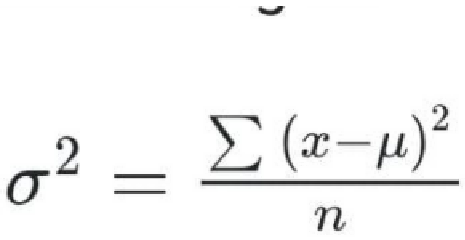
ASSIGNMENT 2

1. **WHY THE DENOMINATOR OF SAMPLE VARIANCE IS (n-1) WHEN THE DENOMINATOR OF THE POPULATION VARIANCE IS N?**

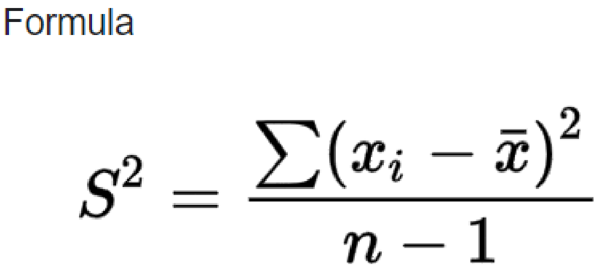
Let us have a look at the formulas of population and sample variance:

Population variance formula:



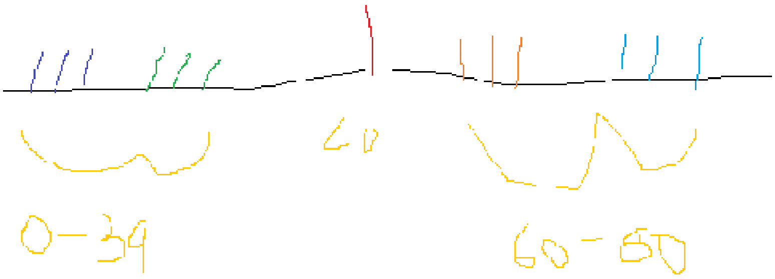
$$\sigma^2 = \frac{\sum (x-\mu)^2}{n}$$

Sample variance formula:



Formula

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



There are two reasons for using n-1 as the denominator for sample variance:

First, we need to understand that we are approaching the term sample as a part of the population from which we need to make the required predictions. Hence, whatever result we get from the sample in the calculation is reflected in the population as an opinion.

Let's take the above graph as an example of finding the average age of people in a city and the red mark in the centre is the mean age. If we take the samples only from the left side or right side of the spread data then we can't use the same denominator i.e. n as we use in population variance, so (n-1) is considered the perfect choice which brings a similar output value to the population variance.

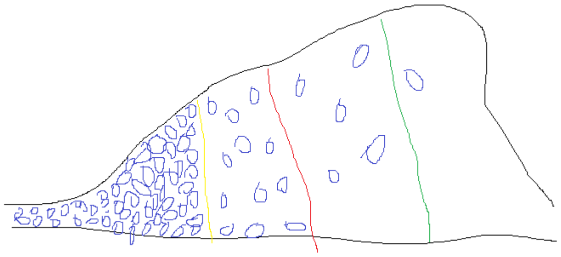
And on the other hand, even if we take more sample size, still we can able to get a similar output for sample variance when compared with population variance with the (n-1) approach which brings an unbiased approach to the calculation.

ASSIGNMENT-3

WHAT ARE LEFT SKEW, RIGHT SKEW, AND SYMMETRICAL DISTRIBUTION, AND HOW THE MEAN, MEDIAN, AND MODE ARE RELATED TO THESE DISTRIBUTION METHODS?

Let's have a look at the skews separately and discuss their relationship with mean, median, and mode with an example.

Left skew:



Yellow- mean

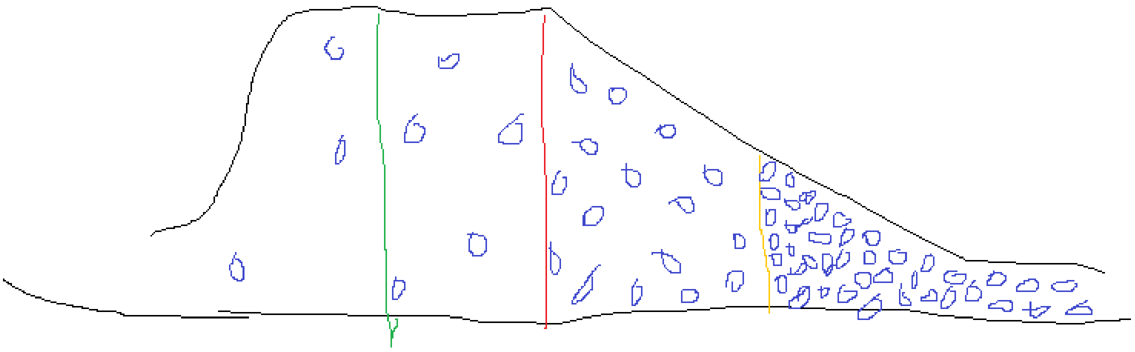
Red- median

Green- mode

This is called left skew or negative skew, where most of the values et distributed in the range of the lowest limit to the median point. Let’s take the height of people from a particular country as an example and approach this concept, the above diagram states that the majority or maximum of the values in the dataset are spread in the lower limit and only a few values are above the upper limit, hence we can say that the above diagram the height Japanese people are spread over and the majority of their height lies below the mean value and somewhere near the right-hand side of the mean value. And close to 10% or fewer values i.e. from the mode (green colour) is spread on the right side. Hence this is called left-skewed distribution means most of the values are on the left side or the minimal side.

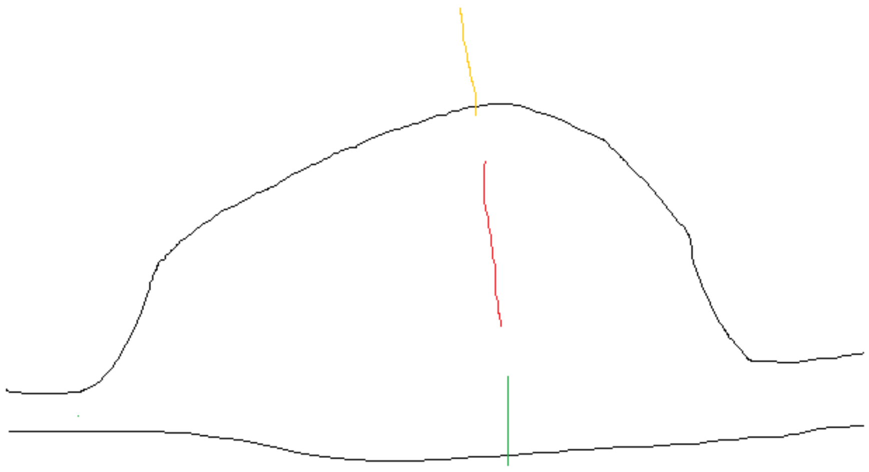
If the average height of the Japanese is 5.2 feet then most of the values fall within it(below the red mark) and there are minimal possibilities that the height of Japanese might fall from 5.2-5.5 feet (maybe for example) that gets rarely spread in the right side. Hence here the value of mode will be the highest (i.e. lesser number of tall people) and in terms of mean and median, the distribution range will be quite close with more number of people.

Right skew:



The right or the positive skew is opposite to that of the left skew concept, here the mean is the highest one followed by the median and then mode. We can also see more data got spread in the right-skewed region. Coming to our example we can say the Dutch people are the tallest race of humankind and hence we can see that more number of tall people lies on the right side of the distribution and it gradually getting decreased when it moves towards the left.

Symmetrical or normal distribution



In the symmetrical or normal distribution, the value of the mean, median and mode will almost be similar. It states that as per example there will be not much difference between the average height of a person when compared to the lower side that falls on the left side and the upper or above the average height and this scenario we can take for India.