Analysis of Variance

The two primary methods for analysing experimental data are Analysis of Variance (ANOVA) and Regression. ANOVA is generally used to identify which variables should be included in the experimental model and regression is used to estimate the coefficients for the variables, but either ANOVA or regression can be used to identify the important variables and determine the coefficients for these variables. Another consideration is that, when using ANOVA data must be collected in a systematic manner, usually from a designed experiment, while regression can generally be used on any data set.

Consider the industrial rolling of aluminium sheet metal, and there is an attempt to determine which factors influence the variability of the thickness of the sheet metal. This may not be possible by analysing existing production data. The factors of interest are often co-linear when analysing non-experimental data. In a designed experiment, great care is taken to be able to prevent confounding (puzzling) of the main effects and some if not all the interactions, but this is not the case for production data. Back to the aluminium sheet example, the factors that are of interest are speed, roll pressure, sheet temperature, and the rate of injection for coolant. After analysing the data, it is found the rate of coolant injection is statistically significant, but the speed is not. However, if the rate of coolant injection is not included in the model, the speed is the most significant variable. How can this happen? This is common occurrence when two variables are co-linear. Unlike a designed experiment where the settings of the rate of coolant injection and speed would be carefully set, in the historical data, when the speed is reduced, the rate of coolant injection was also reduced, thus, it is impossible to separate the effects of these two variables.

Analysis of Variance (ANOVA) is used to determine which factors have a statistically significant effect on the response variable. Any ANOVA has the following assumptions:

- the treatment data is normally distributed,
- the variance is the same for all treatments, and
- all samples are randomly selected and are independent.

One-Way ANOVA

A one-way ANOVA is used to determine if the mean is equal for different levels of a **single** variable. If there are only 2 levels for the variable, then a *t*-test is used, but when there are more than 2 levels, ANOVA is required. For example, is the mean hardness of a rubber compound equal if a component of the rubber is supplied by 5 different vendors. The null hypothesis being tested in a one-way ANOVA is:

H_0: All treatment means are equal.

The alternate hypothesis is:

H_1: At least one of the treatment means is different.

The one-way ANOVA table is shown below.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Between Treatments	SST	k-1	MST = SST/(k-1)	F = MST/MSE
Error	SSE	N-k	MSE = SSE/(N-k)	
Total	TSS	N-1		

SST = sum of squares between treatments

$$SST = \sum_{j=1}^{k} \frac{(Y_{\bullet j})^2}{n} - \frac{Y_{\bullet\bullet}^2}{N}$$

$Y_{.j}$ = the sum of observations for treatment j.
$Y_{..}$ = the sum of all observations.
SSE = sum of squares due to error.
SSE = TSS - SST
TSS = total sum of squares.

$$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n} y_{ij}^2 - \frac{Y_{\bullet\bullet}^2}{N}$$

MST = mean square for treatments.
MSE = mean square for error.
k = number of treatment levels.
n = number of runs at a particular level
N = total number of runs
F = the calculated F statistic with k-1 and N-k degrees of freedom.

If the calculated F is greater than the critical F value then reject the null hypothesis (i.e., all the means are not statistically the same at the α value chosen.)

Example

An automobile bearing is experiencing a higher-than-expected failure rate due to low strength. Upon investigation, it is discovered that the supplier produces this bearing on 4 different machines. The strength of randomly selected bearings from the 4 machines is given below; at 5% significance, is there any difference in the bearing strength based on the production machine?

Machine A	Machine B	Machine C	Machine D

47.6	52.8	50.4	52.4
52.5	51.4	55.8	51.6
45.9	49.7	49.8	57.8
51.6	54.2	51.8	53.2
51.7	53.6	51.8	49.1
50.4	52.8	53.2	51.6
48	54	51.2	44.2
50.3			49.7

Solution

The sum of all 30 observations is 1,540.1. The sum of the square of the 30 observations is 79,281.35. The total sum of squares is:

$$TSS = 79{,}281.35 - \frac{1{,}540.1^2}{30} = 217.75$$

The sum of the 8 observations for Machine A is 398.0. The sum of the 7 observations for Machine B is 368.5. The sum of the 7 observations for Machine C is 364.0. The sum of the 8 observations for Machine D is 409.6. The sum of squares between treatments is:

$$SST = \frac{398.0^2}{8} + \frac{368.5^2}{7} + \frac{364.0^2}{7} + \frac{409.6^2}{8} - \frac{1{,}540.1^2}{30} = 35.31$$

The sum of squares for the error term is the difference between the total sum of squares and the sum of squares between treatments.

SSE = 217.75 - 35.31 = 182.44

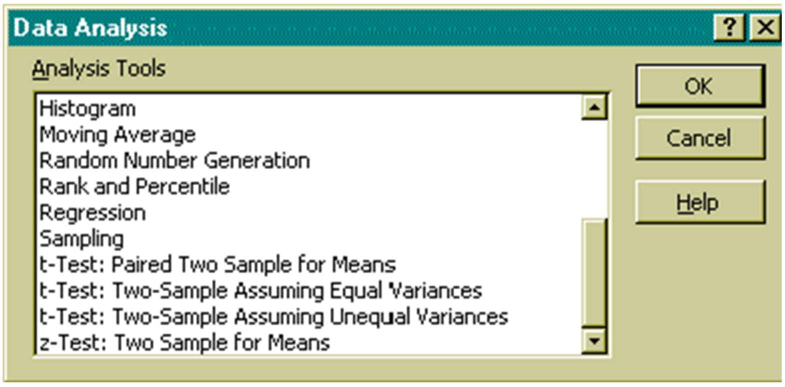
The degrees of freedom between treatments is equal to the number of treatments minus 1, which equals 3. The degrees of freedom for the error is the total number of observations less the number of treatments which is 30-4 = 26. The critical value for the *F*-distribution at 5% significance with 3 and 26 degrees of freedom is 2.98. The complete ANOVA table is shown in the table below. Since the computed *F*-statistic is less than the critical value, the null hypothesis is not rejected - there is not enough evidence to show that the mean strength is not the same for bearings produced on all machines.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Between Treatments	35.31	3	11.77	1.68
Error	182.44	26	7.02	
Total	217.75	29		

Detailed calculations are available in the Microsoft Excel file "**Chapt-9 ANOVA.xlsx**".

Solution Using Excel's Analysis Tool Pack

This problem is solved much easier with the assistance of Microsoft Excel. Microsoft Excel contains an *Analysis Tool Pack* that can be used to perform one-way ANOVA. To access this package, click the *Tools* from the menu then click *Data Analysis*. If *Data Analysis* is not on the menu the *Analysis Tool Pack* has not been activated. To activate the *Analysis Tool Pack,* click the *Tools* menu and select *Add-Ins*, then check the *Analysis Tool Pack* box and click *OK*. Click the "Tools" menu, select "Data Analysis" and the screen shown below will appear.



Select "ANOVA: Single Factor", then click "OK". The configuration screen will appear. Simply follow instructions.

Multi-Factor ANOVA

When more than one factor is of interest, a more general ANOVA is required for analysis. The 2-way and 3-way ANOVA tables are shown below, and the 3-way ANOVA can be expanded to accommodate any desired number of factors. With any ANOVA the mathematics is tedious, but there are several commercial software packages available to perform these computations.

Two-Way ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Factor A	SSA	a-1	MSA = SSA/(a-1)	F = MSA/MSE
Factor B	SSB	b-1	MSB = SSB/(b-1)	F = MSB/MSE
Interaction	SSAB	(a-1)(b-1)	MSAB=SSAB/[(a-1)(b-1)]	F = MSAB/MSE
Error	SSE	ab(n-1)	MSE = SSE/[ab(n-1)]	
Total	TSS	N-1		

$$SSA = \sum_{i=1}^{a} \frac{(Y_{i..})^2}{bn} - \frac{Y_{...}^2}{N}$$

$$SSB = \sum_{j=1}^{b} \frac{(Y_{.j.})^2}{an} - \frac{Y_{...}^2}{N}$$

$$SSAB = \sum_{i=1}^{a} \sum_{i=1}^{b} \frac{(Y_{ij.})^2}{n} - \frac{Y_{...}^2}{N} - SSA - SSB$$

$$TSS = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} Y_{ijk}^2 - \frac{Y_{...}^2}{N}$$

a = number of treatment levels for factor A
b = number of treatment levels for factor B
n = the number of repetitions for each experimental setting
N = total number of runs (*abn*)

Three-Way ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-Statistic
Factor A	SSA	a-1	MSA = SSA/(a-1)	F = MSA/MSE
Factor B	SSB	b-1	MSB = SSB/(b-1)	F = MSB/MSE
Factor C	SSC	c-1	MSC = SSC/(c-1)	F = MSC/MSE
Interaction AB	SSAB	(a-1)(b-1)	MSAB=SSAB/[(a-1)(b-1)]	F = MSAB/MSE
Interaction AC	SSAC	(a-1)(c-1)	MSAC = SSAC/[(a-1)(c-1)]	F = MSAC/MSE
Interaction BC	SSBC	(b-1)(c-1)	MSAB = SSBC/[(b-1)(c-1)]	F = MSBC/MSE
Interaction ABC	SSABC	(a-1)(b-1)(c-1)	MSABC = SSABC/[(a-1)(b-1)(c-1)]	F = MSABC/MSE
Error	SSE	abc(n-1)	MSE = SSE/[abc(n-1)]	
Total	TSS	N-1		

$$SSA = \sum_{i=1}^{a} \frac{Y_{i...}^2}{bcn} - \frac{Y_{....}^2}{N}$$

$$SSB = \sum_{j=1}^{b} \frac{Y_{.j..}^2}{acn} - \frac{Y_{....}^2}{N}$$

$$SSC = \sum_{k=1}^{c} \frac{Y_{..k.}^2}{abn} - \frac{Y_{....}^2}{N}$$

$$SSAB = \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{Y_{ij..}^2}{cn} - \sum_{i=1}^{a}\frac{Y_{i...}^2}{bcn} - \sum_{j=1}^{b}\frac{Y_{.j..}^2}{acn} + \frac{Y_{....}^2}{N}$$

$$SSAC = \sum_{i=1}^{a}\sum_{k=1}^{c}\frac{Y_{i.k.}^2}{bn} - \sum_{i=1}^{a}\frac{Y_{i...}^2}{bcn} - \sum_{k=1}^{c}\frac{Y_{..k.}^2}{abn} + \frac{Y_{....}^2}{N}$$

$$SSBC = \sum_{j=1}^{b}\sum_{k=1}^{c}\frac{Y_{.jk.}^2}{an} - \sum_{j=1}^{b}\frac{Y_{.j..}^2}{acn} - \sum_{k=1}^{c}\frac{Y_{..k.}^2}{abn} + \frac{Y_{....}^2}{N}$$

$$SSABC = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{Y_{ijk.}^2}{n} - \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{Y_{ij..}^2}{cn} -$$
$$\sum_{i=1}^{a}\sum_{k=1}^{c}\frac{Y_{i.k.}^2}{bn} - \sum_{j=1}^{b}\sum_{k=1}^{c}\frac{Y_{.jk.}^2}{an} + \sum_{i=1}^{a}\frac{Y_{i...}^2}{bcn} + \sum_{j=1}^{b}\frac{Y_{.j..}^2}{acn}$$
$$+ \sum_{k=1}^{c}\frac{Y_{..k.}^2}{abn} - \frac{Y_{....}^2}{N}$$

$$TSS = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\sum_{l=1}^{n} Y_{ijkl}^2 - \frac{Y_{....}^2}{N}$$

a = number of treatment levels for factor A
b = number of treatment levels for factor B
c = number of treatment levels for factor C
n = the number of repetitions for each experimental setting
N = total number of runs (*abcn*)

Example

To understand the variables that impact the brightness of rolled sheet metal, the engineering department conducted an experiment using 6 factors with 2 levels each. The experimental results are shown in the table below. Which of these factors affect brightness at 5% significance?

Trial	Temperature	Speed	Angle	Pressure	Water	Kerosene	Brightness
1	400	1000	0	1.5	0%	2%	4.2
2	400	1500	0	1.5	1%	5%	5.1
3	400	1000	15	2.5	0%	5%	3.7
4	400	1500	15	2.5	1%	2%	5.6
5	600	1000	15	1.5	1%	2%	2.9
6	600	1500	15	1.5	0%	5%	3.8
7	600	1000	0	2.5	1%	5%	4.7
8	600	1500	0	2.5	0%	2%	3.6

Solution

Since there are only 8 trials, there is 1 degree of freedom for estimating the overall mean, 6 degrees of freedom for estimating the effect of the 6 factors, and 1 degree of freedom for experimental error. There are no degrees of freedom available for estimating the effect of any interactions. The brightness sum for all 8 trials is 33.6. The total sum-of-squares is:

$$TSS = \sum_{i=1}^{8} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 4.2^2 + 5.1^2 + 3.7^2 + 5.6^2 + 2.9^2 + 3.8^2$$

$$+4.7^2 + 3.6^2 - \frac{33.6^2}{8} = 5.48$$

The sum-of-squares due to temperature is computed from the squared differences between the average of all experimental trials, 4.2, and the average of the trials at each setting of temperature. When the temperature is set at 400, the average of the 4 trials was 4.65; when temperature 600, the average of the 4 trials was 3.75. The sum-of-squares for temperature is

$SS_{temp} = 4(4.65-4.2)^2 + 4(3.75-4.2)^2 = 1.62$

The sum-of-squares for the remaining factors are:

$SS_{speed} = 4(3.875-4.2)^2 + 4(4.525-4.2)^2 = 0.845$

$SS_{angle} = 4(4.4-4.2)^2 + 4(4.0-4.2)^2 = 0.32$

$SS_{pressure} = 4(4.0-4.2)^2 + 4(4.4-4.2)^2 = 0.32$

$SS_{water} = 4(3.825-4.2)^2 + 4(4.575-4.2)^2 = 1.125$

$SS_{kerosene} = 4(4.075-4.2)^2 + 4(4.325-4.2)^2 = 0.125$

The sum-of-squares error is the difference between the total sum-of-squares and the sum-of-squares for the factors and is 1.125. The degrees-of-freedom for the sum-of-squares for each factor is equal to the number of levels for that factor minus 1. The degrees-of-freedom for the total is equal to the number of experimental trials minus 1. The mean square for each factor is equal to the sum-of-squares divided by the degrees-of-freedom. The F-statistic for each factor is equal to the mean square for that factor divided by the mean square error. This is summarized in the table below. The critical value of the F-statistic for each of the factors at 5% significance using 1 and 1 degrees of freedom is 161.4. Since none of the computed values from the F-distribution exceed the critical value, none of the factors are considered significant.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Statistic
Temperature	1.62	1	1.62	1.44
Speed	0.845	1	0.845	0.75
Angle	0.32	1	0.32	0.28
Pressure	0.32	1	0.32	0.28
Water	1.125	1	1.125	1
Kerosene	0.125	1	0.125	0.11

Error	1.125	1	1.125	
Total	5.48	7		

Detailed calculations are available in the Microsoft Excel file "**Chapt-9 ANOVA.xlsx**".