Statistics: numbers that have been collected in order to provide information about something.

Also defined as science of collecting and studying numbers.

In [ ]:
```
1
```

Types of Statisitics

1) **Descriptive Statistics**: In descriptive statistics the data is described in a summarized way. The summarization is done from the sample of the population using different parameters such as **Mean or Standard Deviation**.

Descriptive statistics are a way of using charts, graphs & summary measures to organize, represent and explain the set of data.

Data is typically arranged and displayed in table or graphs summarizing detail such as histogram, pie chart,bars,scatter plots etc.

Descriptive statistics are just descriptive and thus does not require any generalization beyond the data collected.

In [ ]:
```
1
```

2) **Inferential Statisitics**:

In Inferential statistics we try to interpret the outcome. After the data has been collected,analysed & summarized we use inferential statistics to describe the meaning of the collected data.

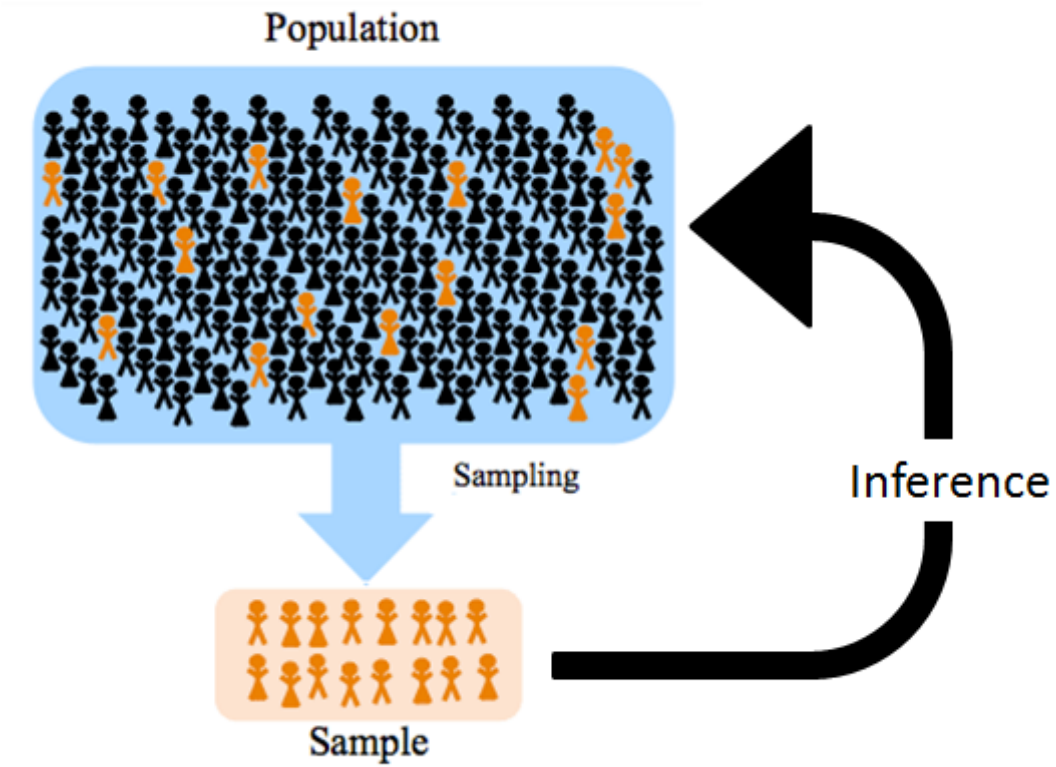
In Inferential statistics, it is majorly intended to test hypothesis and investigate relationship between variables and can be used to make predictions.

They are typically used to draw conclusion and inference, i.e to make a valid generalization from sample data

In [ ]:
```
1
```

Population: It is actually a collection of a set of individual or object or events whose properties are yet to be analyzed.

Sample : It is the subset of the population



In [ ]:
```
1
```

Measure of Central Tendency

Measure of Central Tendency is also known as summary in statistics that is used to represent the center point of a particular value of data set or sample set.

1) Mean

2) Median

3) Mode

```
1) Mean (Average)


Formula: sum of all the terms / total no of terms


Example:


Cars    Milage    Cylinder
Swift    21.3        3
BMW      20.8        4
Audi     18          5
```

```
Milage Mean = 21.3+20.8+18 / 3

            = 20.03 ===> Mean value for milage
```

```
Cylinder Mean =   3+4+5/3

              =   4 ====> Mean value for cylinder
```

In [1]:
```
21.3+20.8+18
```

Out[1]: 60.1

In [2]:
```

```

```
2) Median :


Cars    Milage    Cylinder
Swift    21.3        3
BMW      20.8        4
Audi     18          5
Ertiga   15          4
```

If your number of data point is even:

Step 1: Arrange data in ascending order

15 18 20.8 21.3 10 20

==========

10 15 18 20 20.8 21.3

18+20 ===> 38/2 ===> 19

Step 2: Check the midpoint : 18 + 20.8 ===> 38.8/2 ===> 19.4 (median value)

In case of Odd Numbers:

21.3 20.8 18 10 122

Median Value : 18

```
Mode:

10
2
3
4
2
6
4
2
7
7
10
4
2
```

In [ ]:

```

```

In [ ]:

```

```

Measure of variability

In [ ]:
```
1
```

What is variability : It describes how far apart the data points lie from each other and from a center of distribution

Variability is also refferd as spread OR scatter OR dispersion

In [ ]:
```
1
```

Why variability matters

1) While measure of central tendency tells us where most of the point lie, variance on the other hand tells how far apart the data points are

2) This is important because it tells you whether the points tends to be clustered or more widely spread out.

3) Low variability is ideal beacuse it means that you can better predict information about sample data based on population.

4) High Variability means that the value are less consistent so it is harder to make predictions.

In [ ]:
```
1
```

Range:

It shows the spread of the data from the lowest to the highest value in distribution.

It is the easiest way of finding the measure of variability.

```
1 Example:
2
3 Data
4 72
5 110
6 134
7 190
8 238
```

```
287
305
324
```

Range = Highest value - lowest value

```
324 - 72

= 252
```

In [ ]:

```

```

Inter Quartile Range (IQR) : It gives you the spread of the data from middle of your distribution

```
Q1 ==> Quartile 1 ==> 25% of value from data.
Q2 ==> (the median) ==> median value (mid value)
Q3 ==> (Quartile 3) ==> 75% of the value
Q4 ==> (Quartile 4) ==> Highest value


Formula : Q3 - Q1

Example:

72
110
134
190
238
287
305
324


Q1 ==> 0.25 * 8 ==> 2

       2 ===> (value which is at the second position)

```

```
       value is for q1 ==> 110


Q3 ==> 0.75 * 8 ==> 6

       6 ==> (value which is at the 6th Position)

       value for q3 is 287



IQR = Q3 - Q1
    = 287 - 110

    = 177

```

```
72
110
134
190
238
287
305

Q1 = 0.25 * 7
   = 1.75  ==> 2

   == 110


Q3 =  0.75 * 7

   =  5.25 ==> 5

   =  238


IQR = 238 - 110

```

```
    = 128


```

In [5]:
```
import numpy as np
```

In [ ]:
```

```

Standard Deviation

It finds out the average of variability in your data set

It tells you on an average how far the score (data points lie) from the mean

The larger the standard deviation is the more variance (spread) the data is

Steps:

1) List each score and find out the mean value.

2) substract mean value from each data to get the deviation from the mean

3) square each of those deviation.

4) Add up all the squared deviation

5) Divide the sum of squared deviation by n-1 (where n is the sample data size)

6) find the square root of the number

```
Example:

72
110
134
190
238
287
```

```
305
324



Step 1:

72+110+134+190+238+287+305+324 / 8

==> 207.5



Step 2:                          Deviation from mean

72                               72 - 207.5  ==> -135.5
110                              110-207.5   ===> -97.5
134                              134-207.5  ===>  -73.5
190                              190-207.5  ====> -17.5
238                              238-207.5  ===>  30.5
287                              287-207.5  ===> 79.5
305                              305-207.5  ===> 97.5
324                              324-207.7  ===> 116.5


Step 3

Square Of Deviation

18360.25
9506.25
5402.25
306.25
930.25
6320.25
9506.25
13572.25


Step 4

Add up all the squared deviation
```

```
18360.25
9506.25
5402.25
306.25
930.25
6320.25
9506.25
13572.25

=========
63904.0



Step 5:

Divide the sum of squared deviation by n-1


63904.0 / 7

==> 9129.14


Step6 : Find square root

sqrt(9129.14)

===> 95.54




It means that the average sum of deviation from mean is 95.54
```

In [6]:
```
1
```

Out[6]: -97.5

In [8]:
```
np.sum([18360.25,
9506.25,
5402.25,
306.25,
930.25,
6320.25,
9506.25,
13572.25])
```

Out[8]: 63904.0

In [10]:
```
import math
math.sqrt(9129.14)
```

Out[10]: 95.54653316578262

In [ ]:
```

```

Covariance

Covariance provides a measure of stregth of the relation between two or more sets of random variates.

In covariance it tells how two variables vary together.

```
Cov(X,Y) = SUM E((X-U) E(Y-v)) / n-1



where

X is random variable
E(X) = U is the expected value (the mean) of the random variable x
E(Y) = v is the expected value (th mean) of the random variable Y
```

```
n = number of item  in the data set

SUM ==> Summation Sign
```

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \overline{X})(Y_j - \overline{Y})}{n}$$

```
X           Y

2.1         8
2.5         10
3.6         12
4.0         14


mean of X  =  2.1+2.5+3.6+4.0 / 4   ===> 3.05

mean of Y  =  8+10+12+14/4  ==>   11

COV(X,Y) = (2.1-3.05)(8-11) + (2.5-3.05)(10-11) + (3.6-3.05)(12-11) + (4.0-3.05)(14-11) / 4-1

         = (-1)(-3) + (-0.6)(-1) + (0.5)(1) + (1)(3) / 3

         =  3 + 0.6 + 0.5 + 3 / 3

         = 7.1/3

         = 2.36
```

In [17]:
```
2.5-3.05
```

Out[17]:
```
-0.5499999999999998
```

In [16]:
```
(-1)*(-3)
```

Out[16]:
```
3
```

In [18]:
```
(-0.6)*(-1)
```

Out[18]:
```
0.6
```

In [ ]:
```

```

corelation

Corelation is a statistical technique that measure the degree to which two or more variables are related to each other.

The value of corelation co-efficient must fall between -1.0 to 1.0

It can only measure association, but does not show whether X can cause effect on Y OR Vice versa

In [30]:
```
57*57
```

Out[30]:
```
3249
```

```
AGE(x)       SUGAR LEVEL(y)

43          99
21          65
25          79
42          75
57          87
59          81
```

```
Step 1 is to calcualte X * Y

```

```

XY

4257
1365
1975
3150
4959
4779

```

```
Step 2 is to calcualte (X)**2


X**2

1849
441
625
1764
3249
3481
```

```
Step 3 is to calculate (Y)**2


Y**2

9801
4225
6241
5625
7569
6561
```

In [35]:
```
np.sum([9801,
4225,
6241,
5625,
7569,
6561])
```

Out[35]: 40022

In [ ]:
```

```

In [ ]:
```

```

```
Step 4 is to do the sum of all columns


SUM X      SUM Y      SUM XY     SUM X**2   SUM Y**2

247        486        20485       11409      40022
```

In [ ]:
```

```

Formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

In [ ]:
```
6(20485) - (247*486)
------------------------
sqrt([6(11409)-(247)**2]*[6(40022)-(486)**2])
```

In [ ]:
```

```

The corelation coefficient is : **0.5298091541711386**

In [ ]:
```

```

In [ ]:
```

```

In [38]:
```
1 6*(20485) - (247*486)
```

Out[38]:
```
2868
```

In [40]:
```
1 math.sqrt((6*(11409)-(247)**2) * (6*(40022)-(486)**2))
```

Out[40]:
```
5413.272577655775
```

In [41]:
```
1 2868/5413.27
```

Out[41]:
```
0.5298091541711386
```

In [ ]:
```
1
```