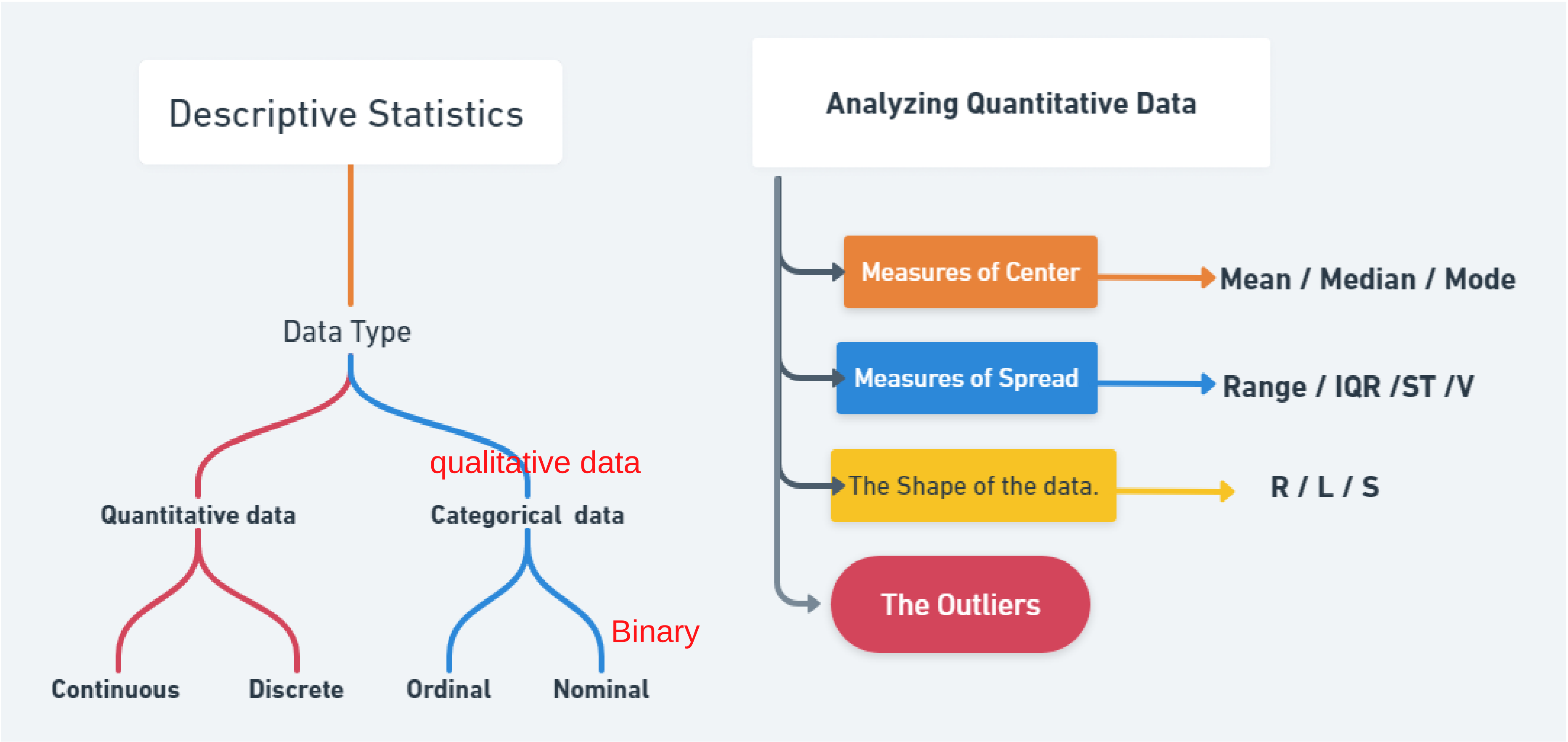


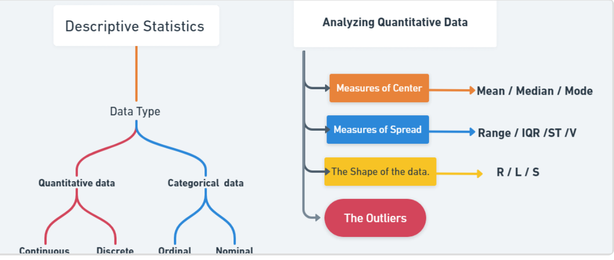
# Descriptive Statistics



Descriptive Statistics Summary Of Udacity Course

Descriptive Statistics Data Types Quantitative and Categorical. Quantitative data takes on numeric values that allow us to perform mathematical operations (like the number of dogs) Categorical are used to label a group or set of items (like dog breeds - Collies, Labs, Poodles, etc.

https://www.linkedin.com/pulse/descriptive-statistics-summary-udacity-course-engy-wahpa



## Data Types

### Quantitative and Categorical.

Quantitative data takes on **numeric values** that allow us **to perform mathematical operations** (like the number of dogs).

Categorical are used to **label a group** or **set of items** (like dog breeds - Collies, Labs, Poodles, etc.).

### Categorical Ordinal vs Categorical Nominal

We can divide categorical data further into two types: **Ordinal** and **Nominal**.

Categorical Ordinal data take on **a ranked ordering** (like a ranked interaction on a scale from `Very Poor` to `Very Good` with the dogs).

Categorical Nominal data **do not have an order** or **ranking** (like the breeds of the dog سلالات من الكالاب).

## Quantitative Continuous vs Quantitative Discrete

We can think of quantitative data as being either **continuous** or **discrete**.

Continuous **data can be split into smaller** and **smaller** units**,** and **still a smaller unit exists**. An example of **this is the age of the dog** - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

Discrete data **only takes on countable values.** The number of dogs we interact with is an example of a discrete data type.

## Quantitative: Examples

Continuous **: Height, Age, Income**

Discrete **: Pages in a Book, Trees in Yard, Dogs at a Coffee Shop**

## Categorical: Examples

Ordinal **: Letter Grade, Survey Rating**

Nominal **: Gender, Marital Status, Breakfast Items**

## Analyzing Quantitative Data

### Four Aspects for Quantitative Data

There are four main aspects to analyzing **Quantitative** data.

1. **Measures of** `Center`
2. **Measures of** `Spread`
3. **The** `Shape` **of the data.**
4. `Outliers`

## Analyzing Categorical Data

if we were looking at the breeds of the dogs, we would care about **how many dogs are of each breed, or what proportion of dogs** are of each breed type.

Categorical data is analyzed usually be **looking at the counts or proportion of individuals that fall into each group.**

## 1-Measures of Center

There are three measures of center:

1. `Mean`
2. `Median`
3. `Mode`

## 1- The Mean

The mean is often called the **average** or the **expected value** in mathematics.

We calculate the mean by adding all of our values together, and dividing by the number of values in our dataset.

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

## 2- The Median

The **median** splits our data so that **50%** of our values are **lower** and **50%** are **higher**.

### Median for Odd Values

If we have an **odd** number of observations, the **median** is simply the number in the **direct middle**.

### Median for Even Values

If we have an **even** number of observations, the **median** is the **average of the two values in the middle**.

## 3-The Mode

**The mode** is the most **frequently** observed value in our dataset.

There might be **multiple modes** for a particular dataset, or **no mode** at all.

## Notation

Notation : **Think of notation as a universal language used by academic and industry professionals to convey mathematical ideas. 5+3**

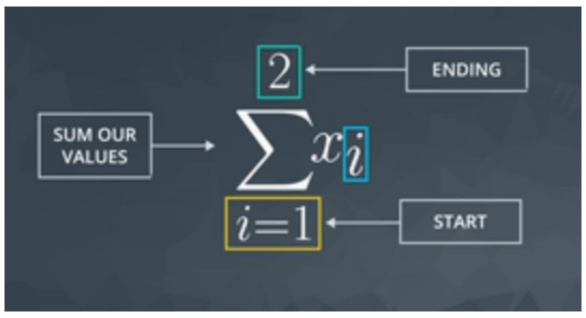
### Random Variables

A **random variable** is a **placeholder** for the possible values of some process

## Aggregations

An **aggregation** is a way to turn multiple numbers into fewer numbers (commonly one number).

**Summation** is a common aggregation. The notation used to sum our values is a **greek** symbol called sigma **Σ.**



## 2- Measures of Spread

**Measures of Spread** are used to provide us an idea of how spread out our data are from one another. Common measures of spread include:

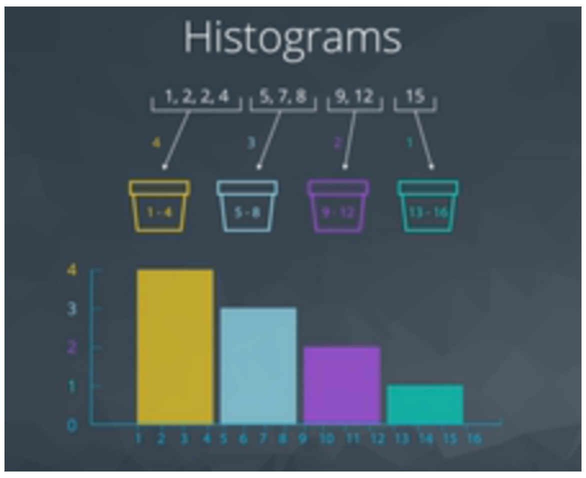
1. **Range**
2. **Interquartile Range (IQR)**
3. **Standard Deviation**
4. **Variance**

## Histograms المدرج التكرارى

**Histograms :** are super useful to understanding the different aspects of quantitative data. In the upcoming concepts, you will see histograms used all the time to help you understand **the four aspects** we outlined earlier regarding **a quantitative variable:**

- **center** - **Spread** - **Shape** - **Outliers**

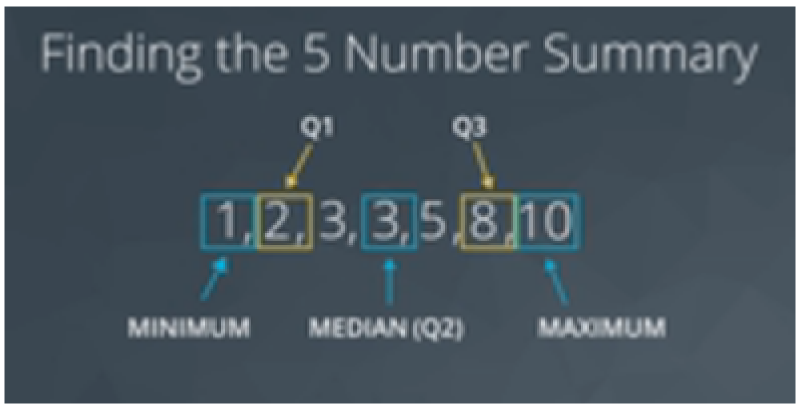
المدرج التكرارى هى مجموعة من البيانات بتتقسم لفئات و بتتحول لشكل بيانى



## Calculating the 5 Number Summary (Outliers Or Skewed )

The five number summary consist of 5 values:

1. **Minimum:** The **smallest** number in the dataset.
2. **Q1**: The value such that **25**% of the data **fall below.**
3. **Q2**: **(Median)** The value such that **50**% of the data fall below.
4. **Q3**: The value such that **75**% of the data **fall below.**
5. **Maximum:** The **largest** value in the dataset.



## 1- The Range = ( Max - Min )

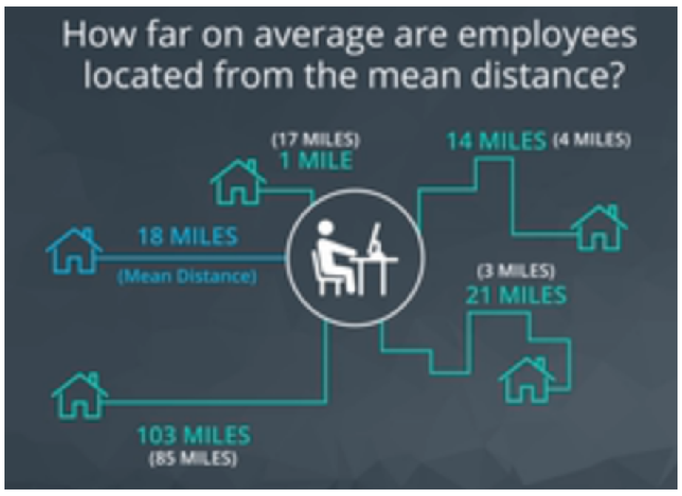
The **range** is then calculated as the **difference** between the **maximum** and the **minimum**.

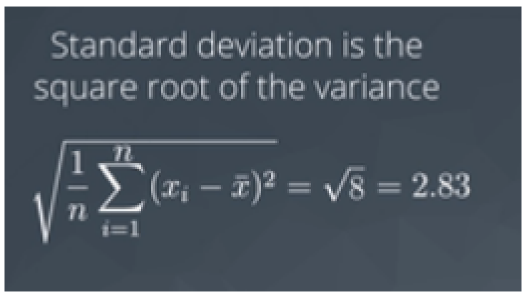
## 2- Interquartile Range (IQR) Q3 - Q1

The **interquartile range** is calculated as the **difference** between **Q3** and **Q1**.

### 3-The Standard Deviation

The **standard deviation** is one of the most common measures for talking about the spread of data. It is defined as **the average distance of each observation from the mean**.




Standard deviation is the square root of the variance

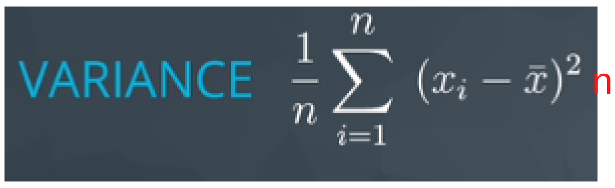
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\sqrt{8}=2.83$$

- **The standard deviation** is associated with risk in finance, **assists in determining the significance of drugs in medical studies**, and **measures the error of our results** for **predicting anything from the amount of rainfall we can** expect **tomorrow to your predicted commute time tomorrow.**

### 4-The Variance التفاوت

The Variance : **is the average squared difference of each observation from the mean**

.


VARIANCE $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ ni

- **The variance** is **used to** **compare the spread of two different groups**. A set of data with **higher variance is more spread out than a dataset with lower variance**. Be careful though, there might just be **an outlier** (or outliers) **that is increasing the variance,** when most of the data are actually very close.

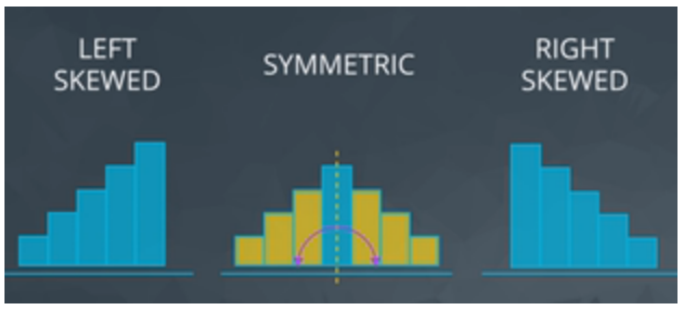
## 3- The Shape Of Data

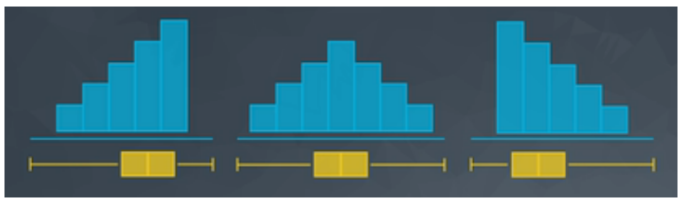
From a **histogram** we can quickly identify **the shape of our data,** which helps influence all of the measures we learned in the previous concepts. We learned that **the distribution of our data is frequently associated with one of the three shapes:**

**1. Right-skewed**

**2. Left-skewed**

**3. Symmetric** (frequently normally distributed)

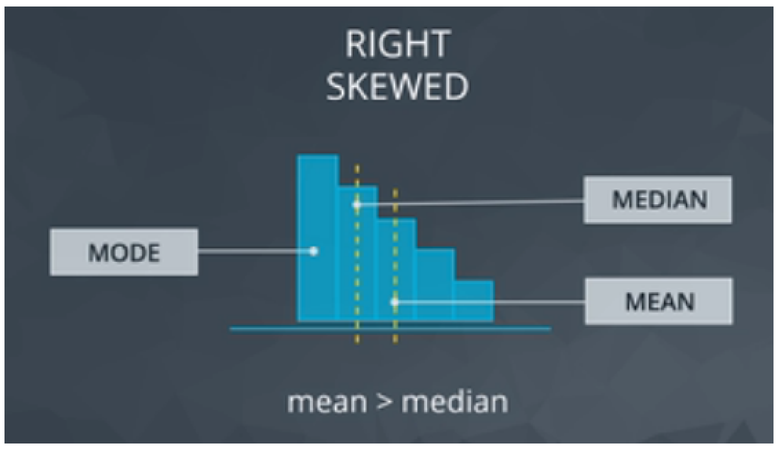




## 1- Right skewed  Mean > Median

### Real World Applications

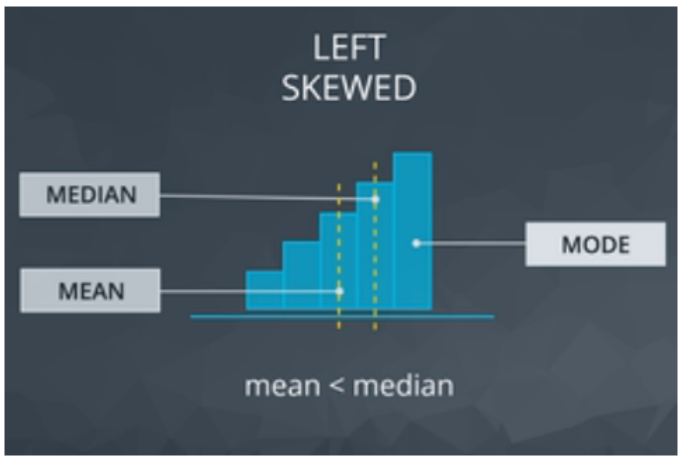
- **Amount of drug remaining in a blood stream,**
- **Time between phone calls at a call center**,
- **Time until light bulb dies**



## 2 - Left skewed   Median > Mean

### Real World Applications

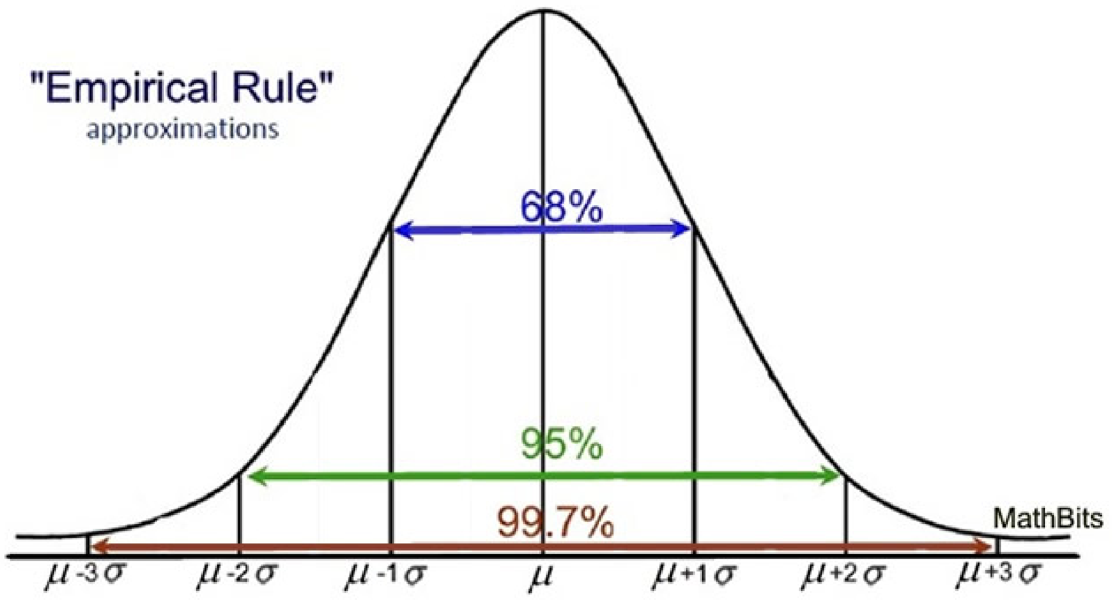
- Grades as a percentage in many universities,
- Age of death,
- Asset price changes



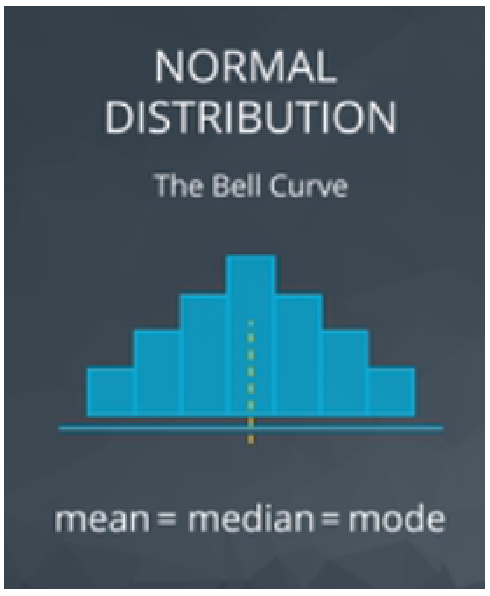
## 3 - Symmetric (frequently normally distributed) Median = Mean

### Real World Applications   ( Mean And Standard Deviation)

- Height,
- Weight, Errors,
- Precipitation



- 68% of the distribution lies within **one** standard deviation of the mean.
- 95% of the distribution lies within **two** standard deviations of the mean.
- 99.7% of the distribution lies within **three** standard deviations of the mean.



## 4- Outliers

**outliers : are points that fall very far from the rest of our data points.** This influences measures like the mean and standard deviation **much more than measures associated with the five number summary.**

### Outliers Advise

**1. Plot** your data to identify if **you have outliers.**

**2. Handle outliers** accordingly via the methods above.

**3. If no outliers** and your data follow **a normal distribution** - use the **mean** and **standard deviation** to describe your dataset, and report that the data are normally distributed.

**4.** If you **have skewed** data or **outliers**, use **the five number summary** to **summarize your data and report the outliers.**

## Descriptive Statistics

`Descriptive statistics` **is about describing our collected data**.

---

## Inferential Statistics

`Inferential Statistics` **is about using our collected data to draw conclusions to a larger population**.

We looked at specific examples that allowed us to identify the

1. **Population -** our entire group of interest.
2. **Parameter** - numeric summary about a population
3. **Sample -** subset of the population
4. **Statistic -** numeric summary about a sample