<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

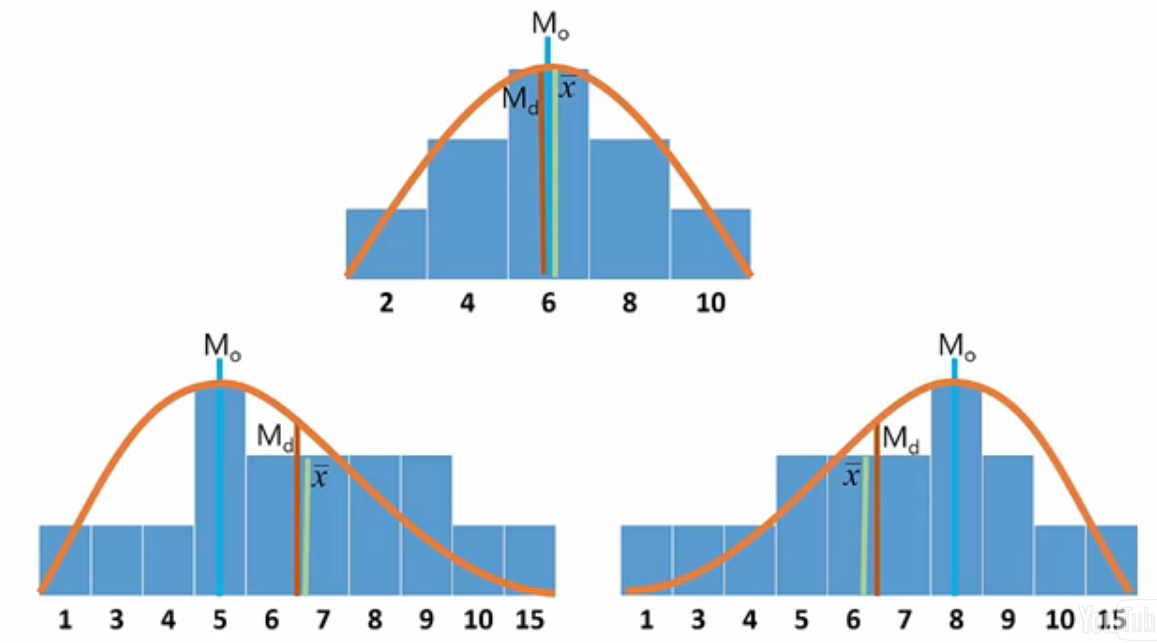
Median: middle value (after sorting the values). If the total number of observations is even, then we get 2 middle values. The average of those 2 values is median

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.

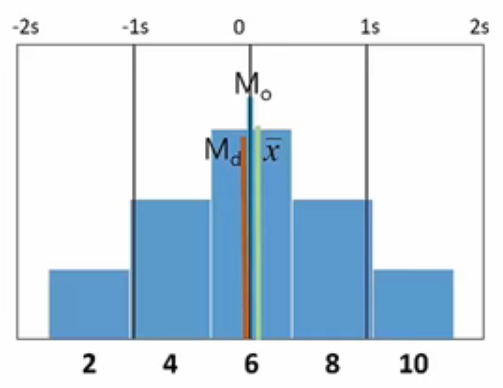
Variance for population: σ2 = Σ ( Xi - μ )2 / N where N is the population size

Variance for sample: [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). P-value measures your null hypothesis. It should be very small for that predictor/feature to have correlation with the response/label/output.

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.

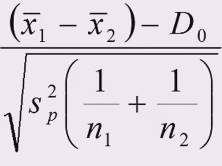
So say we know the sales data from last 5 years. And we want to know if the sales that she got this year are different from what she’s seen historically. Over the last 5 years, she’s sold on average 120 glasses in the month of july. Take mean from you current sample and subtract it from mean got form 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.

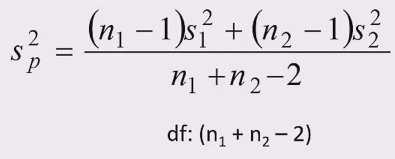


df = n-1

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. In excel use z.test and select all the values in the column ‘Sales’(that is all the values for month of july I think) and then we want the value in this analysis to be compared to which in this case is 200 (is 200 the average for month of july over last 5 years or average total sales in last 5 years) and we get the result of 0.638251553. Now this number is the probability of the t results or z result that was obtained from this test. Most statisticians use a p-value value of less .05 as the criteria with which they will look at a statistic as being significant. So, p-value is essentially the probability of being wrong. If I say something is statistically significant, then what is the probability that I might be wrong. Now in this case, .68 is clearly more than .05 meaning. What does that mean. If the number was, say, .02 which is less than .05 then that means our result is statistically significant and she in fact has more sales than she had previous years.

2-sample t-test: where we can compare different variable within the same dataset. So you want to know that are you selling more lemonade or more orangeade. To do that we use 2 sample t-test. Essentially what the formula is doing is that it is comparing the mean of your orangeade sales and the mean of your lemonade sales and it is diving it by the pooled standard deviation.





In excel, use data analysis pack. You will see that there are several different options for 2 sample t-tests. For most part, you will use the equal variances version. But if you suspect that your 2 sample have very-very different variances, then in that case use the unequal variances version. Now if have proposed or hypothesized that for example that she is selling more lemonade than orangeade, then you have what is called a 1 tailed test. If you weren’t sure about what she selling more (lemonade or orangeade), then you actually have a 2-tailed test.

Paired sample t-test: more stats..ugh. So does posting flyers and handing out leaflets make a difference in sales.

Analysis of variance (ANOVA): she sell or’ade and lem’ade at both beach and the park. So is there a difference where sells her product. Again, can be done in excel.

Regression: for a lot of people the goal of data science is the prediction. So we have some variables or features like flyers, price and temperature. So can we use to predict sales. So we are essentially creating a function that has these variables as input and gives us sales as output. We already have some data, so using regression we are finding the function that will best fit the data.

Again in excel, use data analysis and select regression. The y(output being total sales), select the column. For x(input being temperature, leaflets and price variable), select the corresponding columns.

Excel stats actual course:

Histogram is a fancy bar chart. Histogram shows bars for range of numbers (buckets/bin) rather than single values shown by bar charts. Histograms are used to display continuous data whereas bar charts are used for discrete data. Histograms are valuable tools as they help us understand the shape or distribution of our data. Normal distribution of data (not skewed towards one end) is a favorable for many stats analyses.

The skewness measure reported by the descriptive statistics output indicates whether a data set is highly skewed, in the following ways:

1. A skew greater than +1 indicates a high degree of positive skew.
2. A skew less than 1 indicates a high degree of negative skew.
3. A skew between 1 and +1 inclusive indicates a relatively symmetric data set.

Assuming that a histogram follows a Gaussian, or normal population, the rule of thumb (set of related math rules) tells us the following:

1. Approximately 68 percent of all observations are between

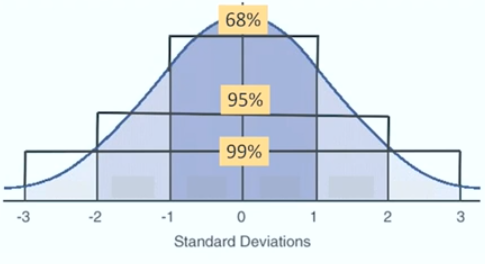
x–s and x+s.

1. Approximately 95 percent of all observations are between

x–2s and x+2s.

1. Approximately 99.7 percent of all observations are between

x–3s and x+3s.



Box plots are also good visual tools for distribution, percentiles as well as outliers(good for comparative stats as well). Boxplot show the bold horizontal line is the median value of the data, the box shows the range of values between the first quartile and third quartile, and the whiskers (the dotted lines extending outside the box) show the minimum and maximum values, excluding any outliers (which are plotted as circles). Outliers are defined by first computing the difference between the first and third quartile values, or the height of the box. This number is called the Inter-Quartile Range (IQR). Any point that is greater than the third quartile plus the IQR or less than the first quartile minus the IQR is considered an outlier.