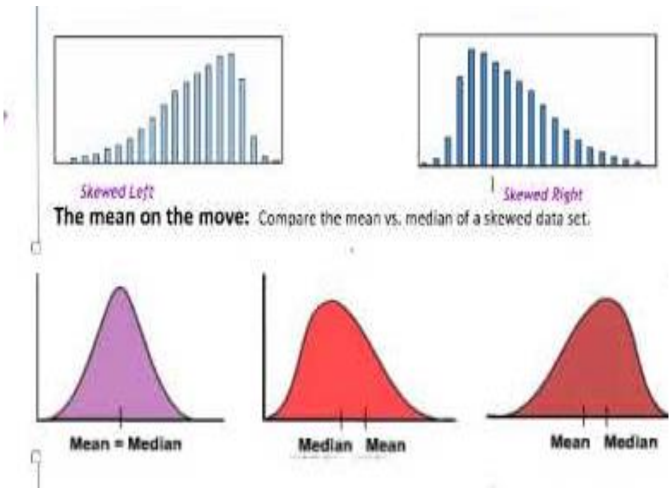
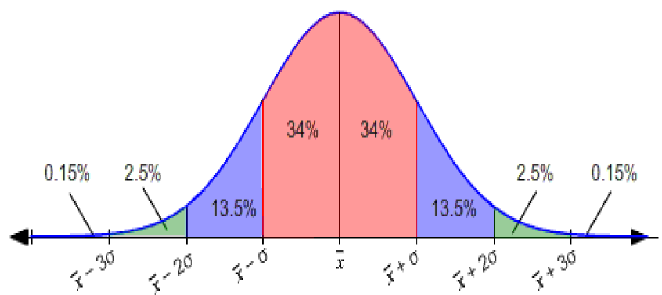
Data Distribution

Data Distribution is a function or a listing which shows all the possible values (or intervals) of the data. It also (and this is important) tells you how often each value occurs. Often, the data in a distribution will be ordered from smallest to largest, and graphs and charts allow you to easily see both the values and the frequency with which they appear.





Remember
50% of data below the mean
50% of data above the mean
Mean=median=model

Normal Distribution

It is a 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. Or data is symmetrically distributed with no skew.

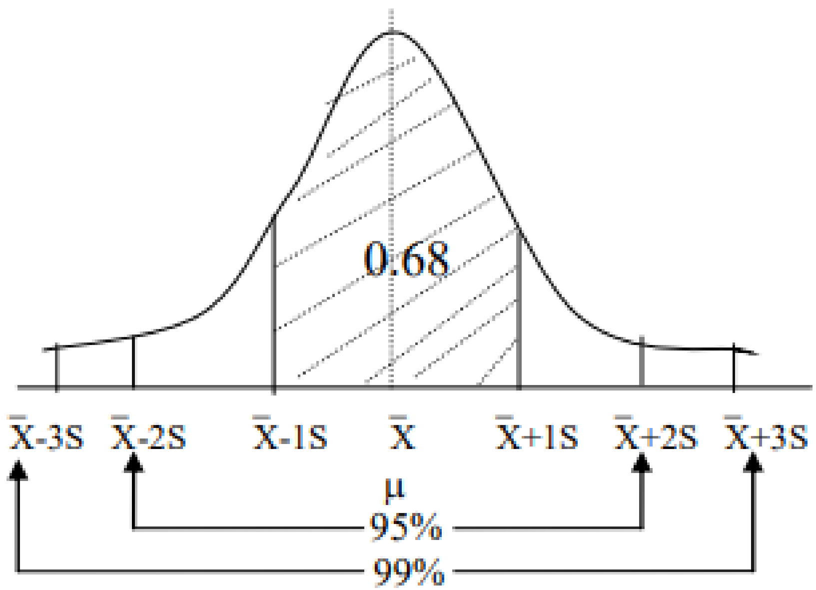
It is a graphical presentation and a most important continuous distribution (frequency polygon) of any Quantitative variables.

This distribution is important in the theory and practice of statistics,

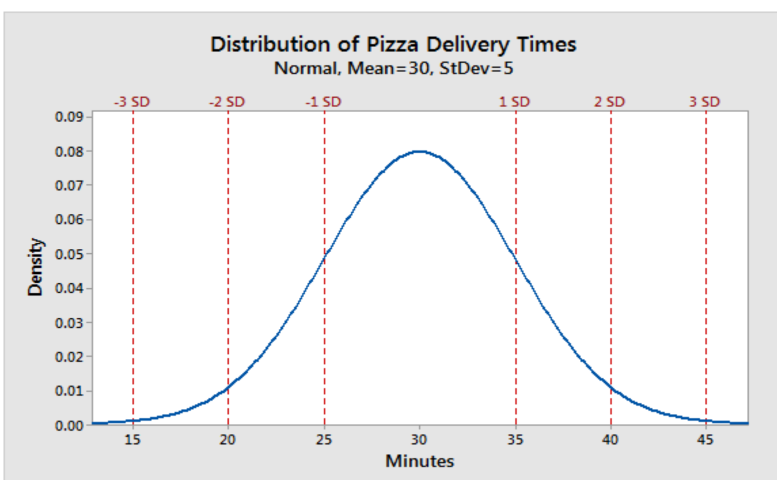
Properties of normal distribution:

- Bell-shaped
- Symmetric about mean
- Continuous

- Never touches the x-axis (The ends of curve are parallel to x-axis.)
- Total area under curve is 1.00
- Mean=median=model
- The area between two columns on points ($\bar{X}$-1S , $\bar{X}$+1S) = 0.68
- The area between two columns on points ($\bar{X}$-2S , $\bar{X}$+2S) = 0.95
- The area between two columns on points ($\bar{X}$-3S , $\bar{X}$+3S) = 0.99within 3 standard deviations of the mean.
- Data values represented by x which has mean mu and standard deviation sigma.



Let's look at this example. Assume that a pizza restaurant has a mean delivery time of 30 minutes and a standard deviation of 5 minutes.



we can determine that 68% of the delivery times are between 25-35 minutes (30 +/- 5), 95% are between 20-40 minutes (30 +/- 2*5), and 99.7% are between 15-45 minutes (30 +/-3*5). The chart below illustrates this property graphically.

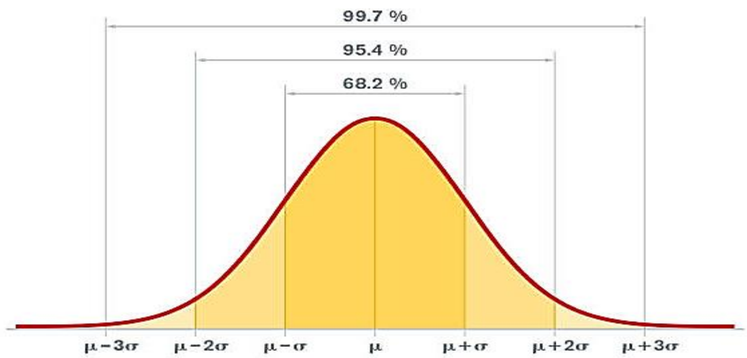
Example Assume a theoretical normal distribution of body weight of calves at age 6 months with average 200 kg and the variance is 400 kg^2, how you can draw the normal distribution of body weight?

Solution: $\bar{X}$= 200 kg, σ = $S = \sqrt{400}$ = 20 kg

$\bar{X}$-3S	$\bar{X}$-2S	$\bar{X}$-1S	$\bar{X}$	$\bar{X}$+1S	$\bar{X}$+2S	$\bar{X}$+3S
140	160	180	200	220	240	260

From the above distribution, we can conclude:

a. 50 % of calves have body weights less than 200 kg and 50 % of calves have weights higher than 200 kg.
b. 68 % of calves have body weight arranged between ($\bar{x}$-1S , $\bar{x}$+1S) = (200-1(20), 200 + 1(20) = 180 – 220 kg.
c. 95 % of calves have body weight arranged between ($\bar{x}$-2S , $\bar{x}$+2S) = (200 - 2(20), 200 + 2(20) = 160 – 240 kg.
d. 99 % of calves have body weight arranged between ($\bar{x}$-3S , $\bar{x}$+3S) = (200 - 3(20), 200 + 3(20) = 140 – 260 kg.



140 160 180 200 220 240 260

- **Standard Normal Distribution (Z)** The standard normal distribution is a special case of the normal distribution where **the mean = zero and the standard deviation = 1.** This distribution is also known as the Z-distribution.

Probability Function given by

$$Z = \frac{X-\mu}{\sigma}$$

Example Suppose we want to compare apples to oranges. Specifically, let's compare their weights. Imagine that we have an apple that weighs 110 grams and an orange that weighs 100 grams.

If we compare the raw values, it's easy to see that the apple weighs more than the orange. However, let's compare their **standard scores.** To do this, we'll need to know the properties of the weight distributions for apples and oranges. Assume that the weights of apples and oranges follow a normal distribution with the following parameter values:

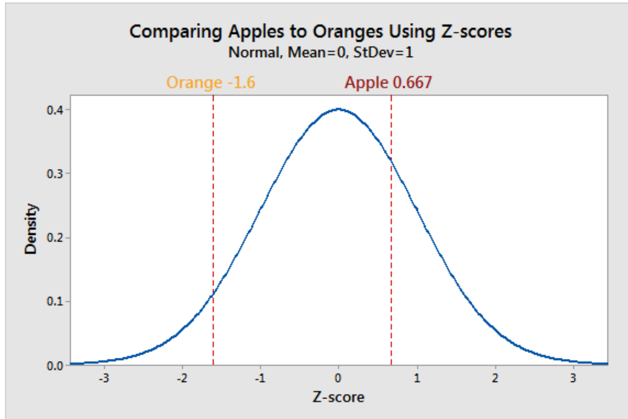
	Apples	**Oranges**
Mean weight grams	100	140
Standard Deviation	15	25

To standardize your data, you need to convert the raw measurements into Z-scores. calculate the standard score for an observation, the formula for that process is the following:

$$Z = \frac{X-\mu}{\sigma}$$

- Apple = (110-100) / 15 = 0.667
- Orange = (100-140) / 25 = -1.6

The Z-score for the apple (0.667) is positive, which means that our apple weighs more than the average apple. It's not an extreme value by any means, but it is above average for apples. On the other hand, the orange has fairly negative Z-score (-1.6). It's pretty far below the mean weight for oranges. Look at this Z-values in the standard normal distribution below.

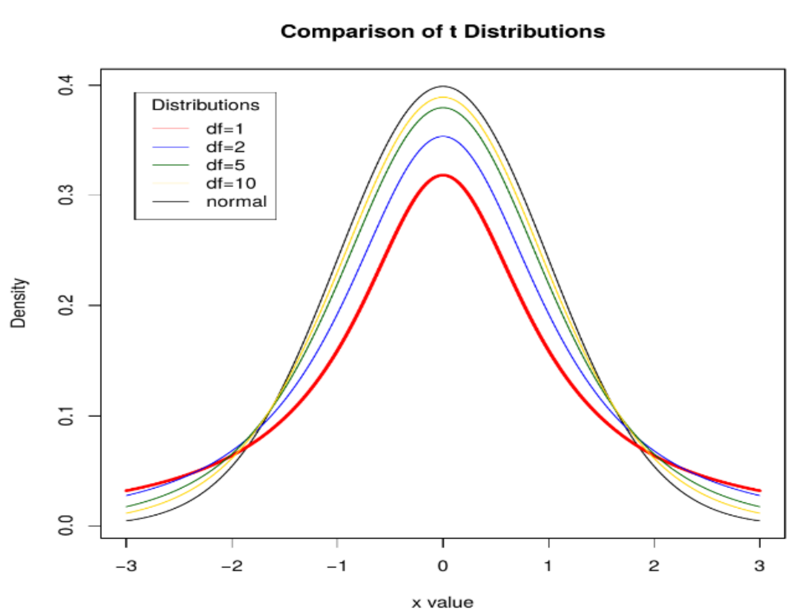


Student's t-Distribution (Normal t- Distribution)

The t-distribution describes the standardized distances of sample means to the population mean when the **population standard deviation is not known**, and the observations come from a normally distributed population.

This distribution used only for **small samples.**

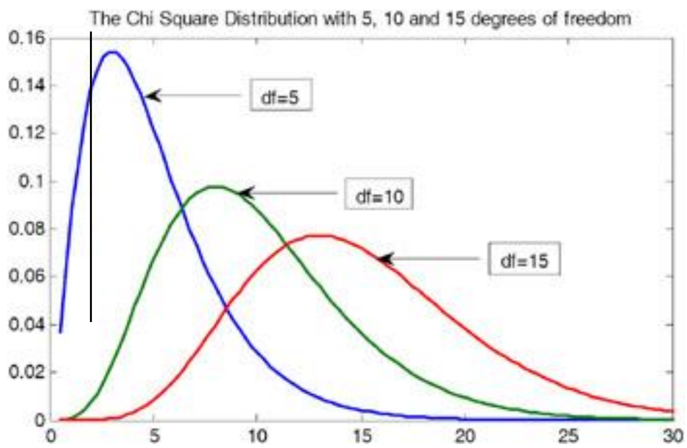
a. The standard normal or z-distribution assumes that you know the population standard deviation. The *t*-distribution is based on the sample standard deviation.
b. The *t*-distribution is defined by the ***degrees of freedom*** **(n-1),** related to the sample size.
c. The *t*-distribution is most useful for small sample sizes, when the population standard deviation is not known.



Chi-square Distribution (χ^2)

Chi-square distribution is a continuous distribution with k degrees of freedom.

- Independence of two criteria of classification of **qualitative variables.**
- used to test the **goodness of fit** of a distribution of data, whether data series are independent,
- Relationships between **categorical variables (contingency tables).**
- Tests of deviations of differences between expected and observed frequencies (one-way tables).
- The shape of χ^2 - distribution depends on degrees of freedom **(df =no. of groups – 1).**



- ## F- Distribution
- A continuous statistical distribution which arises in the testing of whether two observed samples have the same variance**.** It is the ratio of two variances:
- $F = \dfrac{Large\ variance}{Small\ variance}$
- The shape of F- Distribution depends on ***df*** of both variances and level of significant **(α=0.05 or 0.01).**
-