13.12: Factor analysis and ANOVA

Authors: Alexander Voice, Andrew Wilkins, Rohan Parambi, Ibrahim Oraiqat
Stewards: Irene Brockman, Chloe Sweet, Rob Chockley, Scott Dombrowski

Introduction

First invented in the early 1900s by psychologist Charles Spearman, factor analysis is the process by which a complicated system of many variables is simplified by completely defining it with a smaller number of “factors.” If these factors can be studied and determined, they can be used to predict the value of the variables in a system. A simple example would be using a person’s intelligence (a factor) to predict their verbal, quantitative, writing, and analytical scores on the GRE (variables).

Analysis of variance (ANOVA) is the method used to compare continuous measurements to determine if the measurements are sampled from the same or different distributions. It is an analytical tool used to determine the significance of factors on measurements by looking at the relationship between a quantitative “response variable” and a proposed explanatory “factor.” This method is similar to the process of comparing the statistical difference between two samples, in that it invokes the concept of hypothesis testing. Instead of comparing two samples, however, a variable is correlated with one or more explanatory factors, typically using the F-statistic. From this F-statistic, the P-value can be calculated to see if the difference is significant. For example, if the P-value is low (P-value<0.05 or P-value<0.01 - this depends on desired level of significance), then there is a low probability that the two groups are the same. The method is highly versatile in that it can be used to analyze complicated systems, with numerous variables and factors. In this article, we will discuss the computation involved in Single-Factor, Two-Factor: Without Replicates, and Two-Factor: With Replicates ANOVA. Below, is a brief overview of the different types of ANOVA and some examples of when they can be applied.
Overview and Examples of ANOVA Types

ANOVA Types

**Single-Factor ANOVA (One-Way):**

One-way ANOVA is used to test for variance among two or more independent groups of data, in the instance that the variance depends on a single factor. It is most often employed when there are at least three groups of data, otherwise a t-test would be a sufficient statistical analysis.

**Two-Factor ANOVA (Two-Way):**

Two-way ANOVA is used in the instance that the variance depends on two factors. There are two cases in which two-way ANOVA can be employed:

- Data *without replicates*: used when collecting a single data point for a specified condition
- Data *with replicates*: used when collecting multiple data points for a specified condition (the number of replicates must be specified and must be the same among data groups)

When to Use Each ANOVA Type

- Example: There are three identical reactors (R1, R2, R3) that generate the same product.
- One-way ANOVA: You want to analyze the variance of the product yield as a function of the reactor number.
- Two-way ANOVA without replicates: You want to analyze the variance of the product yield as a function of the reactor number and the catalyst concentration.
- Two-way ANOVA with replicates: For each catalyst concentration, triplicate data were taken. You want to analyze the variance of the product yield as a function of the reactor number and the catalyst concentration.

ANOVA is a Linear Model

Though ANOVA will tell you if factors are significantly different, it will do so according to a linear model. ANOVA will always assume a linear model, it is important to consider strong nonlinear interactions that ANOVA may not incorporate when determining significance. ANOVA works by assuming each observation as overall mean + mean effect + noise. If there are non-linear relationship between these (for example, if the difference between column 1 and column 2 on the same row is that column2 = column1^2), then there is the chance that ANOVA will not catch it.

**Key Terms**

Before further explanation, please review the terms below, which are used throughout this Wiki.
Comparison of Sample Means Using the F-Test

The F-Test is the ratio of the sample variances. The F-statistic and the corresponding F-Test are used in single-factor ANOVA for purposes of hypothesis testing.

Null hypothesis ($H_0$): all sample means arising from different factors are equal

Alternative hypothesis ($H_a$): the sample means are not all equal

Several assumptions are necessary to use the F-test:

- The samples are independent and random
- The distribution of the response variable is a normal curve within each population
- The different populations may have different means
- All populations have the same standard deviation

Introduction to the F-Statistic

The F-statistic is the ratio of two variance estimates: the variance between groups divided by the variance within groups. The larger the F-statistic, the more likely it is that the difference between samples is due to the factor being tested, and not just the natural variation within a group. A standardized table can be used to find $F_{critical}$ for any system. $F_{critical}$ will depend on alpha, which is a measure of the confidence level. Typically, a value of alpha = 0.05 is used, which corresponds to 95% confidence. If $F_{observed} > F_{critical}$, we conclude with 95% confidence that the null hypothesis is false. For an explanation of how to read an F-Table, see Interpreting the F-statistic (below). In a similar manner, F tables can also be used to determine the p-value for a given data set. The p-value for a given data set is the probability that you could obtain this data set if the null hypothesis were true: that is, if the results were strictly due to chance. When $H_0$ is true, the F-statistic has an F distribution.

F-Distributions

The F-distribution is important to ANOVA, because it is used to find the p-value for an ANOVA F-test. The F-distribution arises from the ratio of two Chi squared distributions. Thus, this family has a numerator and denominator degrees of freedom.
freedom. (For information on the Chi squared test, click here.) Every function of this family has a skewed distribution and minimum value of zero.

Figure 1 - F distribution with alpha and $F_{\text{critical}}$ indicated

**Single-Factor Analysis of Variance**

In the case of single-factor analysis, also called single classification or one-way, a factor is varied while observing the result on the set of dependent variables. These dependent variables belong to a specific related set of values and hence, the results are expected to be related.

This section will describe some of the computational details for the F-statistic in one-way ANOVA. Although these equations provide insight into the concept of analysis of variance and how the F-test is constructed, it is not necessary to learn formulas or to do this analysis by hand. In practice, computers are always used to do one-way ANOVA

**Setting up an Analysis of Variance Table**

The fundamental concept in one-way analysis of variance is that the variation among data points in all samples can be divided into two categories: variation between group means and variation between data points in a group. The theory for analysis of variance stems from a simple equation, stating that the total variance is equal to the sum of the variance between groups and the variation within groups –

$$\text{Total variation} = \text{variation between groups} + \text{variation within groups}$$

An analysis of variance table is used to organize data points, indicating the value of a response variable, into groups according to the factor used in each case. For example, Table 1 is an ANOVA table for comparing the amount of weight lost over a three month period by dieters on various weight-loss programs.

*Table 1 - Amount of weight lost by dieters on various programs over a 3 month period*

<table>
<thead>
<tr>
<th>Program 1</th>
<th>Program 2</th>
<th>Program 3</th>
</tr>
</thead>
</table>

https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_C…
A reasonable question is, can the type of program (a factor) be used to predict the amount of weight a dieter would lose on that program (a response variable)? Or, in other words, is any one program superior to the others?

**Measuring Variation Between Groups**

The variation between group means is measured with a weighted sum of squared differences between the sample means and the overall mean of all the data. Each squared difference is multiplied by the appropriate group sample size, $n_i$, in this sum. This quantity is called the sum of squares between groups or $SS_{Groups}$.

$$SS_{Groups} = n_1(x_1 - x)^2 + n_2(x_2 - x)^2 + \ldots + n_k(x_k - x)^2 = \sum_{groups} n_j(x_j - x)^2$$

The numerator of the F-statistic for comparing means is called the mean square between groups or $MS_{Groups}$, and it is calculated as -

$$MS_{Groups} = \frac{SS_{Groups}}{k - 1}$$

**Measuring Variation Within Groups**

To measure the variation among data points within the groups, find the sum of squared deviations between data values and the sample mean in each group, and then add these quantities. This is called the sum of squares within groups, $SSE$, or sum of squares within groups.

$$SSE = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2 = \sum_{all\, groups} (n_j - 1)s_j^2$$

Where

$$s_j^2 = \frac{\sum_{groupj} (x_{ij} - \bar{x}_j)^2}{n_j - 1} = \text{variance within each group}$$
The denominator of the F-statistic is called the **mean square error, MSE**, or **mean squares within groups**. It is calculated as

\[
MSE = \frac{SSE}{N - k} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n_1 + n_2 + \ldots + n_k - k}
\]

MSE is simply a weighted average of the sample variances for the k groups. Therefore, if all \(n_i\) are equal, MSE is simply the average of the k sample variances. The square root of MSE (\(s_p\)), called the **pooled standard deviation**, estimates the population standard deviation of the response variable (keep in mind that all of the samples being compared are assumed to have the same standard deviation \(\sigma\)).

---

**Measuring the Total Variation**

The total variation in all samples combined is measured by computing the sum of squared deviations between data values and the mean of all data points. This quantity is referred to as the **total sum of squares** or SS Total. The total sum of squares may also be referred to as SSTO. A formula for the sum of squared differences from the overall mean is

\[
SSTotal = \sum_{values} (x_{ij} - \bar{x})^2
\]

where \(x_{ij}\) represents the jth observation within the ith group, and

\(\bar{x}\)

is the mean of all observed data values. Finally, the relationship between SS Total, SS Groups, and SS Error is

\[
SS Total = SS Groups + SS Error
\]

Overall, the relationship between the total variation, the variation between groups, and the variation within a group is illustrated by Figure 2.

---

**Figure 2**: Visual illustration of variance in ANOVA, adapted from www.uwsp.edu/psych/stat/12/anova-1w.htm

https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_C... Updated: Thu, 13 May 2021 16:07:23 GMT

Powered by
A general table for performing the one-way ANOVA calculations required to compute the F-statistic is given below.

**Table 2 - One-Way ANOVA Table**

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Sum of Squares</th>
<th>F-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups</td>
<td>k-1</td>
<td>$SS_{Groups} = \sum_{groups} n_i(\bar{x}_i - \bar{x})^2$</td>
<td>$\frac{SS_{Groups}}{k - 1}$</td>
<td>$F = \frac{MS_{Groups}}{MSE}$</td>
</tr>
<tr>
<td>Within groups (error)</td>
<td>N-k</td>
<td>$SSE = \sum_{groups} (n_i - 1)s_i^2$</td>
<td>$\frac{SSE}{N - k}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>N-1</td>
<td>$SSTO = \sum_{values} (x_{ij} - \bar{x})^2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Interpreting the F-statistic**

Once the F-statistic has been found, it can be compared with a critical F value from a table, such as this one: F Table. This F table is calculated for a value of alpha = 0.05, indicating a 95% confidence level. This means that if $F_{observed}$ is larger than $F_{critical}$ from the table, then we can reject the null hypothesis and say with 95% confidence that the variance between groups is not due to random chance, but rather due to the influence of a tested factor. Tables are also available for other values of alpha and can be used to find a more exact probability that the difference between groups is (or is not) caused by random chance.

**Finding the Critical F value**

In this F Table, the first row of the F table is the number of degrees of between groups (number of groups - 1), and the first column is the number of degrees of freedom within groups (total number of samples - number of groups).
For the diet example in Table 1, the degree of freedom between groups is \((3-1) = 2\) and the degree of freedom within groups is \((13-3) = 10\). Thus, the critical \(F\) value is 4.10.

### Computing the 95% Confidence Interval for the Population Means

It is useful to know the confidence interval at which the means of the different groups are reported. The general formula for calculating a confidence interval is

\[
\text{sample mean} \pm \text{Multiplier} \times \text{Standard error}
\]

Because it is assumed that all populations have the same standard deviation

\[
s_p = \sqrt{(\text{MSE})}
\]

can be used to estimate the standard deviation within each group. Although the population standard deviation is assumed to be the same, the standard error and the multiplier may be different for each group, due to differences in group size and degrees of freedom. The standard error of a sample mean is inversely proportional to the square root of the number of data points within the sample. It is calculated as

\[
\text{SE}(\bar{x}_i) = \frac{s_p}{\sqrt{(n_i)}}
\]

The multiplier is determined using a t-distribution where the degrees of freedom are calculated as \(df = N-k\). Therefore, Insert formula here the confidence interval for a population mean is

\[
\bar{x}_i \pm t \times \frac{s_p}{\sqrt{(n_i)}}
\]

More details on confidence intervals can be found in Comparison of two means

An example for using factor analysis is the following:

You have two assembly lines. Suppose you sample 10 parts from the two assembly lines. \(H_0: s_1^2 = s_2^2\) \(H_a:\) variances are not equal Are the two lines producing similar outputs? Assume \(a=0.05\) \(F_{0.025,9,9} = 4.03\) \(F_{1-.025,9,9} = ?\)
Are variances different?

Test for Equal Variances

Nonparametric confidence intervals for standard deviations

<table>
<thead>
<tr>
<th>Lower</th>
<th>Sigma</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.200019</td>
<td>0.300019</td>
<td>0.700019</td>
</tr>
<tr>
<td>0.200019</td>
<td>0.310019</td>
<td>0.620019</td>
</tr>
</tbody>
</table>

F Test (normal distribution)

Test Statistic: 1.404
P-Value: 0.566

How would we test if the means are different?
Two-Factor Analysis of Variance

A two factor or two-way analysis of variance is used to examine how two qualitative categorical variables (male/female) affect the mean of a quantitative response variable. For example, a psychologist might want to study how the type and volume of background music affect worker productivity. Alternatively, an economist maybe be interested in determining the affect of gender and race on mean income. In both of these examples, there is interest in the effect of each separate explanatory factor, as well as the combined effect of both factors.

Assumptions

In order to use the two-way ANOVA, the following assumptions are required:

- Samples must be independent.
- Population variances must be equal.
- Groups must have same sample size. The populations from which the samples were obtained must be normally distributed (or at least approximately so).
- The null hypothesis is assumed to be true.

The null hypothesis is as follows:

- The population means for the first factor have to be equal. This is similar to the one-way ANOVA for the row factor.
- The population means for the second factor must also be equal. This is similar to the one-way ANOVA for the column factor.
- There isn’t an interaction between the two factors. This is similar to performing an independence test using contingency tables.

More simply, the null hypothesis implies that the populations are all similar and any differences in the populations are caused by chance, not by the influence of a factor. After carrying out two-way ANOVA it will be possible to analyze the validity of this assumption.
Terms Used in Two-Way ANOVA

The interaction between two factors is the most unique part of a two-way analysis of variance problem. When two factors interact, the effect on the response variable depends on the value of the other factor. For example, the statement being overweight caused greater increases in blood pressure for men than women describes an interaction. In other words, the effect of weight (factor) on blood pressure (response) depends on gender (factor).

The term main effect is used to describe the overall effect of a single explanatory variable. In the music example, the main effect of the factor "music volume" is the effect on productivity averaged over all types of music. Clearly, the main effect may not always be useful if the interaction is unknown.

In a two-way analysis of variance, three F-statistics are constructed. One is used to test the statistical significance of the interaction, while the other two are used to test the significance of the two separate main effects. The p-value for each F-statistic is also reported--a p-value of <.05 is usually used to indicate significance. When an F-factor is found to have statistical significance, it is considered a main effect. The p-value is also used as an indicator to determine if the two factors have a significant interaction when considered simultaneously. If one factor depends strongly on the other, the F-statistic for the interaction term will have a low p-value. An example output of two-way analysis of variance of restaurant tip data is given in Table 4.

Table 4 - Two-Way Analysis of Variance of Restaurant Tipping Data

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F-Statistic</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Message</td>
<td>1</td>
<td>14.7</td>
<td>14.7</td>
<td>.13</td>
<td>.715</td>
</tr>
<tr>
<td>Sex</td>
<td>1</td>
<td>2602.0</td>
<td>2602.0</td>
<td>23.69</td>
<td>0.00</td>
</tr>
<tr>
<td>Interaction</td>
<td>1</td>
<td>438.7</td>
<td>438.7</td>
<td>3.99</td>
<td>.049</td>
</tr>
<tr>
<td>Error</td>
<td>85</td>
<td>9335.5</td>
<td>109.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>88</td>
<td>12407.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this case, the factors being studied are sex (male or female) and message on the receipt ( :-) or none). The p-values in the last column are the most important information contained in this table. A lower p-value indicates a higher level of significance. Message has a significance value of .715. This is much greater than .05, the 95% confidence interval, indicating that this factor has no significance (no strong correlation between presence of message and amount of tip). The reason this occurs is that there is a relationship between the message and the sex of the waiter. The interaction term, which was significant with a value of p= 0.049, showed that drawing a happy face increased the tip for women but decreased it for men. The main effect of waiter sex (with a p-value of approximately 0) shows that there is a statistical
difference in average tips for men and women.

Two-Way ANOVA Calculations

Like in one-way ANOVA analysis the main tool used is the square sums of each group. Two-way ANOVA can be split between two different types: with repetition and without repetition. With repetition means that every case is repeated a set number of times. For the above example that would mean that the :-) was given to females 10 times and males 10 times, and no message was given to females 10 times and males 10 times.

Using the SS values as a start the F-statistics for two-way ANOVA with repetition are calculated using the chart below where a is the number of levels of main effect A, b is the number of levels of main effect B, and n is the number of repetitions.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>Adj MS</th>
<th>F-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Effect A</td>
<td>From data given</td>
<td>a-1</td>
<td>SS/df</td>
<td>MS(A)/MS(W)</td>
</tr>
<tr>
<td>Main Effect B</td>
<td>From data given</td>
<td>b-1</td>
<td>SS/df</td>
<td>MS(B)/MS(W)</td>
</tr>
<tr>
<td>Interaction Effect</td>
<td>From data given</td>
<td>(a-1)(b-1)</td>
<td>SS/df</td>
<td>MS(A*B)/MS(W)</td>
</tr>
<tr>
<td>Within</td>
<td>From data given</td>
<td>ab(n-1)</td>
<td>SS/df</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>sum of others</td>
<td>abn-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Without repetition means there is one reading for every case. For example is you were investigating whether or not difference in yield are more significant based on the day the readings were taken or the reactor that the readings were taken from you would have one reading for Reactor 1 on Monday, one reading for Reactor 2 on Monday etc... The results for two-way ANOVA without repetition is slightly different in that there is no interaction effect measured and the within row is replaced with a similar (but not equal) error row. The calculations needed are shown in the table below.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>DF</th>
<th>MS</th>
<th>F-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main Effect A</td>
<td>From data given</td>
<td>a-1</td>
<td>SS/df</td>
<td>MS(A)/MS(E)</td>
</tr>
<tr>
<td>Main Effect B</td>
<td>From data given</td>
<td>b-1</td>
<td>SS/df</td>
<td>MS(B)/MS(E)</td>
</tr>
</tbody>
</table>

https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book%3A_Chemical_Process_Dynamics_and_C…
Updated: Thu, 13 May 2021 16:07:23 GMT
Powered by
Other Methods of Comparison

Unfortunately, the conditions for using the ANOVA F-test do not hold in all situations. In this section, several other methods which do not rely on equal population standard deviations or normal distribution. It is important to realize that no method of factor analysis is appropriate if the data given is not representative of the group being studied.

Hypotheses About Medians

In general, it is best construct hypotheses about a population median, rather than the mean. Using the median accounts for the sample being skewed based on extreme outliers. Median hypotheses should also be used for dealing with ordinal variables (variables which are described only as being higher or lower than one other and do not have a precise value). When several populations are compared, the hypotheses are stated as:

\[ H_0: \text{Population medians are equal} \]
\[ H_a: \text{Population medians are not all equal} \]

Kruskal-Wallis Test for Comparing Medians

The **Kruskal-Wallis Test** provides a method of comparing medians by comparing the relative rankings of data in the observed samples. This test is therefore referred to as a rank test or non-parametric test because the test does not make any assumptions about the distribution of data.

To conduct this test, the values in the total data set are first ranked from lowest to highest, with 1 being lowest and N being highest. The ranks of the values within each group are averaged, and the test statistic measures the variation among the average ranks for each group. A p-value can be determined by finding the probability that the variation among the set of rank averages for the groups would be as large or larger as it is if the null hypothesis is true. More information on the Kruskal-Wallis test can be found [here].

Mood's Median Test for Comparing Medians

Another nonparametric test used to compare population medians is **Mood's Median Test**. Also called the Sign Scores Test, this test involves multiple steps.

1. Calculate the median (M) using all data points from every group in the study
2. Create a contingency table as follows:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3. Calculate the expected value for each data set using the following formula:

\[ \text{expected} = \frac{(r_\text{owtotal})(c_\text{olumnntotal})}{\text{grandtotal}} \]

4. Calculate the chi-square value using the following formula

\[ \chi = \frac{(\text{actual} - \text{expected})^2}{\text{expected}} \]

A chi-square statistic for two-way tables is used to test the null hypothesis that the population medians are all the same. The test is equivalent to testing whether or not the two variables are related.

---

**ANOVA and Factor Analysis in Process Control**

ANOVA and factor analysis are typically used in process control for troubleshooting purposes. When a problem arises in a process control system, these techniques can be used to help solve it. A factor can be defined as a single variable or simple process that has an effect on the system. For example, a factor can be temperature of an inlet stream, flow rate of coolant, or the position of a specific valve. Each factor can be analyzed individually to determine the effect that changing the input has on the process control system as a whole. The input variable can have a large, small, or no effect on what is being analyzed. The amount that the input variable affects the system is called the “factor loading”, and is a numerical measure of how much a specific variable influences the system or the output variable. In general, the larger the factor loading is for a variable the more of an effect that it has on the output variable.

A simple equation for this would be:

\[ \text{Output} = f_1 \ast \text{input}_1 + f_2 \ast \text{input}_2 + ... + f_n \ast \text{input}_n \]

Where \( f_n \) is the factor loading for the \( n^{th} \) input.

Factor analysis is used in this case study to determine the fouling in an alcohol plant reboiler. This article provides some additional insight as to how factor analysis is used in an industrial situation.
Using Mathematica to Conduct ANOVA

Mathematica can be used for one-way and two-way factor analyses. Before this can be done, the ANOVA package must be loaded into Mathematica using the following command:

```
Needs["ANOVA"]
```

Once this command is executed, the 'ANOVA' command can be utilized.

### One-Way Factor Analysis

The basic form of the 'ANOVA' command to perform a one-way factor analysis is as follows:

```
ANOVA[data]
```

An example set of data with five elements would look like:

```
ANOVA[ParseError: EOF expected (click for details)]
```

An output table that includes the degrees of freedom, sum of the squares, mean sum of the squares, F-statistic, and the P-value for the model, error, and total will be displayed when this line is executed. A list of cell means for each model will be displayed beneath the table.

### Two-Way Factor Analysis

The basic form of the 'ANOVA' command to perform a two-way factor analysis is as follows:

```
ANOVA[data, model, vars]
```

An example set of data with seven elements would look like:

```
ANOVA[ParseError: EOF expected (click for details)]
```

https://eng.libretexts.org/Bookshelves/Industrial_and_Systems_Engineering/Book:_Chemical_Process_Dynamics_and_C...
An output table will appear similar to the one that is displayed in the one-way analysis except that there will be a row of statistics for each variable (i.e. x,y).

**ANOVA in Microsoft Excel 2007**

In order to access the ANOVA data analysis tool, install the package:

1. Click on the Microsoft Office button (big circle with office logo)
2. Click 'Excel Options'
3. Click 'Add-ins' on the left side
4. In the manage drop-down box at the bottom of the window, select 'Excel Add-ins'
5. Click 'Go...'
6. On the Add-Ins window, check the Analysis ToolPak box and click 'OK'

To use this package:

1. Click on the 'Data' tab and select 'Data Analysis'
2. Choose the desired ANOVA type- 'Anova: Single Factor', 'Anova: Two Factor with Replication', or 'Anova: Two Factor without Replication'(see note below for when to use replication)
3. Select the desired data points including data labels at top of the corresponding columns. Make sure the box is checked for 'Labels in first row' in the ANOVA parameter window.
4. Specify alpha in the ANOVA parameter window. Alpha represents the level of significance.
5. Output the results into a new worksheet.

**NOTE:** Anova: Two Factor with Replication is used in the cases where there are multiple readings for a single factor. For instance, the input below, there are 2 factors, control architecture and unit. This input shows how there are 3 readings corresponding to each control architecture (FB, MPC, and cascade). In this sense, the control architecture is replicated 3 times, each time providing different data relating to each unit. So, in this case, you would want to use the Anova Two Factor with Replication option.
Anova: Two Factor without Replication is used in cases where there is only one reading pertaining to a particular factor. For example, in the case below, each sample (row) is independent of the other samples since they are based on the day they were taken. Since multiple readings were not taken within the same day, the "without Replication" option should be chosen.

Excel outputs:

**Summary:**
1. **Count**: number of data points in a set  
2. **Sum**: sum of the data points in a set  
3. **Average**: mean of the data points in a set  
4. **Variance**: standard deviation of the data points in a set

**ANOVA:**
1. **Sum of squares (SS)**  
2. **The degree of freedom (df)**  
3. **The mean squares (MS)**  
4. **F-statistic (F)**  
5. **P-value**  
6. **$F_{critical}$**

See the figure below for an example of the inputs and outputs using Anova: Single Factor. Note the location of the Data Analysis tab. The data was obtained from the dieting programs described in Table 1. Since the F-statistic is greater than $F_{critical}$, the null hypothesis can be rejected at a 95% confidence level (since alpha was set at 0.05). Thus, weight loss was not random and in fact depends on diet type chosen.
Worked out Example 1

Determine the fouling rate of the reboiler at the following parameters:

\[ T = 410 \text{K} \]
\[ C_c = 16.7 \text{g/L} \]
\[ R_T = 145 \text{min} \]

Which process variable has the greatest effect (per unit) on the fouling rate of the reboiler?

Note that the tables below are made up data. The output data for a single input was gathered assuming that the other input variables provide a negligible output. Although the factors that affect the fouling of the reboiler are similar to the ones found the the article linked in the "ANOVA and Factor Analysis in Process Controls" section, the data is not.

<table>
<thead>
<tr>
<th>Temperature of Reboiler (K)</th>
<th>400</th>
<th>450</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fouling Rate (mg/min)</td>
<td>0.8</td>
<td>0.86</td>
<td>0.95</td>
</tr>
<tr>
<td>Catalyst Concentration (g/L)</td>
<td>10</td>
<td>20</td>
<td>30</td>
</tr>
</tbody>
</table>
Solution:

1) Determine the "factor loading" for each variable.

This can be done using any linearization tool. In this case, the factor loading is just the slope of the line for each set of data. Using Microsoft Excel, the equations for each set of data are the following:

**Temperature of Reboiler**

\[ y = 0.0015 \times x + 0.195 \]

Factor loading: 0.0015

**Catalyst Concentration**

\[ y = 0.0805 \times x - 0.2833 \]

Factor loading: 0.0805

**Residence Time**

\[ y = 0.0238 \times x - 0.5067 \]

Factor loading: 0.0238

2) Determine the fouling rate for the given process conditions and which process variable affects the fouling rate the most (per unit). Note that the units of the factor loading value are **always** the units of the output divided by the units of the input.
Plug in the factor loading values into the following equation:
\[ \text{Output} = f_1 \ast \text{input}_1 + f_2 \ast \text{input}_2 + \ldots + f_n \ast \text{input}_n \]

You will end up with:
\[ \text{FoulingRate} = 0.0015 \ast T + 0.0805 \ast C_c + 0.0238 \ast RT \]

Now plug in the process variables:
\[ \text{FoulingRate} = 0.0015 \ast 410 + 0.0805 \ast 16.7 + 0.0238 \ast 145 \]
\[ \text{FoulingRate} = 5.41 \text{ mg / min} \]

The process variable that affects the fouling rate the most (per unit) is the catalyst concentration because it has the largest factor loading value.

**Worked out Example 2**

**Problem:**

The exit flow rate leaving a tank is being tested for 3 cases. The first case is under the normal operating conditions, while the second (A) and the third (B) cases are for new conditions that are being tested. The flow value of 7 (gallons/hour) is desired with a maximum of 10. A total of 24 runs are tested with 8 runs for each case. The tests are run to determine whether any of the new conditions will result in a more accurate flow rate. First, we determine if the new conditions A and B affect the flow rate. The results are as follows:

<table>
<thead>
<tr>
<th>Case 1</th>
<th>7</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>6</th>
<th>2</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 2</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Case 3</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

The recorded values for the 3 cases are tabulated. Following this the values for each case are squared and the sums for
all of these are taken. For the 3 cases, the sums are squared and then their means are found.

These values are used to help determine the table above (the equations give an idea as to how they are calculated). In the same way with the help of ANOVA, these values can be determine faster. This can be done using the mathematica explained above.

Conclusion:

\[ F_{critical} \] equals 3.4668, from an F-table. Since the calculated F value is greater than \( F_{critical} \), we know that there is a statistically significant difference between 2 of the conditions. Thus, the null hypothesis can be rejected. However we do not know between which 2 conditions there is a difference. A post-hoc analysis will help us determine this. However we are able to confirmed that there is a difference.

**Worked out Example 3**

As the new engineer on site, one of your assigned tasks is to install a new control architecture for three different units. You test three units in triplicate, each with 3 different control architecture: feedback (FB), model predictive control (MPC) and cascade control. In each case you measure the yield and organize the data as follows:

<table>
<thead>
<tr>
<th>control architecture</th>
<th>unit1</th>
<th>unit2</th>
<th>unit3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FB</td>
<td>52.59</td>
<td>51.48</td>
<td>49.63</td>
</tr>
<tr>
<td>FB</td>
<td>48.12</td>
<td>50.56</td>
<td>50.10</td>
</tr>
<tr>
<td>FB</td>
<td>50.00</td>
<td>49.22</td>
<td>49.83</td>
</tr>
<tr>
<td>MPC</td>
<td>49.67</td>
<td>50.61</td>
<td>50.07</td>
</tr>
<tr>
<td>MPC</td>
<td>50.21</td>
<td>51.33</td>
<td>50.44</td>
</tr>
<tr>
<td>MPC</td>
<td>49.84</td>
<td>51.81</td>
<td>51.41</td>
</tr>
<tr>
<td>cascade</td>
<td>51.29</td>
<td>49.23</td>
<td>50.85</td>
</tr>
<tr>
<td>cascade</td>
<td>50.91</td>
<td>48.47</td>
<td>49.95</td>
</tr>
<tr>
<td>cascade</td>
<td>50.04</td>
<td>47.70</td>
<td>51.16</td>
</tr>
</tbody>
</table>

Do the units differ significantly? Do the control architectures differ significantly?

Answer: This problem can be solved using ANOVA Two factor with replication analysis.
Multiple Choice Question 1

ANOVA analysis works better for?

A. Non-linear models
B. Linear models
C. Exponential models
D. All of the above

Multiple Choice Question 2

Two-Way ANOVA analysis is used to compare?

A. Any two sets of data
B. Two One-Way ANOVA models to each other
C. Two factors on their effect of the output
D. B and C
Multiple Choice Answers

Question 1: B

Question 2: C

Sage's Corner

If you are a sage for this page, please link your narrated powerpoint presentation here.

For just the slides: controls.engin.umich.edu/wiki/index.php/Image:Anchor2.ppt
video.google.com/googleplayer.swf?docid=-3237677123667838406
video.google.com/googleplayer.swf?docid=6201287739553253294

For just the slides: controls.engin.umich.edu/wiki/index.php/Image:ANOVA_Slides.ppt
www.youtube.com/v/fA7xPxbM6NY

For just the slides: controls.engin.umich.edu/wiki/index.php/Image:Mn-anova-presentation.ppt
www.youtube.com/v/AEbYrMZEyac

References