Statistics Learning – Day 5

The entire statistics is divided into the 2 parts namely :

- Descriptive Statistics,
- Inferential Statistics.

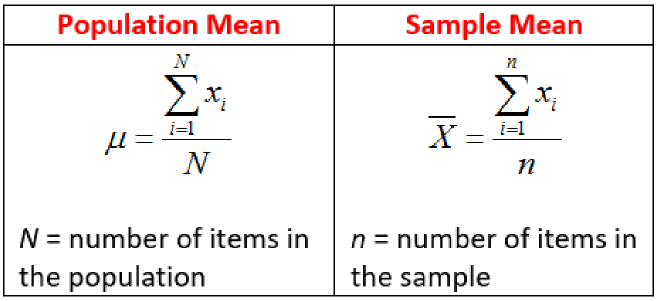
A, DESCRIPTIVE STATISTICS

1. Measure of Central Tendency: (Mean, Median, Mode)

(i)Mean:

Mean means nothing but it's just average value. (add all the values and divide by no. of values)

Population Size (N), Sample Size (n)



Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	*n* = number of items in the sample

X = {1,1,2,2,3,3,4,5,5,6} = 32/10 = 3.2

(ii) Median:

Mean means nothing but it's just a middle value.

How to find median?

(i) Sort the random Variables (1,2,2,3,4,5)	(ii) No. of elements	iii) If count = even Here I have 6 elements so I take middle two elements and finding median. {1,2,2,3,4,5} = (2+3)/2 = 5/2 = 2.5	If count = add Here I have 7 element so I take the middle one element and that is my median value. {1,2,2,3,4,5,6} = Median = 3

Why Median?

It is because of outliers, for example x is my random variable, x={1,2,3,4,5}
(1+2+3+4+5)/5 = 3

Let suppose If I have 1 outlier in my data.   x = (1+2+3+4+5+100)

Mean = (1+2+3+4+5+100) / 6 = 19.1667
Median = (1+2+3+4+5+100) = 3.5

Here I got my mean is nearly 20, but my median is just 3.5, so median will work perfectly in outliers data.

Conclusion:

Median is used to find the central tendency when the outlier is present.

(iii) Mode:

Mean means nothing but it's a most repeated elements.

How to find mode?

I should find which element is presented many times in my dataset.
For example,

X = {2,1,1,1,4,5,7,7,7,7,7,7,7,8,9,9,10,}

MODE = 7

Here, My mode is 7, because 7 is present maximum times presented in my data.

Mean, Median, Mode is specifically used in EDA, & Feature engineering. I will let you know about everything in future writings.

Where do we use Mean, Median, & Mode?

This is specifically Used in EDA and Feature Engineering,

For example, This is my dataset and I am having a lot of missing values.

We can handle our missing values in feature engineering steps, we did not drop the entire row where missing value is available. Suppose If we drop the entire row, we loss some data and it is our loss.

AGE	WEIGHT	SALARY	GENDER	DEGREE
24	70	40k	M	-
25	-	33k	-	BE
-	25	60k	F	MSC
40	72	55k	F	-

MODE is specifically use for categorical value replacement, Mean & Median is specifically used for numerical value replacement.

In numerical column first we should check whether the age has outliers or not, if it has outliers we will specifically use median, if not by using mean we can replace that numerical missing values.

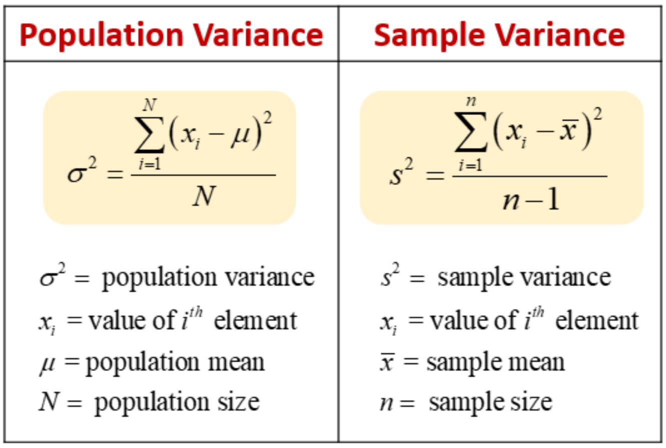
2. Measure of Dispersion: (Spread of the data)

In measure of variance two different things. There are **Standard Deviation (σ), Variance (σ^2).**

(whenever we are talking about measure of dispersion - we are talking about spread of the data, how much data is spread, how much data is distributed).

(i) Variance:

Variance means how much well your data is basically spread.



Population Variance	Sample Variance
$\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
$\sigma^2 =$ population variance $x_i =$ value of i^{th} element $\mu =$ population mean $N =$ population size	$s^2 =$ sample variance $x_i =$ value of i^{th} element $\bar{x} =$ sample mean $n =$ sample size

In measure of variance two different things. There are **Standard Deviation (σ), Variance (σ^2).**

(whenever we are talking about measure of dispersion - we are talking about spread of the data, how much data is spread, how much data is distributed).

(ii) Standard Deviation:

Standard deviation is nothing but how much of data deviate from the mean.

Standard deviation is the root of variance.

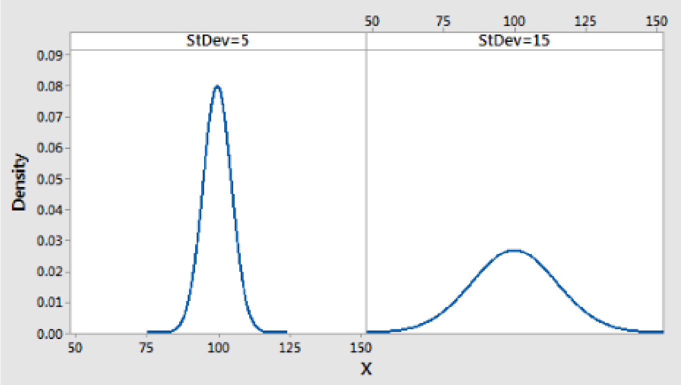
Population Standard Deviation	Sample Standard Deviation
$\sigma = \sqrt{Variance}$	S = $\sqrt{s^2}$

What does the variance basically specified?

Let suppose I have a 2 distribution random variable X and Y.

For example, let suppose,

My x distribution sample variance is 2.5 &
My y distribution sample variance is 7.5



When the variance is less, our spread will be less,
When the variance is high, our spread will be high.

Random Variable:

Random variable is a process of mapping the output of a random process / experiment to a number.

Example:

(i) Tossing A Coin:

{Head, Tail}

X = { 0 if Head, 1 if Tail}

(The outcome of the process is converted to a number & this number is fixed)

Histogram:

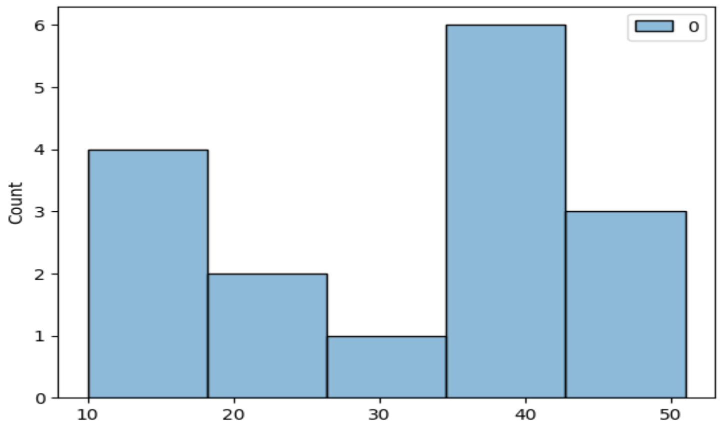
Histogram -----> Whenever we are talk about histogram, it's a **frequency**

The best visualization diagram we using is called as histogram.

(Example)

Ages = {12,10,14,18,24,26,30,35,36,37,40,41,42,43,50,51}

If I want to visualize this ages, specifically we use histogram.

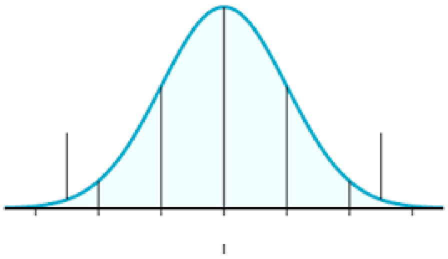


Explanation: In above chart from age 10 to 20 - 4 persons are in our data, 20 to 30 2 persons are in our data.

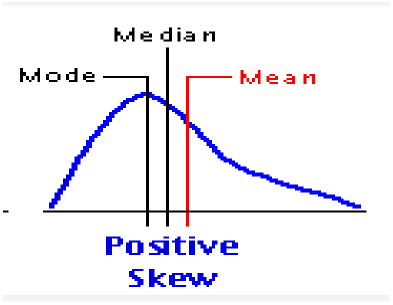
Skewness:

What exactly skewness?

Skewness basically means that right hand side elongated or left hand side elongated, its not a symmetrical distribution in shot. (symmetrical distribution means normal distribution)



(ii) Right Skewed:



(iii) Left Skewed:

Left skewed is called as negative skewed,
My left hand side will be elongated
Here my relationship is mode > median > mean

Percentile:

Before understanding percentile and quantile, we should know about percentage %.

Percentage:

1,2,3,4,5,6

My question is what is the % of numbers in odd?

%of no. in add = no. of odd no. / no. of total no. = 3/6 =1/2 = 50%
50% of numbers in odd.

Percentile:

Definition : Percentile is the value below which certain % of data points lie. Lets say my problem is
X = {2,3,3,4,6,6,6,7,8,8,9,9,10,11,12}

Now my question is what is the percentile rank of 10?

Formula :

Percentile rank of 10 = (No. of values below 10 / n) * 100
(12 / 15) * 100 = 80 percentile

What is 80 percentile basically means?

80% of the distribution fall below the value of 10

X = {2,3,3,4,6,6,6,7,8,8,9,9,10,11,12}

Now my question is what value exist at 25 percentile?

Formula :

(Percentile / 100) * (n+1)

(25/100) * (15+1)   ----------> n is nothing but is count of value

(1/4) * 16 = 4th element

X = {2,3,3,4,6,6,6,7,8,8,9,9,10,11,12}

Here my 4th element is 4. 25% of this values are below 4

Remember : If data is not in ascending format -------> First sort it.

Quantile:

Q1 = 25 Percentile
Q2 = 50 Percentile (Median)
Q3 = 75 Percentile

(Before understand this lets understand about outliers, why outliers playing very big role?)

FIVE NUMBER SUMMARY:

1. Minimum
2. First Quartile (25 Percentile) - Q1
3. Median - Q2
4. Third Quartile (75 Percentile) - Q3
5. Maximum

(Lets go & see how to remove the outliers with the help of this)

X = {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,29}

Out of all this number which number is outlier?

Outlier simply means that for example:

(i) In a class almost every student is same height approximately, but only one student is very tall or very short, that person is a outlier

(ii) In class 12 every student age is 15 or 16, but only one student age is 22, that person is a outlier

(Lets go & see how to remove the outliers with the help of this)

In Five number summary we should focus on two things one is lower bound, another one is upper bound.

Below the lower bound value and above the upper bound value, everything is consider as a outlier.
Lets find out how to find the outlier. (Because lower bound and upper bound is a border we are focusing in the inside values range only).

Formula:

Lower Bound = Q1 - (1.5*+IQR)
Upper Bound = Q3 + (1.5*IQR)

IQR Means Inter Quartile Range

Formula:

Lower Bound = Q1 - (1.5*+IQR)
Upper Bound = Q3 + (1.5*IQR)
IQR = Q3 - Q1

X = {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,29}

1st Quantile:

Q1 = 25 Percentile

Percentile / 100 * (n+1) = 25 / 100 * (19+1) = (4/100) * 20= 5th element = 3 is my 1st Quantile

3rd Quantile:

Q3 = 75 Percentile

= Percentile / 100 * (n-1) = 75/100 * (19+1) = 15th element = 7

IQR = Q3 - Q1

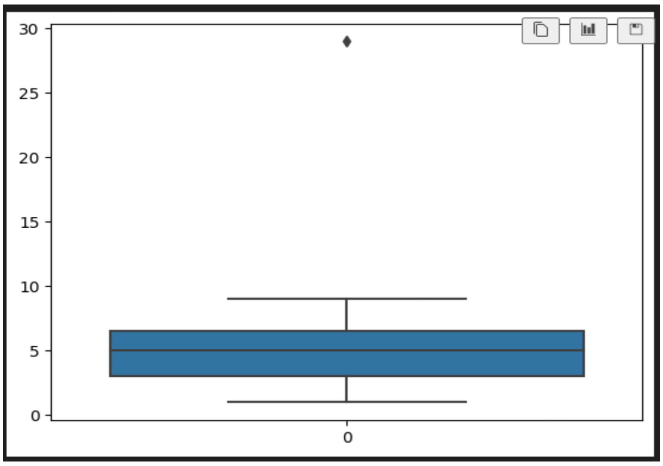
7-3=4 IQR = 4

Lower Bound = Q1 - 1.5*IQR = 3-1.5*4=-3
Upper Bound = Q3 - 1.5*IQR = 7+1.5*4 = 13

So from -3 to 13 is my good value and rest all of my other values are consider as a outliers

X = {1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,29}

Here 29 is outlier we should remove that.



In a similar way, we do one more sum,

X = {1,2,4,6,7,12,18,34,77,66,108,94,14}

First sort it X = {1,2,4,6,7,12,14,18,34,66,77,99,108}

Minimum = 1	1st Quantile (Q1) = 25 Percentile = (25/100) * (n+1) = 1/4 * (14) = 3.5th element = 5	Median = 14	3rd Quantile (Q3) = 75 Percentile = (75/100) * (n+1) = 3/4 * 14 = 10.5th element = 71.5	Maximum=108

Formula:

IQR = Q3 - Q1 = 71.5 - 5 = 66.5
Lower Bound = Q1 - (1.5*+IQR) = 5 - (1.5 * 66.5) = -94.75
Upper Bound = Q3 + (1.5*IQR) = 71.5 + (1.5 * 66.5) = 171.25
IQR = Q3 - Q1

So from -94.75 to 171.25 is my good value and
Rest all of my other values are consider as a outliers.

X = {1,2,4,6,7,12,14,18,34,66,77,99,108} Here 108 is outlier we should remove that.

Will Continue
Santhosh Kumar