<https://academy.microsoft.com/en-us/dashboard>

Extract from csv or other sources, Format as table, data cleanup, formatting (dates, currency, etc.), scatter lines (inside a column), filtering data using either column header filters or by slicer, various BI charts (scatter, line, column, pie), statistics charts (histogram, box and whisker chart), pivot table (along with pivot charts) used to create multidimensional data summaries

Difference between population(total dataset) vs sample.

Mean: average (sum of all the observations divided by total number of observations). μ = ( Σ Xi ) / N

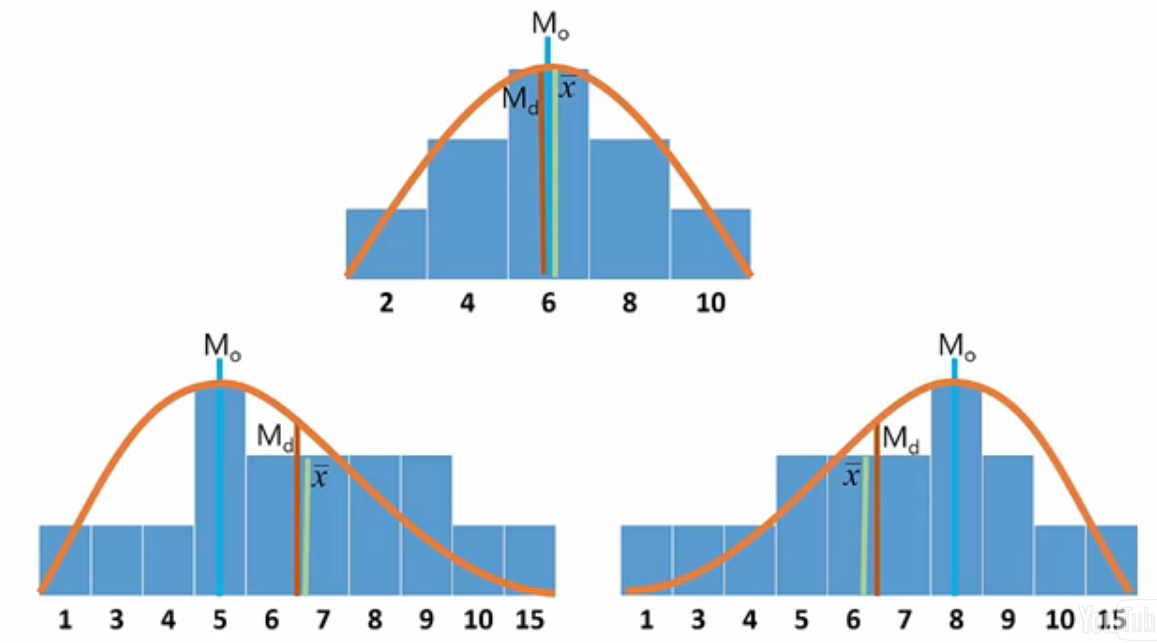
Median: middle value (after sorting the values). If the total number of observations is even, then we get 2 middle values. The average of those 2 values is median

Mode: most frequent (easily seen if you do a histogram)

A normal distribution is one in which mean, median and mode are same.

If mode and median are less than mean, then it is called right skewed distribution.

If mode and median are more than mean, then it is called left skewed distribution.



Range: max observation value – min observation value

Variance: sum of squares of each individual observation value subtracted from mean divided by number of observations.

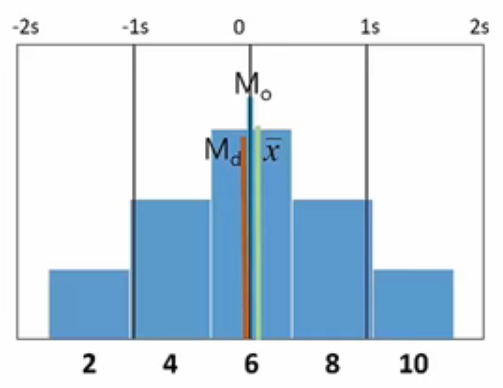
Variance for population: σ2 = Σ ( Xi - μ )2 / N where N is the population size

Variance for population: [s2 = Σ ( xi - x )2 / ( n - 1 )](http://stattrek.com/AP-Statistics-1/Variability.aspx?Tutorial=stat) where n is the sample size and x is the sample mean

Here we are dividing my n-1 to correct some bias that might be inherent in the sample.

Standard deviation: σ = sqrt(σ2) and s = sqrt( [s2](http://stattrek.com/AP-Statistics-1/Variability.aspx?Tutorial=stat) )

For normal distribution, the percentage of data that falls between +1 and -1 of standard deviation is 68.2% , that falls between +2 and -2 is 95.4% and that between +3 and -3 is 99.7%



Standard Error: SE = s/sqrt(n) and that is because we are dealing with samples and not the complete dataset

Descriptive statistics: use the data analysis add-in and select the columns you want stats on. Kurtosis is an estimate of normality of the data. The closer it is to 0, the more likely the data is normally distributed. Positive Skewness means data is right skewed and negative skewness means data is left skewed

Correlation: +1 to -1. Again use the data analysis add-in. Pick correlation. Highlight all the columns in which you are interested in finding a correlation. Correlation is not causation.

Hypothesis testing: null hypothesis, alternate hypothesis. Suppose rosie sells more lemonade in august and less in September. So we are trying to find some answers here as to why that happened. The question is there a correlation between months and the amount of lemonade sold. That is the hypothesis we are testing. The null hypothesis for this case would, no, the months do not make a difference on the lemonade sold (the data we are seeing is just due to change or randomness). Alternate hypothesis tells that:1) yes correlation between months and lemonade exists and actually the difference in the amount of lemonade sold should have been greater 2) or lesser 3) or we do not know what direction the difference just that there should be difference.

Another example:

Effect of temperature on evapotranspiration: So we are trying to hypothesize that there is a correlation between temp and evapo. The null hypothesis says there is no correlation while alternate hypothesis says the correlation could be in 3 different ways 1) the effect of temp should have been greater than what we see (from the observed data) or 2) the effect of temp should have been less than what we see (from the observed data) or 3) the effect of temp should have been greater or less than what we see (from the observed data) just that we are not sure.

Another example: effect of drug on a medical condition.

And how much risk are you willing to assume that you could be wrong is called the p-value in you hypothesis. Most statisticians would say they are comfortable with p=.05 or 5% risk (that they are wrong and the result that see is purely due to chance or randomness).

One-sample t-test or z-test: comparing data from 2 different samples. Can do it in a number of different ways. Essentially 4 different statistics that allow comparisons of different groups of data. The simplest statistic is called the 1 sample t-test or a z-test.

So say we know the sales data from last 5 years. And we want to know if the sales that she got this year are different from what she’s seen historically. Over the last 5 years, she’s sold on average 120 glasses in the month of july. Take mean from you current sample and subtract it from mean got form previous samples and divide it by the standard error which essentially is the standard deviation of the sample means. Having got the t statistic, we are going to apply the t distribution to determine if that T statistic is statistically significant. Something new here is the df which is the degrees of freedom (how many numbers of items in the data can be changed and you can still end up the same mean) that is used to determine if you value is statistically significant.



df = n-1

Math is exactly the same for z-test and one sample t-test. Z-test is built into excel. Now coming back to the example, lets say the average sales for last 5 years is 200. So we want to know if she has sold more this year that she has historically.