# Statistics 101:

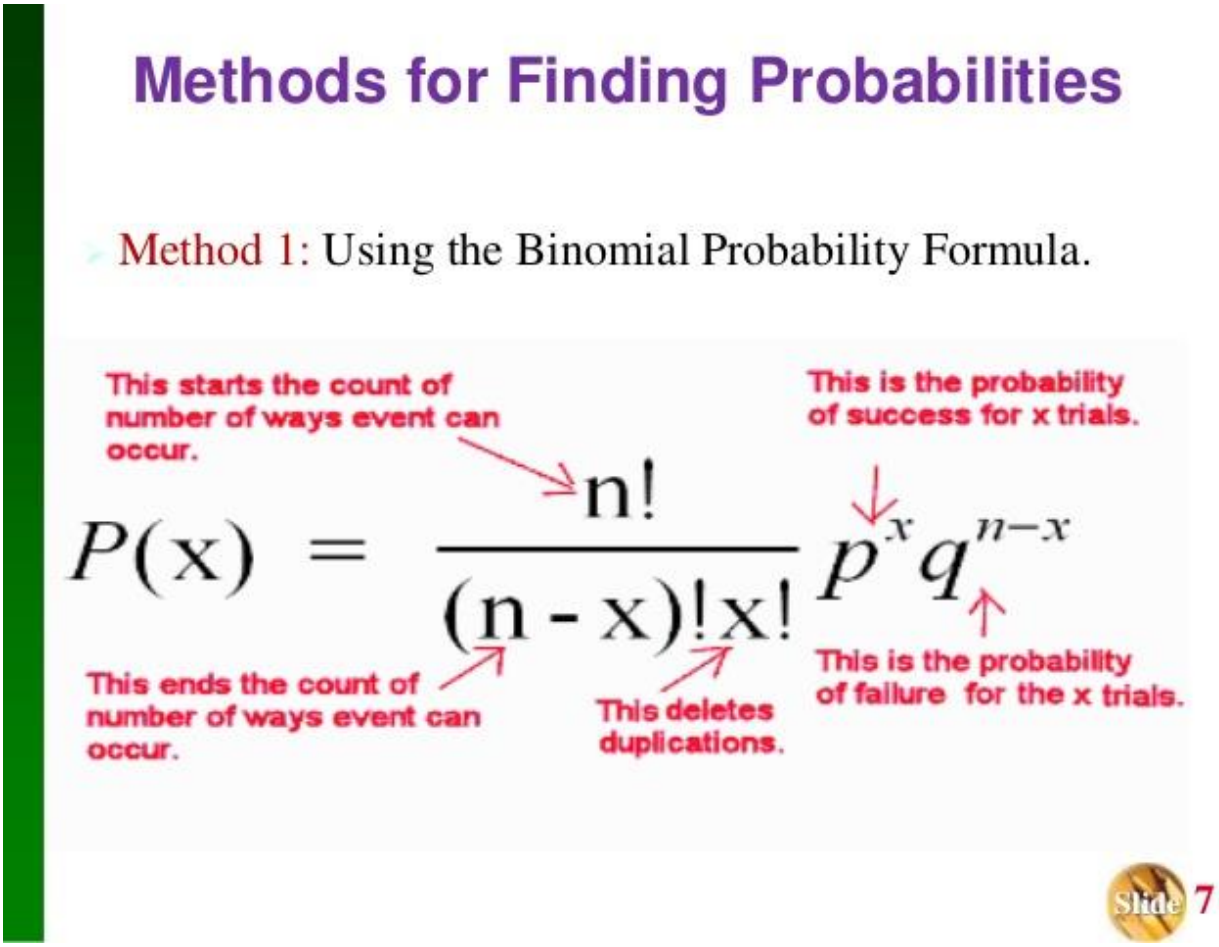
**Binomial Distribution:** Binomial distribution can be thought of as simply the probability of a **SUCCESS** or **FAILURE** outcome in an experiment or survey that is repeated multiple times. For example, *A coin is tossed 10 times. What is the probability of getting exactly 6 heads?*

A real world example. *A company drills 9 wild-cat oil exploration wells, each with an estimated probability of success of 0.1. All nine wells fail. What is the probability of that happening?*

```
Let's do 20,000 trials of the model, and count the number that generate zero positive
results.

sum(np.random.binomial(9, 0.1, 20000) == 0)/20000.

answer = 0.38885, or 38%
```

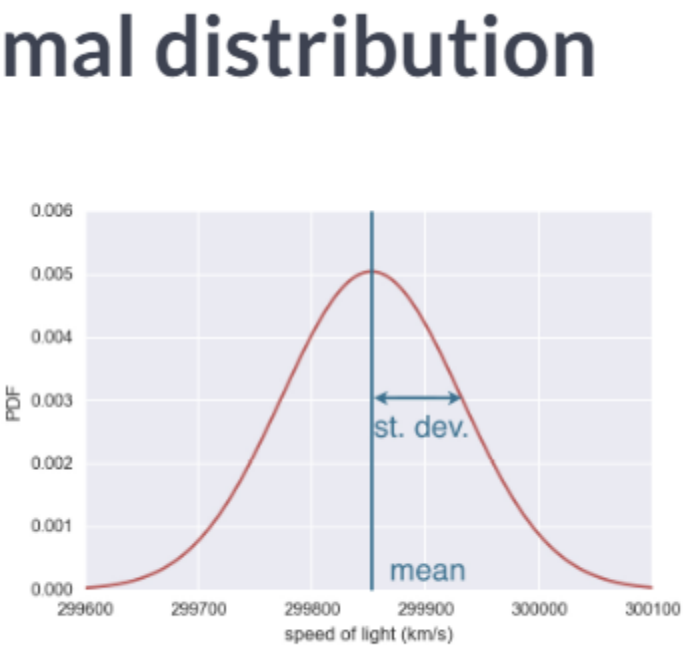


Where n is the number of event, x is the number of success and p is probability of success.

**Normal Distribution:** Calculating binomial distribution for large value of n can be a pain back in 18th century so James Bernoulli and Abraham De Moivre invented **Normal distribution.**

Normal distribution, also known as the Gaussian distribution, is a continuous probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graph form, normal distribution will appear as a bell curve. In simple terms**.**

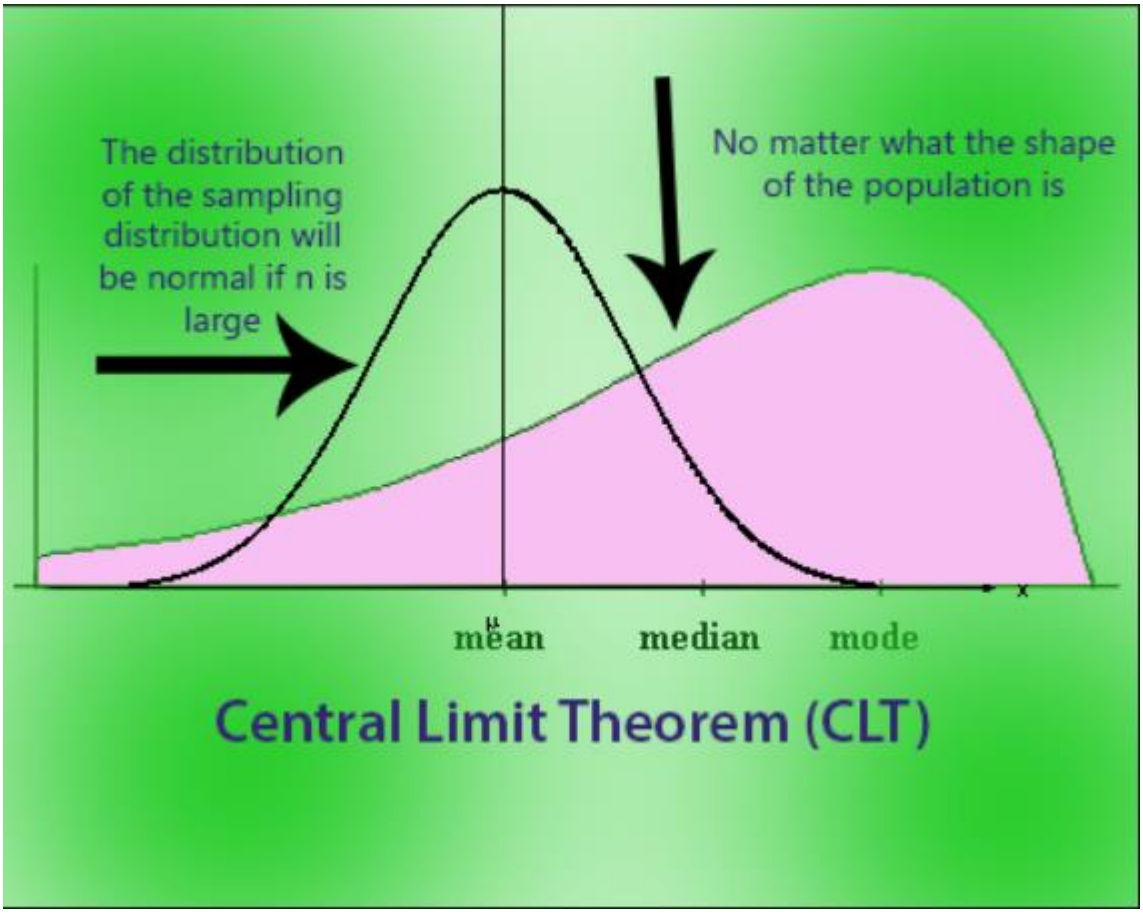
Normal Distribution is a probability distribution that is solely dependent on **mean** and **standard deviation.** Normal distribution parameterized by two parameters, **mean** describe where the center of the peak is and **standard deviation** describe how spread out data are.




```
mean = numpy.mean(speed_of_light)
std = numpy.std(speed_of_light)
samples = numpy.random.normal(mean, std, size=10000)
plot.hist(samples, bins=100, density = True, histtype = 'step')
```

**Central Limit Theorem:** *It just says that with a large sample size, sample means are normally distributed even* if the population distribution is not normal.

CLT has a core idea in stats that lets you use data to evaluate your ideas, even with incomplete information, hence it is one of the pillars in hypothesis testing, an important decision making statistics



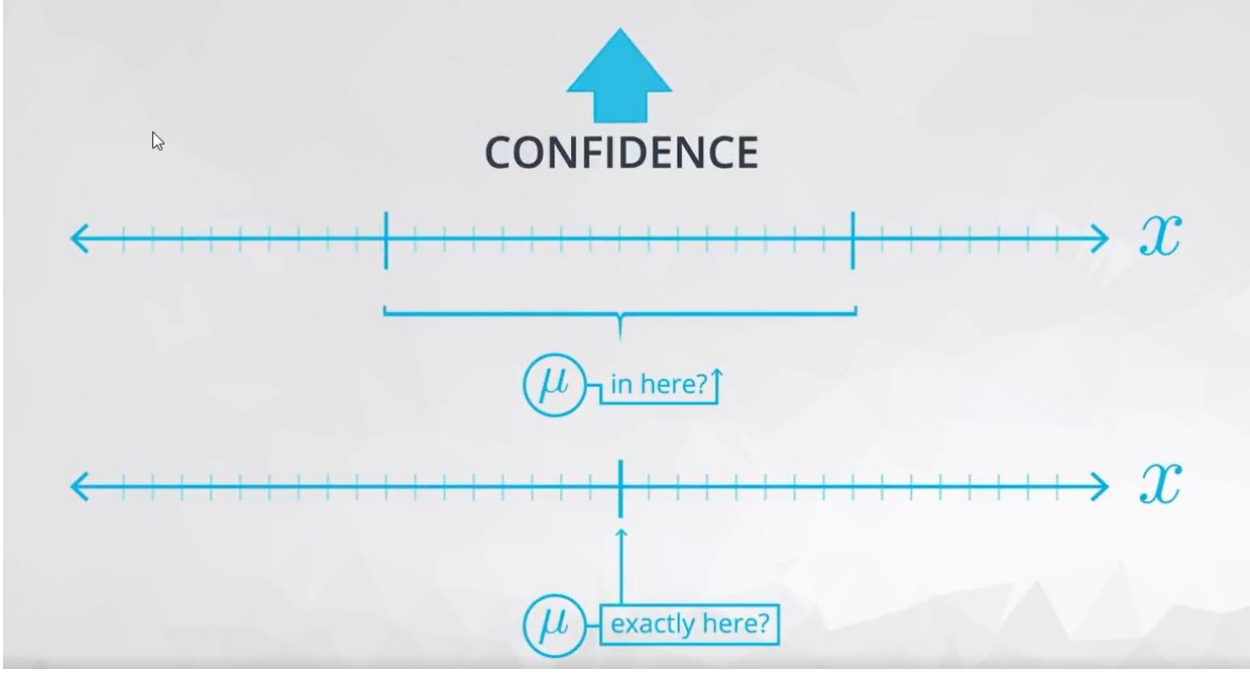
**Why Central Limit Theorem is important:** The field of statistics is based on fact that it is highly impossible to collect the data of entire population. Instead of doing that we can gather a subset of data from a population and use the statistics of that sample to draw conclusions about the population.

In practice, the unexpected appearance of normal distribution from a population distribution is skewed (even heavily skewed). Many practices in statistics such as hypothesis testing make this assumption that the population on which they work is normally distributed.

This assumption basically simplifies matters and we overcome the problem of the data belonging to the population which is not normal.

Thus, even we don't know the shape our distribution where our data comes from but according to Central Limit Theorem we can treat sampling distribution of any population as normal. Of-course, for the conclusions of the Central Limit Theorem to hold we need sample size to be large enough.

**Confidence Interval:** Let's assume that you are catching fish with hook, in this case the confidence is much lesser than catching fish with smaller net. The confidence will be much higher if you have larger net for catching fish. This is same idea behind confidence interval.




Confidence interval communicates how accurate our estimate is likely to be. Confidence intervals are usually reported with point estimates to show how reliable the estimates are.

*For example: we have to find the mean weight of all the apples in the orchard (population mean)?*

We take the sample and calculate the sample mean. we use confidence interval to express the range in which we are pretty sure the population mean will lies.

**It is reasonable to say that the mean weight of the apples in the orchard lies between 147g to 151g or we can say we are 95% confident the weight lies between 147g to 151g.**

**Hypothesis Testing:** During our analysis of the different datasets, we are often concerned with questions like whether males default more than females? Do self-driving cars crash more than normal cars? Does drug X help prevent/treat disease Y? To answer these questions, we can use another statistical technique known as **Hypothesis Testing**.

The aim of the hypothesis test is to determine whether the **null hypothesis** can be rejected or not. *The null hypothesis is a statement that assumes that nothing interesting is going on, or no relationship is present between two variables, or that there is no difference between a sample and a population*.

For instance, if we suspect that males default more than females, the null hypothesis would be that *males do not default more than females*. If there is little or no evidence against the null hypothesis, we accept the null hypothesis. Otherwise, we reject the null hypothesis in favor of the alternate hypothesis, which states that something interesting is going on, or there is a relationship between two variables, or that the sample is different from the population.

Common hypothesis tests include:

1. Testing a population mean (One sample t-test).
2. Testing the difference in means (Two sample t-test)
3. Testing the difference before and after some treatment on the same individual (Paired t-test)
4. Testing a population proportion (One sample z-test)
5. Testing the difference between population proportions (Two sample z-test)