ANOVA

In most realistic scenarios, the difference between prediction (explanatory) variables and response variables is cloudy. In most cases the predictors are not independent but influence each other. The method used to compare the means of all the variables is known as the **analysis of variance**, or **ANOVA**. We are essentially comparing the variance of the *Within*-sample variation and *Between*-sample variation.

Let $n \times b$ be the total number of data where b is the number of groups (usually columns) and n is the number of samples (usually rows)

$$\begin{aligned} \textit{Within}\text{-sample variation} &= \frac{\sum_{i=1}^{n_1}(y_{i1}-\bar{y}_1)^2+\sum_{i=1}^{n_2}(y_{i2}-\bar{y}_2)^2+\cdots+\sum_{i=1}^{n_b}(y_{ib}-\bar{y}_b)^2}{\left(\sum_{i=1}^{b} n_i\right)-b}, \\ &= \frac{\text{SSE}}{\left(\sum_{i=1}^{b} n_i\right)-b}, \\ &= \frac{\text{SSE}}{b(n-1)}, \quad \text{(if all columns have the same number of data).} \end{aligned}$$

where SSE is **Sum of Squared Errors** and the denominator is the *degrees of freedom*.

$$\textit{Between}\text{-sample variation} = \frac{n_1(\bar{y}_1-\bar{y})^2+n_2(\bar{y}_2-\bar{y})^2+\cdots+n_b(\bar{y}_b-\bar{y})^2}{(\sum b_i)-1} = \frac{\text{SST}}{(b-1)}.$$

SST is the **Sum of Squares for Treatments** and the denominator is the *degrees of freedom*. Both SSE and SST sum to a known total.

$$\text{SS(total)} = \sum (y_i-\bar{y})^2.$$

The value of the F statistic is the following ratio:

$$F = \frac{\text{Between-sample variation}}{\text{Within-sample variation}}.$$

Now we can place everything into a table.

Source	Expression	df (degs. of freedom)	SS (Sum of Squares Err.)	MS (Mean Square Err.)	F	p-value
Between groups B	$[B] = \frac{\sum_i (\sum_j B_{i,j})^2}{n}$	$b-1$	$[B]-[T]$	$\frac{SS_B}{df_b}$	$\frac{MS_B}{MS_{s/b}}$	F table lookup
Within Groups S/B $[W]$	$[W] = \sum\sum W_{i,j}^2$	$b(n-1)$	$[W]-[B]$	$\frac{SS_{S/B}}{df_{s/b}}$		
Total	$[T] = \frac{(\sum\sum T_{i,j})^2}{bn}$	$bn-1$	$[W]-[T]$			

Let's look at an example where $n = 4$ and $b = 5$.

	Y_1	Y_2	Y_3	Y_4	Y_5
1	6	4	5	8	10
2	4	5	7	9	8
3	4	3	6	6	5
4	5	2	6	7	3

First calculate the mean of the total SS:

$$[T] = \frac{(6+4+4+5+4+5+3+2+5+7+6+6+8+9+6+7+10+8+5+3)^2}{20} = 638.45$$

Calculate between groups:

$$[B] = \frac{(6+4+4+5)^2 + (4+5+3+2)^2 + (5+7+6+6) + (8+9+6+7)^2 + (10+8+5+3)^2}{4} = 677.25$$

Calculate within groups:

$$[W] = 6^2+4^2+4^2+5^2+4^2+5^2+3^2+2^2+5^2+7^2+6^2+6^2+8^2+9^2+6^2+7^2+10^2+8^2+5^2+3^2 = 721$$

Now we can place everything into a table

Source	Expression	df (degs. of freedom)	SS (Sum of Squares Err.)	MS (Mean Square Err.)	F	p-value
Between groups B	$[B] = 677.25$	$5 - 1 = 4$	$677.25 - 638.45 = 38.8$	$\frac{38.8}{4} = 9.7$	3.3257	F table lookup
Within Groups S/B $[W]$	$[W] = 721$	$20 - 5 = 15$	$721 - 677.25 = 43.75$	$\frac{43.75}{15} = 2.917$		
Total	$[T] = 638.45$	$20 - 1 = 19$	721-638.45 = 82.55			

The F critical value for 4 degrees of freedom in the numerator and 15 degrees of freedom in the denominator is 3.0556828, so we are above the critical value. The p value is 0.03868965 which is above the 95th percentile, so the results are significant.

We can also calculate a different way:
Let's look at the same example where $n = 4$ and $b = 5$.

	Y_1	Y_1^2	Y_2	Y_2^2	Y_3	Y_3^2	Y_4	Y_4^2	Y_5	Y_5^2
1	6	36	4	16	5	25	8	64	10	100
2	4	16	5	25	7	49	9	81	8	64
3	4	16	3	9	6	36	6	36	5	25
4	5	25	2	4	6	36	7	49	3	9
total	19	93	14	54	24	146	30	230	26	198

$$\begin{aligned}
\text{SSE-between} &= \left(\frac{19^2}{4} + \frac{14^2}{4} + \frac{24^2}{4} + \frac{30^2}{4} + \frac{26^2}{4}\right) - \frac{(19+14+24+30+26)^2}{20.0} = 38.8 \\
\text{SSE-within} &= \left(93 - \frac{19^2}{4}\right) + \left(54 - \frac{14^2}{4}\right) + \left(146 - \frac{24^2}{4}\right) + \left(230 - \frac{30^2}{4}\right) + \left(198 - \frac{26^2}{4}\right) = 43.75 \\
\text{SSE-total} &= (93+54+146+230+198) - \frac{(19+14+24+30+26)^2}{20} = 82.55 \qquad (1)
\end{aligned}$$

We have 38.8 in the numerator and 43.75 in the denominator which is the same result as the first method.
The total SSE is 82.55

To get the total SSE from the SSs in first method above take the difference 721 -638.45 = 82.55