MEAN-

Mean is simply another name for average. To calculate the mean of a data set, divide the sum of all values by the number of values.  
Consider the following set of numbers: {5,2,2,7}. The mean is  (5 + 2 + 2 + 7) / 4 = 16 / 4 = 4. We use the symbol “x-bar” to represent the mean of a sample data. The formula to compute the mean for a set of n values is:

We will explain terms like standard deviation and normal distribution in subsequent blogs. For now, all we need to keep in mind is the sample size (10,000), and the mean (25,000). Don’t worry about other components like numpy for code, or the criteria for calculation.

Code:-

import numpy as np

expenditure = np.random.normal(25000, 15000, 10000)

np.mean(expenditure)

Median-

Median, in simple words, is the number that lies in the middle of a list of ordered numbers. The numbers may be in the ascending or descending order. Let us consider the following data set:  
0,2,3,4,5,1,2,0,6  
After sorting these numbers in the ascending order, we get the following list:  
0,0,1,2,2,3,4,5,6  
2 – the number in the center (fifth from either side) – is the median in this example.

The median is easy to find when there are odd number of elements in the data set. When there are even number of elements, you need to take the average of the two numbers that fall in the center of the ordered list. So, if we consider the following data set:  
0,0,1,4,2,3  
After sorting the numbers, we get the following list:  
0,0,1,2,3,4  
The average of 1 and 2, in this case, is the median.

Median = (1 + 2) / 2  
         = 1.5  
Median is 1.5.

Code:-

expenditure = np.append(expenditure, [1000000000])

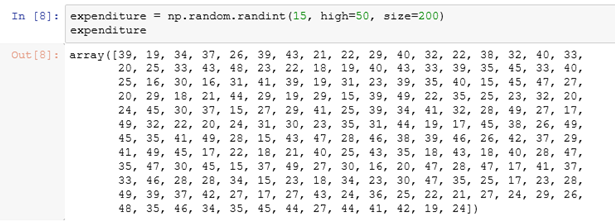
np.median(expenditure)

np.mean(expenditure)

The median of expenditures from the previous example is 25,179.05. In this case, it is not very far from the mean, which is 25,120.24.  
Before we discuss mode, let us understand what outliers are, and how they impact the mean of a data set.

* Any value in a dataset that is at an abnormal distance from all other values can be termed as an outlier. Outliers generally tend to skew the mean radically.
* Outliers can be present in the dataset with very high value or with a very low value.

Mode:-

Let’s generate a random expenditure set data using the script below.  
expenditure = np.random.randint(15, high=50, size=200)  
expenditure  


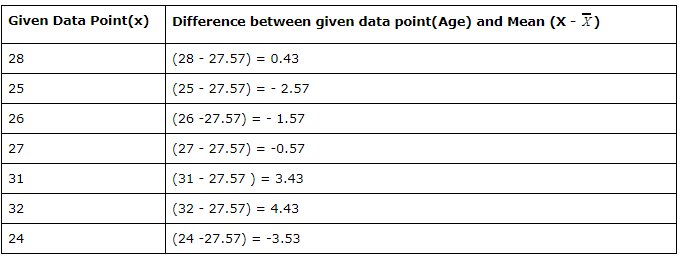
from scipy import stats

stats.mode(expenditure)

Variance:-

Variance is the measure of dispersion in a data set. In other words, it measures how spread out a data set is. It is calculated by first finding the deviation of each element in the data set from the mean, and then by squaring it.   
Variance is the average of all squared deviations.

Steps to Calculate Variance:

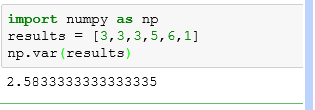
1. List elements of data set.  
The following are ages of students pursuing a Master’s degree:  
Data set 1: 28,25,26,27,31,32,24  
2. Calculate the mean.  
(28 + 25 +26 +27 +31 +32 + 24) / 7 = 27.57  
3. Find the deviation from the mean for each data point.  
  
   4. Square it.  
  
5. The average of all squared differences is the variance. To find it, add all squared variances and divide the sum by a number of elements in data set (n).  
(0.1849 + 6.6049 + 2.4649 + .3249 + 11.76 + 19.6249 + 12. 4609) / 7   
53.4303 /7 = 7.6329  
Now that we know how to calculate the variance of a data set, let us look at how to find the same using Python.  
Consider a list of random integers (data set 2) – 3,3,3,5,6,1. We will now calculate the variance using numpy library.

Code:

import numpy as np

results = [3,3,3,5,6,1]

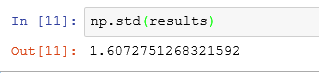
np.var(results)

  
As we can see, the variance of the random data set is 2.58.

Standard Deviation-

Standard deviation tells about the concentration of the data around the mean of the data set. Standard deviation is inversely proportional to the concentration of the data around the mean i.e with high concentration, the standard deviation will be low, and vice versa. It cannot be negative. The value of standard deviation can be easily impacted by outliers as a single outlier (abnormal value) distorts the overall mean, and thereby, deviation from the mean of all elements.  
If, to find variance we square the deviations of individual elements from the mean, then to calculate standard deviation, we need to calculate the square root of the variance.  
We calculate the square root of the variance:  
https://s3.amazonaws.com/acadgildsite/wordpress_images/Data+Science/variance/image4.PNG  
Standard deviation of the data set 1 is 2.76.  
To find standard deviation using Python, we will use data set 2. The numbers are listed below, and we already know the variance.  
results = [3,3,3,5,6,1]  
To calculate standard deviation use the inbuilt function “std” from numpy as shown below:

Code:



Gaussian Distribution-

Normal distribution represents the behavior of most of the situations in the universe (That is why it’s called a “normal” distribution. I guess!). The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application. Any distribution is known as Normal distribution if it has the following characteristics:

1. The mean, median and mode of the distribution coincide.
2. The curve of the distribution is bell-shaped and symmetrical about the line x=μ.
3. The total area under the curve is 1.
4. Exactly half of the values are to the left of the center and the other half to the right.

A normal distribution is highly different from Binomial Distribution. However, if the number of trials approaches infinity then the shapes will be quite similar.

The PDF of a random variable X following a normal distribution is given by:

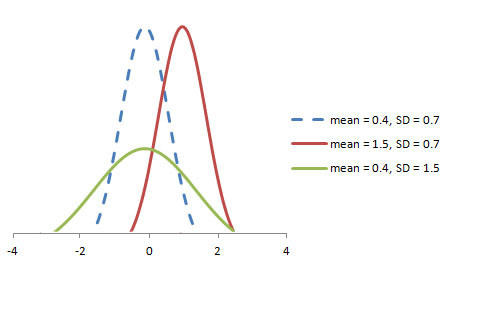
https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/09/16161923/image62.png

The mean and variance of a random variable X which is said to be normally distributed is given by:

Mean -> E(X) = µ

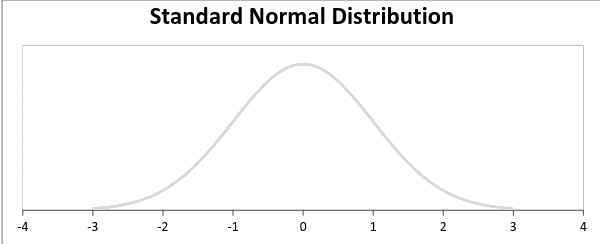
Variance -> Var(X) = σ^2

Here, µ (mean) and σ (standard deviation) are the parameters.  
The graph of a random variable X ~ N (µ, σ) is shown below.

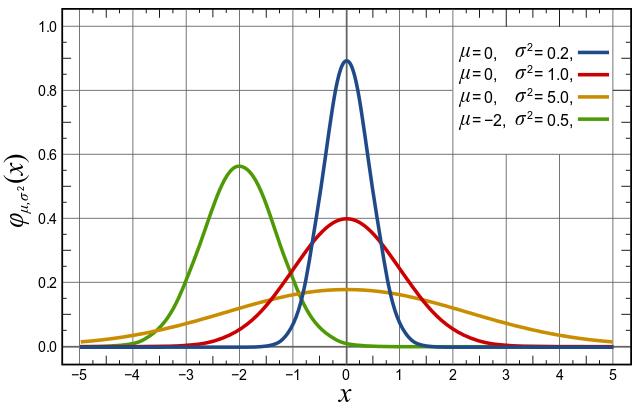


A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1.  For such a case, the PDF becomes:

https://s3-ap-south-1.amazonaws.com/av-blog-media/wp-content/uploads/2017/09/16162317/image63.png



Normal Distributions are the most common distributions in statistics primarily because they describe a lot of natural phenomena. Normal distributions are also known as ‘Gaussian distributions’ or ‘bell curve’, because of the bell shaped curve.



Samples of heights of people, size of things produced by machines, errors in measurements, blood pressure, marks in an examination, wages payed to employees by a company, life span of a species, all of these follows a normal or nearly normal distribution.

I don’t intend to cover a lot of mathematical background regarding normal distributions, still it won’t hurt to know just a few simple mathematical properties of normal distributions:

* Bell curve is symmetrical about mean(which lies at the center)
* mean = median = mode
* Only determining factors of normal distributions are its mean and standard deviation

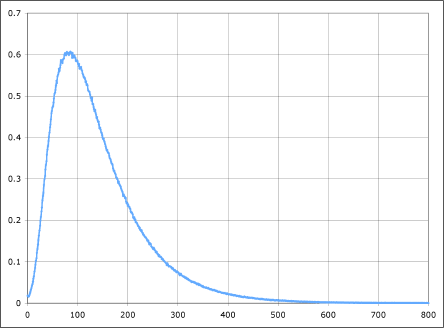
We can also get a normal distribution from a lot of datasets using [Central Limit Theorem](http://en.wikipedia.org/wiki/Central_limit_theorem)(CLT). In layman’s language CLT states that if we take a large number of samples from a population, multiple times and go on plotting these then it will result in a normal distribution(which can be used by a lot of statistical and machine learning models).

A lot of machine learning models assumes that data fed to these models follows a normal distribution. So, after you have got your data cleaned, you should definitely check what distribution it follows. Some of the machine learning and Statistical models which assumes a normally distributed input data are:

* Gaussian naive Bayes
* Least Squares based (regression)models
* LDA
* QDA

It is also quite common to transform non-normal data to normal form by applying log, square root or similar transormations.

If plotting the data results in a skewed plot, then it is probably a log-normal distribution(as shown in figure below), which you can transform into normal form, simply by applying a log function on all data points.



Once it is transformed into normal distributions, you are free to use this dataset with models assuming a normal input data(as listed in above section).

As a general approach, Always look at the statistical/probability distributions as your first step in data analysis.

why gausian is most favouritte?

Unlike many other distribution that changes their nature on transformation, a Gaussian tends to remain a Gaussian.

* Product of two Gaussian is a Gaussian
* Sum of two independent Gaussian random variables is a Gaussian
* Convolution of Gaussian with another Gaussian is a Gaussian
* Fourier transform of Gaussian is a Gaussian

For every Gaussian model approximation, there may exist a complex multi-parameter distribution that gives better approximation. But still Gaussian is preferred because it makes the math a lot simpler!

* Its mean, median and mode are all same
* The entire distribution can be specified using just two parameters- mean and variance

#### Measures to describe shape of distribution:

* Skewness–Skewness is a measure of the asymmetry. Negatively skewed curve has a long left tail and vice versa.
* Kurtosis–Kurtosis is a measure of the “peaked ness”. Distributions with higher peaks have positive kurtosis and vice-versa

