**ANOVA**

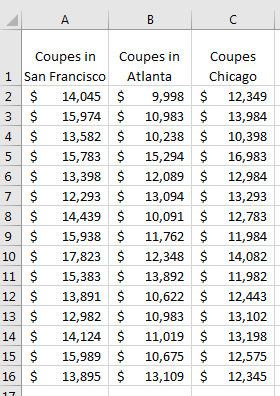
Analysis of Variance (ANOVA) is used for testing statistical differences among 3 or more means (groups). This document discusses both one-factor ANOVA and two-factor ANOVA.

The following focuses on the use of Excel for carrying out an ANOVA, while the accompanying video shows how to run ANOVA using MegaStat. An advantage of MegaStat is that it besides providing an overall ANOVA analysis, MegaStat also teases out whether specific pairs of groups (or conditions) are significantly different from each other.

One-Factor ANOVA using Excel

A one-factor ANOVA involves a single factor or independent variable (e.g., age group or region or semester, etc.) with 3 or more levels (e.g., for Region, the levels could be North, South, East, and West).

Let’s look at the example from Module 3.10 in the course material. Below are prices of a particular car (Coupe) model from dealerships in each of 3 different cities. A sample of 15 car prices were obtained for each city.

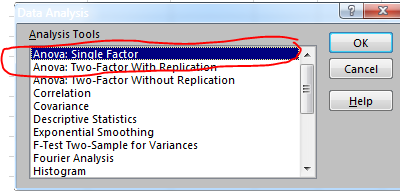


In this case what is the null hypothesis? The null hypothesis is that the average car price is not statistically different for the 3 cities (i.e., µSan Francisco = µAtlanta = µChicago). If the results of the ANOVA turn out to be statistically significant (i.e., the observed F-value is above the critical F-value and p < .05) then we will reject the null hypothesis and conclude that the average car prices are not all equal across the 3 cities.

A one-factor ANOVA can be performed with Excel if you have the Analysis ToolPak add-in included (please refer to this [link](https://srm--c.na60.visual.force.com/apex/coursearticle?Id=kA0a000000158l9CAA) if you need to install the Data Analysis option).

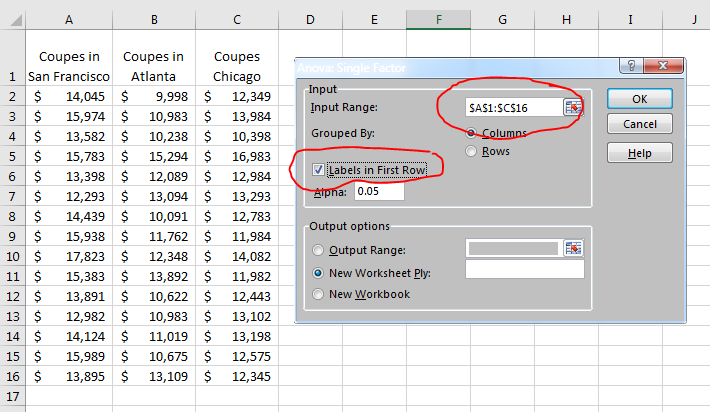
Within Excel 🡪 Data 🡪 Data Analysis

…within the pop-up menu, you should get something like the following….



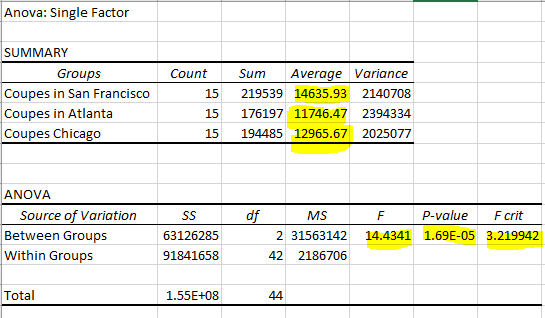
We select “ANOVA Single Factor” and then click “OK”.

The subsequent pop-up box should be filled out as shown below given this example…



*Note: Because the first row in the Input Range contains the column labels (i.e, the city names), we check the “Labels in First Row” box.*

Then we click “OK” to get the ANOVA results…



Above, I’ve highlighted the key information, including the mean car prices for each of the 3 cities, and the results of the statistical ANOVA test. This type of output should be included in your Task 2.

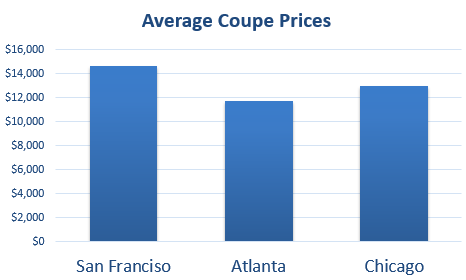
For the ANOVA test, notice that the obtained F-value of 14.43 is larger than the critical F-value of 3.22. This means we can reject the null hypothesis that average prices across the 3 cities are all the same. Also note that the P-value is 1.69E-05. The “E-05” is Excel’s way of stating that the decimal place should be moved 5 places to the left, resulting in the p-value 0.0000169. Clearly, this p-value is less than 0.05 (the standard to measure against), so we reject the null hypothesis (note: because the obtained F-value is greater than the critical F-value, it is necessarily the case that the p-value is less than .05; for statistical significance you want a large F-value and a small p-value).

An appropriate interpretation and write-up of these results would be:

We can reject the null hypothesis that average coupe prices are the same for the three cities because the obtained F-value of 14.43 is greater than the critical F-value of 3.21, and because p<.05. San Francisco has the highest mean price ($14,636), Atlanta the lowest ($11,746), and Chicago is in between these ($12,966). While we can conclude that the mean prices are not the same across the three cities, the ANOVA itself does not specify which means are significantly different from each other.

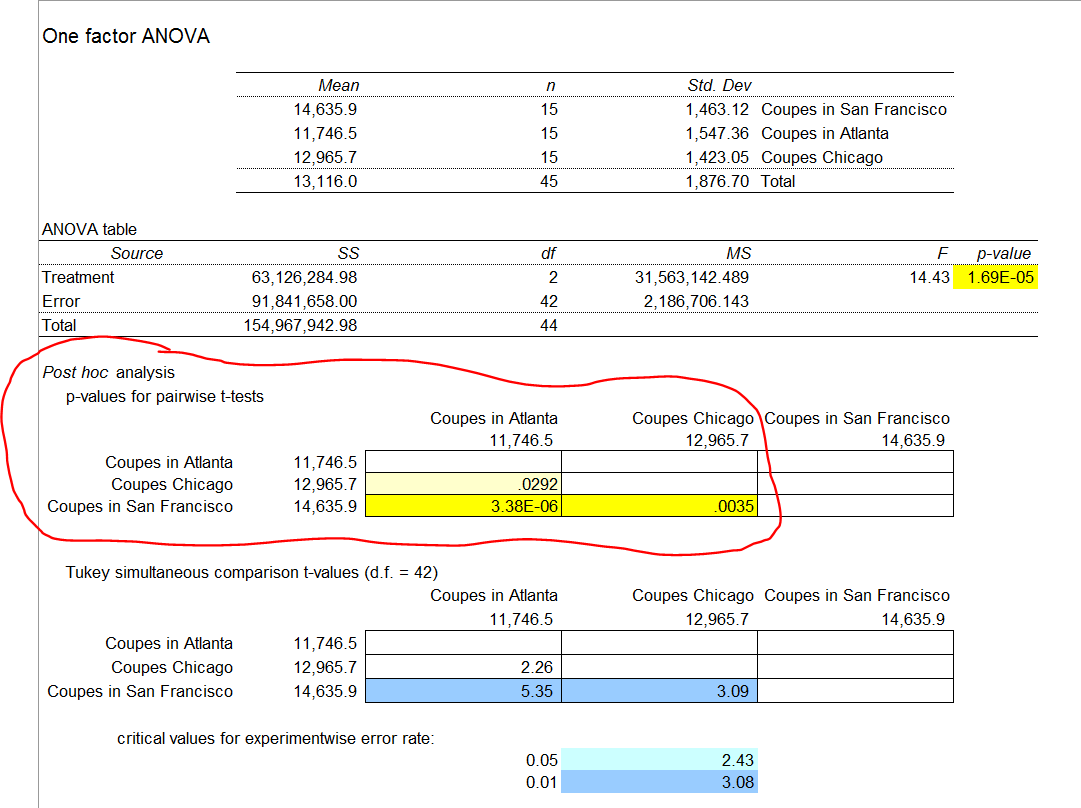
\Again, a limitation of the ANOVA is that it doesn’t indicate which cities are significantly different from each other. This requires additional statistical testing (such as a series of t-tests or other post-hoc test). Excel will only carry out the overall ANOVA, while MegaStat (shown on the next page) will also conduct these post-hoc tests. Note that for Task 2, you are not required to conduct such post-hoc tests.

An appropriate visual chart in this case would be something like the following, showing the average prices for each of the 3 cities.

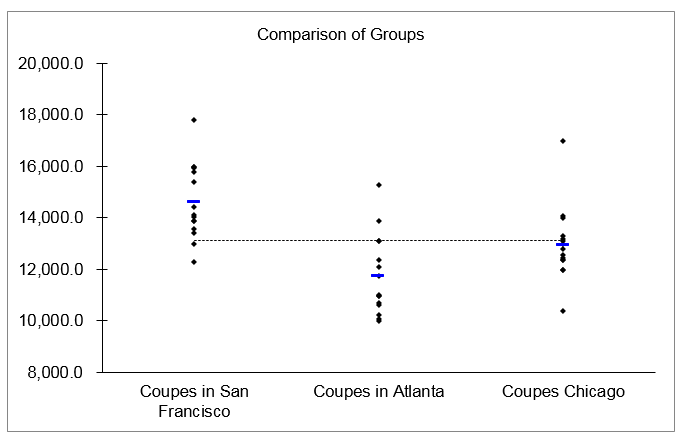


One-Factor ANOVA using **MegaStat**

Below is Output, for the same data, using MegaStat (see [video](http://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=228ef731-f0f6-41ca-bf1d-2856fca8bf1b) for instructions and this [tip sheet](https://srm--c.na60.visual.force.com/apex/coursearticle?Id=kA0a000000158xZCAQ) to install MegaStat). Note the means, F-value and p-value are the same as was obtained using Excel. In addition, MegaStat conducts pairwise t-tests (post-hoc tests) of all possible pairs of the 3 cities. For all 3 pairwise comparisons, the p-value is less than 0.05 (see circled section below), indicting significant differences between every pair. So we can conclude that San Francisco has the highest coupe prices which are significantly higher than that of Chicago, and Chicago’s prices are, in turn, significantly higher than that of Atlanta.



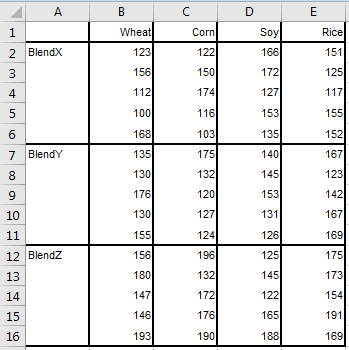
MegaStat automatically produces a graph (that you can include with your Task 2 write-up).



Two-Factor ANOVA using Excel

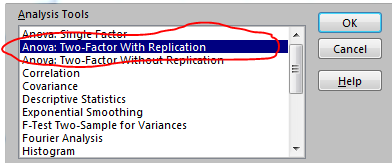
A two-factor ANOVA involves two factors or dimensions.

The example below shows crop yields (the dependent variable) for two different independent variables: type of crop and fertilizer blend. Note that there are 4 types (levels) of crop and 3 types (levels) of fertilizer blend. So, this is a 4x3 design. Also note that there are 5 observations within each cell or crossing of the 2 factors.



To analyze this within Excel…

Data 🡪 Data Analysis….

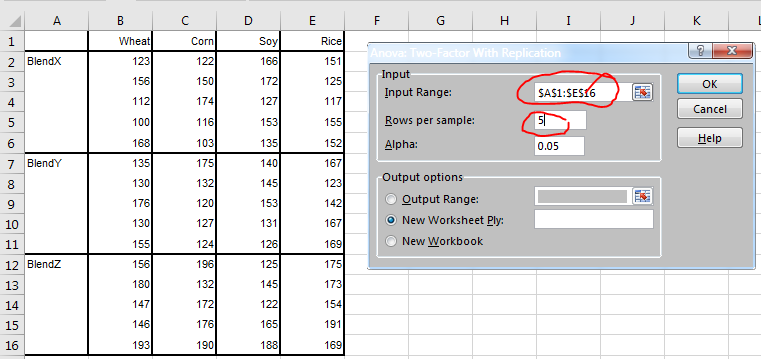


…and select “ANOVA: Two-Factor With Replication”, given that there are 2 Factors and more than one observation per cell.

For this option, Excel assumes that each cell has exactly the same number of observations (5 in this case).

Below is how the pop-up box is filled in for this example. Note that…

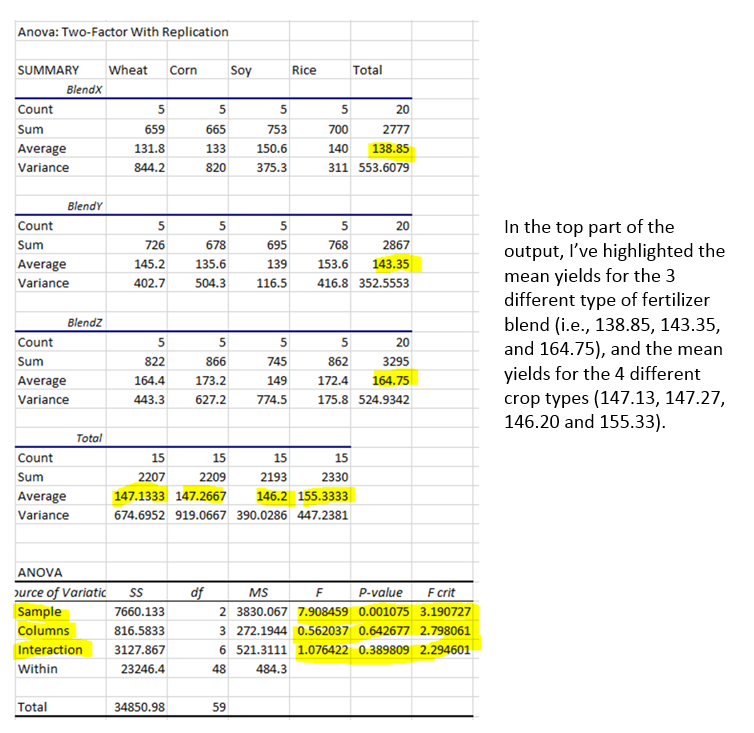
* The first row (row 1) must contain the labels for the levels of the first factor.
* The first column (column A) must contain the labels for the levels of the 2nd factor.
* The “Input Range” includes the labels.
* The number of observations per cell (i.e., “rows per sample”) is specified.



On the next page are the results of the 2-Factor ANOVA.

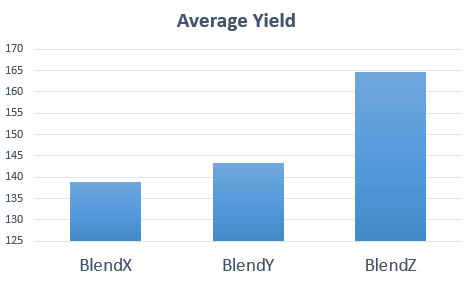
Not all of the labels of the output are particularly clear. For the ANOVA output (which you would include in your Task 2 report) …

* “Sample” refers to the Fertilizer Blend (row) dimension
  + Note that the obtained F-value of 7.91 is greater than the critical F-value of 3.19, and the p-value is less than .05. This indicates that the yields for the 3 fertilizer blends are not all the same (i.e., there are significant differences in yield among the types of fertilizer blend).
* “Columns” refers to the type of Crop dimension
  + Here, the obtained F-value of 0.56 is less than the critical F-value of 2.80 (and the p-value of 0.64 is greater than 0.05). This indicates that there are no significant differences in yields across the 4 types of crops.
* “Interaction” refers to the crossing of fertilizer blend by crop type (i.e., cell means).
  + Because the interaction is not significant, this indicates that there is no significant relationship (pattern) across fertilizer blend by crop type.

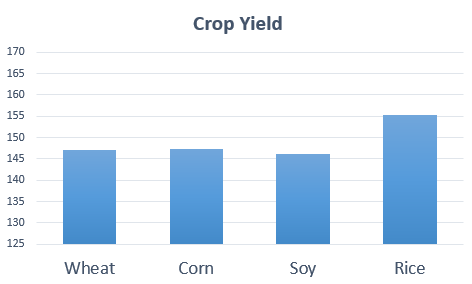


Because there is a significant difference among fertilizer blend, but not among crop type, the first chart below would be appropriate to include in Task 2.

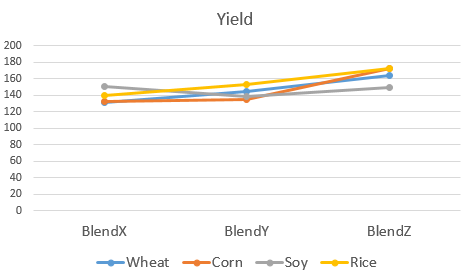
The means below suggest that yield is higher for BlendZ than for the other two fertilizer blends. However, a limitation of ANOVA is that we can only conclude that the yield are not all the same for the 3 blends. Subsequent statistical tests (such as a series of t-tests or other post-hoc tests) are needed to make definitive conclusions about which means are significantly different from each other (note: you are not expected to conduct such follow up analyses).



For completeness, below is a bar chart of the average yield for the 4 different crop types. Recall, the ANOVA determined that there are no significant differences in yield for the 4 crop types.



Below is a plot of the interaction of Crop type by Blend type. Recall, the ANOVA indicated no significant interaction (i.e., the pattern of results is approximately the same for each crop type).



Two-Factor ANOVA using **MegaStat**

Below is Output, for the same data, using MegaStat. Note that the means, F-values and p-values are the same as was obtained using Excel. In addition, MegaStat conducts pairwise t-tests (post-hoc tests) of all possible pairs of the first variable (Fertilizer Blend Type, shown as Factor 1 below) because this is the only variable showing statistical significance (p<.05).

Notice that the post-hoc analyses reveal that crop yield values for BlendZ are significantly higher than that of values of BlendX or BlendY (i.e., the p-values of these pairwise comparisons are less than 0.05; see the circled section below). In contrast, there is no significant difference in crop yields for BlendX versus BlendY because the p-value is greater than 0.05 (p=.521). Refer to the bar chart on page 8.

