

Analysis of Variance for a One-Way Classification of Data

Consider a single factor or treatment done at k levels (i.e., there are 1, 2, 3, ... k different variations on the prescribed treatment). Within a given treatment level there are n_i measurements or scores. The subscript or index i is between 1 and k and labels the different factor levels or treatment variations. The j 'th score in the i 'th level is designated as

$$y_{i,j}$$

i is the the treatment factor index; $1 \leq i \leq k$

j labels the score the within the i 'th treatment: $1 \leq j \leq n_i$.

Note: It is not required and it is typically not the case that you have an equal number of measurements from each treatment group.

The null hypothesis asserts that no treatment population differs from any other. Thus, the k populations of scores of the treatment levels should have the same mean and variance. For theoretical convenience, we will assume the scores are normally distributed with a common variance within each treatment level, but the tests which follow are fairly robust for departures from normality.

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3 = \dots \mu_k$

This is equivalent to saying that the k different treatments are really just k different samples all taken from a single population. Thus, all of the variation seen in the measured scores is due to sample variations, i.e., randomness.

The total number of scores is $n_1 + n_2 + n_3 + \dots n_k = n$. In R. Johnson's text, this is designated as upper case N , however, to be consistent with our earlier notation that lower case n is associated with a sample and upper case with a population, these notes will use n .

$$\sum_{i=1}^k n_i = n \quad (1)$$

The average of all scores is often called the "grand mean" and is given by

$$\bar{y} = \frac{y_{1,1} + y_{1,2} + y_{1,3} + \dots y_{1,n_1} + y_{2,1} + y_{2,2} + y_{2,3} + \dots y_{2,n_2} + \dots y_{k,1} + y_{k,2} + y_{k,3} + \dots y_{k,n_k}}{n}$$

For notational convenience, this is written as a double summation.

$$\bar{y} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j}}{n} \quad (2)$$

If the null hypothesis is true and all of the treatments give the same mean result, then \bar{y} should be a fair estimate of every score. Formally, we can write the following.

$$y_{i,j} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{i,j} - \bar{y}_i) \quad (3)$$

Here \bar{y}_i is the mean of treatment variation i .

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{i,j}}{n_i} \quad (4)$$

In equation (3), $\bar{y}_i - \bar{y}$ can be looked upon as a correction term to the grand mean to take account of how treatment group i differs from the grand mean, while $y_{i,j} - \bar{y}_i$ is an additional correction term to take care of variations **within** treatment group i . If the null hypothesis is true, both of these terms should be of the same magnitude and both reflect random deviations from the grand mean. From equation (4) within each treatment the sum of the deviations about the treatment mean vanishes.

$$\sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i) = 0 \quad (5)$$

Furthermore,
$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}) = \sum_{i=1}^k n_i \bar{y}_i - \bar{y} \cdot n = \sum_{i=1}^k n_i \frac{\sum_{j=1}^{n_i} y_{i,j}}{n_i} - \bar{y} \cdot n = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j} - \bar{y} \cdot n = 0,$$

where the zero is a consequence of equation (2).

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y}) = \sum_{i=1}^k n_i \bar{y}_i - \bar{y} \cdot n = 0 \quad (6)$$

So, the sum of the deviations of the treatment means about the grand mean vanishes. This means that the sample means have $k - 1$ degrees of freedom since only $k - 1$ of them can be arbitrarily specified given the value of the grand mean.

Consider the total variation of y , the sum of squared deviations of each score about the grand mean. The deviation of any score is given by the following.

$$y_{i,j} - \bar{y} = (y_{i,j} - \bar{y}_i) + (\bar{y}_i - \bar{y}) \quad (7)$$

So the sum of squared deviations or total variation in y is

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \left[(y_{i,j} - \bar{y}_i)^2 + 2(\bar{y}_i - \bar{y})(y_{i,j} - \bar{y}_i) + (\bar{y}_i - \bar{y})^2 \right] \quad (8)$$

Now from equation (5),
$$\sum_{i=1}^k \sum_{j=1}^{n_i} 2(\bar{y}_i - \bar{y})(y_{i,j} - \bar{y}_i) = 2 \sum_{i=1}^k (\bar{y}_i - \bar{y}) \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i) = 0.$$

So the total variation in y is made up of two contributions.

$$SS_y = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} \left[(y_{i,j} - \bar{y}_i)^2 + (\bar{y}_i - \bar{y})^2 \right] = SSE + SS_{Tr} \quad (9)$$

$$SS_{Tr} = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 \quad (10)$$

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2 \quad (11)$$

The term SS_{Tr} is the **Treatment** sum of squares and represents that part of the total variation in y that is due to the differences between the treatments (factor levels). This variation has $k - 1$ degrees of freedom. The second term, SSE , is the residual (left over) variation due to differences in scores within the different treatment groups. The i 'th treatment level has $n_i - 1$ degrees of freedom due to the constraint imposed by equation (5). The total degrees of freedom of the residual or **Error** sum of squared deviations is given by

$$\sum_{i=1}^k (n_i - 1) = \sum_{i=1}^k n_i - k = n - k \quad (12)$$

The sum of the degrees of freedom of the treatment sum of squares and the error sum of squares is $k - 1 + n - k = n - 1$. This is the degrees of freedom of the total variation SS_y . This is $n - 1$ since the grand mean "uses up" a degree of freedom. R. Johnson's text uses the notation SST for SS_y and $SS(Tr)$ for SS_{Tr} .

If the null hypothesis is indeed true, then the mean squares of treatment and error both estimate the same common population of each treatment population. The mean squares are computed as a variation divided by associated degrees of freedom.

The Treatment or Between Mean Square: $MS_{Tr} = \frac{SS_{Tr}}{k - 1} = \frac{\sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2}{k - 1} \quad (13)$

The Error or Within Mean Square: $MSE = \frac{SSE}{n - k} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2}{n - k} \quad (14)$

To test the null hypothesis that all treatment levels have the same mean, we compute the observed Fisher F score.

$$F = \frac{MS_{Tr}}{MSE} \quad (15)$$

For a given level of significance, α , this value is compared against a critical score calculated from an F distribution used to compare two variances obtained from sampling variances from two normally distributed populations. The numerator degrees of freedom is $k - 1$ and the denominator degrees of freedom is $n - k$. If $F > F_{\alpha}(k - 1, n - k)$, the null hypothesis is rejected, while if $F < F_{\alpha}(k - 1, n - k)$ we fail to reject H_0 .

To facilitate the actual calculations, we define the following intermediate variables.

$$T_i = \sum_{j=1}^{n_i} y_{i,j} \quad (16)$$

$$A = \sum_{i=1}^k \frac{T_i^2}{n_i} \quad (18)$$

$$B_i = \sum_{j=1}^{n_i} (y_{i,j})^2 \quad (17)$$

$$T = \sum_{i=1}^k T_i \quad (19)$$

$$B = \sum_{i=1}^k B_i \quad (20)$$

From equations (2) and (4),

$$\bar{y}_i = \frac{T_i}{n_i} \quad (21)$$

$$\bar{y} = \frac{T}{n} \quad (22)$$

Also the i 'th sample variance is given by $s_i^2 = \frac{B_i - \frac{T_i^2}{n_i}}{n_i - 1}$. (23)

Now, $SS_{Tr} = \sum_{i=1}^k n_i (\bar{y}_i^2 - 2\bar{y}_i \bar{y} + \bar{y}^2) = \sum_{i=1}^k n_i \left(\frac{T_i}{n_i} \right)^2 - 2 \frac{T}{n} \sum_{i=1}^k n_i \bar{y}_i + \frac{T^2}{n^2} \sum_{i=1}^k n_i = \sum_{i=1}^k \frac{T_i^2}{n_i} - \frac{T^2}{n}$, so

$$SS_{Tr} = A - \frac{T^2}{n}. \quad (24)$$

Similarly,

$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j}^2 - 2\bar{y}_i y_{i,j} + \bar{y}_i^2) = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j}^2 - 2 \sum_{i=1}^k \bar{y}_i T_i + \sum_{i=1}^k n_i \left(\frac{T_i}{n_i} \right)^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j}^2 - \sum_{i=1}^k \frac{T_i^2}{n_i}, \text{ so}$$

$$SSE = B - A. \quad (25)$$

Finally, as a check, the total variation must be the sum of treatment and error variations.

$$\begin{aligned} SS_y &= \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{i,j} - \bar{y})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j}^2 - 2\bar{y} \sum_{i=1}^k \sum_{j=1}^{n_i} y_{i,j} + \bar{y}^2 n = B - \frac{T^2}{n} \\ &= \left(A - \frac{T^2}{n} \right) + (B - A) = SS_{Tr} + SSE \end{aligned}$$

From equations (13) and (14), the mean squares are given by the following.

The Treatment or Between Mean Square: $MS_{Tr} = \frac{SS_{Tr}}{k-1} = \frac{A - \frac{T^2}{n}}{k-1}$ (26)

The Error or Within Mean Square: $MSE = \frac{SSE}{n-k} = \frac{B - A}{n-k}$ (27)

So a scheme to calculate a one-way classification analysis of variance is to lay out the data in columns as in a spreadsheet, with each column representing a different treatment or factor level. Leave room between treatments for a second column which is the square of the first column. Thus, each treatment is associated with a pair of columns. Sum the column of scores to obtain T_i and sum the column of squares of scores to obtain B_i . Then for each treatment, calculate the

sample mean, \bar{y}_i , $\frac{T_i^2}{n_i}$ and the sample variance $s_i^2 = \frac{B_i - \frac{T_i^2}{n_i}}{n_i - 1}$. Then sum over the treatment groups to obtain T , B , A , and \bar{y} . **Note:** The table below should not be interpreted as implying that the number of rows each column is the same. In general, it is **not true** that $n_1 = n_2 = n_3 = \dots = n_k$.

Treatment 1		Treatment 2		Treatment 3		...	Treatment k	
$y_{1,1}$	$y_{1,1}^2$	$y_{2,1}$	$y_{2,1}^2$	$y_{3,1}$	$y_{3,1}^2$		$y_{k,1}$	$y_{k,1}^2$
$y_{1,2}$	$y_{1,2}^2$	$y_{2,2}$	$y_{2,2}^2$	$y_{3,2}$	$y_{3,2}^2$		$y_{k,2}$	$y_{k,2}^2$
$y_{1,3}$	$y_{1,3}^2$	$y_{2,3}$	$y_{2,3}^2$	$y_{3,3}$	$y_{3,3}^2$		$y_{k,3}$	$y_{k,3}^2$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots		\vdots	\vdots
y_{1,n_1}	y_{1,n_1}^2	y_{2,n_2}	y_{2,n_2}^2	y_{3,n_3}	y_{3,n_3}^2		y_{k,n_k}	y_{k,n_k}^2

T_1	B_1	T_2	B_2	T_3	B_3	T_k	B_k	
\bar{y}_1		\bar{y}_2		\bar{y}_3		\bar{y}_k		
s_1^2		s_2^2		s_3^2		s_k^2		

For each pair of columns: $T_i = \sum_{j=1}^{n_i} y_{i,j}$ $B_i = \sum_{j=1}^{n_i} (y_{i,j})^2$

Summing over the columns: $A = \sum_{i=1}^k \frac{T_i^2}{n_i}$ $B = \sum_{i=1}^k B_i$ $T = \sum_{i=1}^k T_i$ $\bar{y} = \frac{T}{n}$

At this point we can construct an ANOVA table.

Source	Sum of Squares	Degrees of Freedom	Mean Square
Treatment	$SS_{Tr} = A - T^2/n$	$k - 1$	$MS_{Tr} = SS_{Tr}/(k - 1)$
Error	$SSE = B - A$	$n - k$	$MSE = SSE/(n - k)$
Total	$SS_y = B - T^2/n$	$n - 1$	$s_y^2 = \frac{SS_y}{n - 1}$

Compute $F = \frac{MS_{Tr}}{MSE}$ and compare its value against $F_\alpha(\nu_1 = k - 1, \nu_2 = n - k)$ for a stated level of significance, α .

If $F < F_\alpha(k - 1, n - k)$, we fail to reject $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$.

If $F > F_\alpha(k - 1, n - k)$, we reject H_0 and conclude that at least two of the population means are different.

To determine which treatment means are different more testing needs to be done. Some procedures which are used are Scheffe's test and Tukey's test. However, the method we will use is multiple - t testing, also known as the Bonferroni procedure. There are $\binom{k}{2} = \frac{k(k-1)}{2}$ distinct pairs of treatments which can be picked from the k different treatment groups. For each such pair an observed t - score is calculated. If this observed score is larger in absolute value than a critical t - score, then the null hypothesis that asserts the equality of the two population means associated with the two treatments is rejected. Specifically, let ℓ and m stand for the treatment group indices of the two treatments being examined. Then,

$$t_{\text{obs}}(\ell, m) = \frac{|\bar{y}_\ell - \bar{y}_m|}{s_p \sqrt{\frac{1}{n_\ell} + \frac{1}{n_m}}} \quad (28)$$

is the by now familiar observed t - score for comparing two independent sample means. The absolute value is used since this is a two-sided test for a difference between two populations. Since we have already assumed that the k treatment populations have measurements which are normally distributed with a common variance, we use the pooled variance which best estimates this common **within** treatment group variance, namely MSE .

$$s_p^2 = MSE = \frac{B - A}{n - k}. \quad (29)$$

Hence, our observed t - score is computed from

$$t_{\text{obs}}(\ell, m) = \frac{|\bar{y}_\ell - \bar{y}_m|}{\sqrt{MSE \left(\frac{1}{n_\ell} + \frac{1}{n_m} \right)}}. \quad (30)$$

Now, one would expect that this $t_{\text{obs}}(\ell, m)$ would be compared against a critical t - score $t_c = t_{\alpha/2}$, where α is the level of significance. However, we want α to be the maximum probability of a Type I error for the **entire** sequence of $m = \binom{k}{2} = \frac{k(k-1)}{2}$ comparisons. If each comparison were made at a level of significance of α , then the probability of a Type I error "somewhere" in the m comparisons made would be $1 - (1 - \alpha)^m$. This probability is larger than α and considerably larger if m is large. To have an actual level of significance of α , we will therefore use $t_c = t_{\alpha/(2m)}$, so that the probability of a Type I error in the m comparisons becomes

$1 - \left(1 - \frac{\alpha}{m}\right)^m \approx \alpha$. Finally, when we find t_c , we need to know the degrees of freedom. Since we are using MSE for s_p^2 , the degrees of freedom is $\nu_2 = n - k$.

Thus, in summary we have

$$t_c = t_{\alpha/(2m)} \text{ with } n - k \text{ degrees of freedom} \quad (31)$$

Consider the following example. Suppose there are five different treatments, i.e., $k = 5$, and suppose we wish to work at a level of significance of 10%, i.e., $\alpha = 0.10$. Then the number of pair wise comparisons is $m = \binom{5}{2} = \frac{5(4)}{2} = 10$. Suppose that the total number of measurements over the five treatments is 26, i.e., $n = 26$. Thus, from equation (31), $t_c = t_{\alpha/(2m)} = t_{.10/20} = t_{.005}$ with $26 - 5 = 21$ degrees of freedom. Hence, $t_c = 2.831$. Now to facilitate the 10 different comparisons, it helps to summarize the results in tabular form. Each open cell in the following table represents one of the 10 possible comparisons.

Treatment	1	2	3	4
2	$t_{\text{obs}}(1, 2)$			
3	$t_{\text{obs}}(1, 3)$	$t_{\text{obs}}(2, 3)$		
4	$t_{\text{obs}}(1, 4)$	$t_{\text{obs}}(2, 4)$	$t_{\text{obs}}(3, 4)$	
5	$t_{\text{obs}}(1, 5)$	$t_{\text{obs}}(2, 5)$	$t_{\text{obs}}(3, 5)$	$t_{\text{obs}}(4, 5)$

In these cells the computed value of t_{obs} for the two treatments being compared is displayed. Whenever this value of t_{obs} exceeds 2.831 one can conclude at a 10% level of significance that the two population means are different.