Compare several group means: ANOVA

Pierre Legendre, Université de Montréal
August 2009

1 - Introduction

Objective: compare the means of several (*k*) groups for a response variable of interest *y*. The groups contain independent observations.

• The name *analysis of variance* (ANOVA) describes the fact that the total, within-group, and among group variances will be computed to test H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, which concerns the means of the groups.

• Instead of using ANOVA, one may be tempted to carry out a series of *t*-tests for all possible pairs of groups. Actually, ANOVA cannot be replaced by a series of *t*-tests because multiple testing significantly increases the probability of type I error if H_0 is true.

Example — Consider 7 groups of observations drawn independently and at random from the same statistical population. H_0 is obviously true.

• 7 (7 – 1)/2 = 21 *t*-tests must be computed to compare all pairs of groups.

• If each *t*-statistic is tested at the $\alpha = 0.05$ significance level, we have, in each case, 5 chances over 100 to reject H_0 when H_0 is true (type I error).

• Using the binomial distribution, the probability of rejecting H_0 at least once in these 21 tests is 0.66 instead of 0.05.

⇒ For a test to be valid, the type I error rate must be $\leq \alpha$.

Analysis of variance has been developed by the British agronomist *Ronald A. Fisher* at the Rothamsted Experimental Station, UK.

The nominal variable (factor) describing the group to which each observation belongs is called the classification criterion. It may represent either a *fixed* or a *random factor.*

• *Fixed factor*: For a particular factor, the *levels* (or *treatments*) of the factor represented in the experiment compose all possible levels of the factor in the population, or at least all those that are of interest in the particular experiment. Fixed factors are usually of interest: one is interested in determining their effect on the response variable.

• *Random factor*: We assume that the levels of a factor in our experiment have been randomly assigned from all possible levels of the factor that exist in the real world. Random factors may have been incorporated simply because they are known to affect the response variable.

The statistical hypotheses are the following for *k* groups for which the response variable *y* has been observed or measured:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. The groups were drawn from the same population or from populations with identical means.

H_1: at least one of the means differs from the others.

After carrying out a global test, a posteriori multiple comparison tests may be conducted to determine which group(s) differ from the others.

Note: we will not compare the variances of *k* groups along *y*. The null hypothesis is not H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$. That hypothesis could be tested using a Bartlett, Levene, or log-ANOVA test of homogeneity of the group variances.

We will, however, use the ratio

$$\frac{\text{Among-group variance}}{\text{Within-group variance}}$$

to compare the *k* means, just like the *t*-test was comparing two means while taking the corresponding within-group variances into account.

Applications: ANOVA is a widely used method to analyze the results of experiments conducted in the lab or the field. Hurlbert* has shown that ecological surveys can often be seen as *mensurative experiments* if they are structured by identifiable factors, and studied by ANOVA. The underlying ecological processes can then be separately studied through *manipulative (controlled) experiments.*

Objective examples:

- Analysis of the effect of a random factor: one often hopes to show that the data support H_0. If the groups do not differ in their means, they can be pooled in subsequent analyses.

- Analysis of the effect of a controlled factor: in most cases, one is hoping to reject H_0 in order to support the hypothesis (H_1) that a portion of the response variable's variation can be explained by the factor.

• Several classification criteria can be considered in the same analysis: next lecture.

* Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54**:187-211.

2 - Notation in single-classification (one-way) anova

Vector **y** represents the response variable, vector **x** the factor (or classification criterion). The values in the response and explanatory vectors are identified by two indices: the first index represents the group {1...*k*}, the second index the replicate within the group. Thus, y_{32} is the value of the response observed for the second replicate of group #3.

Example:

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \\ y_{33} \\ y_{34} \end{bmatrix} = \begin{bmatrix} 2.45 \\ 1.07 \\ 3.42 \\ 4.97 \\ 4.68 \\ 8.92 \\ 6.51 \\ 7.20 \\ 9.43 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ x_{21} \\ x_{22} \\ x_{31} \\ x_{32} \\ x_{33} \\ x_{34} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

In the representation in the centre, **x** contains a group number for each observation. Before ANOVA in R, the classification criterion coded in that way must be declared as a factor by the command: x = as.factor(x).

Vector **x** could equally well be coded by a series of binary (or *dummy*) variables, as shown in the right-hand column. With that coding, the variation of **y** explained by **x** can be analysed by regression.

3 - Sources of variation

• Total dispersion (or *sum-of-squares*) = TSS

• Within-group (or *residual*) dispersion (or *sum-of-squares*) = RSS

• Among-group dispersion (or *sum-of-squares*) = ASS

Estimate the total dispersion, TSS

TSS = sum of squares of the differences to the overall mean $\bar{y}$, without consideration for the groups ($j = 1 \ldots k$) to which the data belong.

$$\text{TSS} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where $n = \sum n_j$. Since $\nu_{\text{Tot}} = n - 1$, we can use TSS to compute the total variance, $\text{Var}_{\text{Tot}} = s_y^2 = \frac{\text{TSS}}{n-1}$. TSS can be transformed as follows:

$$\text{TSS} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j + \bar{y}_j - \bar{y})^2$$

$$\text{TSS} = \sum\sum [\,(y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y})\,]^2 \qquad \text{Form: } [a + b]^2$$

$$\text{TSS} = \sum\sum (y_{ij} - \bar{y}_j)^2 + 2\sum\sum (y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y}) + \sum\sum (\bar{y}_j - \bar{y})^2$$

$$\text{TSS} = \sum\sum (y_{ij} - \bar{y}_j)^2 + 2\sum_j (\bar{y}_j - \bar{y}) \sum_i (y_{ij} - \bar{y}_j) + \sum\sum (\bar{y}_j - \bar{y})^2$$

⇒ The sum of the differences to the mean of a group is zero, by definition of the group mean. Hence the middle element of the above equation is zero. It follows that

$$\text{TSS} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2$$

$$\text{TSS} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

Estimate the within-group (or *residual*) dispersion, RSS

The total variation [sum of (differences to the mean)^2] within the groups is not of primary interest. It is seen as experimental variation, called "error" by statisticians. In ecology, the within-group variation corresponds to "local innovation" at the individual study sites.

For each group *j*, we compute $\text{RSS}_j = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$

Summing these terms over all groups *j*, we obtain:

$$\text{RSS} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 = \sum_j \left[\sum_i y_{ij}^2 \right] - \sum_j \left(\frac{T_j^2}{n_j} \right)$$

where T_j is the sum of the observed values within group *j*.

Degrees of freedom: $\nu_R = (n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1) = n - k$

Within-group (*residual*) mean square (or variance): $\text{MS}_\text{R} = \frac{\text{RSS}}{n-k} = \text{Var}_\text{R}$

Estimate the among-group dispersion, ASS

• For each group *j*, we compute the squared difference between the mean of that group and the overall mean, $(\bar{y}_j - \bar{y})^2$, then we sum these values over all groups.

• Before summing, each sum-of-squares *j* will be weighted by the number of elements in group *j*. Thus, if there are n_j observations in group *j*, the dispersion due to that group is $n_j (\bar{y}_j - \bar{y})^2$.

• For *k* groups,

$$\text{ASS} = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2 = \sum_{j=1}^{k} \left(\frac{T_j^2}{n_j} \right) - \frac{T^2}{n}$$

The number of degrees of freedom associated with a statistic is the number of independent components, i.e. the number of basic components in the calculation minus the number of parameters linking them.

- The number of basic components are the *k* deviations $\bar{y}_j - \bar{y}$.

- The *k* means are linked by a single parameter: the general mean $\bar{y}$.

Hence $\nu_A = k - 1$

and the among-group mean square (or variance) is: $\text{MS}_A = \frac{\text{ASS}}{k-1} = \text{Var}_A$

Interesting relationships: TSS = RSS + ASS

and $\nu_{Tot} = n - 1 = (n - k) + (k - 1) = \nu_R + \nu_A$

Example — Hypothetical experiment: 15 subjects were treated for headache using 3 pain relievers. The effect was noted on a 0 to 5 scale.

Aspirin	Paracetamol	Placebo
3	2	2
5	4	1
4	4	3
5	5	2
5	3	1

The calculation results are presented in an analysis of variance table:

	d.f.	Sums of squares	Mean squares
Brand	2	ASS = 17.7333	$MS_A = 8.8667$
Residual	12	RSS = 11.2000	$MS_R = 0.9333$
Total	14	TSS = 28.9333	$Vat_{Tot} = 2.0667$

Mean squares, abbreviated MS, are variances.

Beware: $MS_A + MS_R \neq Vat_{Tot}$ (i.e., $8.8667 + 0.9333 \neq 2.0667$),
even though ASS + RSS = TSS (17.7333 + 11.2000 = 28.9333).

4 - Two estimates of σ^2 under H_0

The reasoning presented in this section will allow the construction of a test of significance for the differences among the group means.

Assume that the *k* populations, from which the *k* groups of observations were drawn, are normally distributed and that they all have the same variance σ^2 (i.e., $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$).

If H_0 is true (H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$), then the common variance σ^2 can be estimated in two different ways.

First method to estimate σ^2

A fundamental hypothesis of ANOVA is that each of the group variances σ_j^2 estimates the same common variance σ_y^2. This allows us to look for a robust estimate of the common variance by computing the weighted mean of the *k* within-group variances.

⇒ This step introduces the hypothesis of homogeneity of the group variances in the ANOVA test. That hypothesis has to be checked using an appropriate test prior to ANOVA (see section 6).

Variance of a group *j*, weighted by the number of degrees of freedom of that group:

$$(n_j - 1)\frac{\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{(n_j - 1)} = \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

Mean of the weighted variances of the *k* groups:

$$\frac{\sum (y_{i1} - \bar{y}_1)^2 + \ldots + \sum (y_{ik} - \bar{y}_k)^2}{(n_1 - 1) + \ldots + (n_k - 1)}$$

$$\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{(n-k)} = \frac{\text{RSS}}{n-k} = \text{MS}_\text{R}$$

So, MS_R is an estimate of the common variance σ^2, for H_0 true or not.

Second method to estimate σ^2

If H_0 is true, the means $\bar{y}_j$ of the *k* groups are all estimates of the same common mean μ. The variance of these estimates of μ can be written:

$$s^2_{\bar{y}} = \frac{\sum_j (\bar{y}_j - \bar{y})^2}{(k-1)}$$

The square root of that variance estimates the standard error of the mean.

One can also estimate the standard error of the mean from the standard deviation of the data from a single group: $s_{\bar{y}} = s_{y_j} / \sqrt{n_j}$

which can be squared: $s^2_{\bar{y}} = s^2_{y_j} / n_j$

If H_0 is true, one can estimate the population variance σ^2_y by:

$$s^2_{y_j} = n_j s^2_{\bar{y}} = n_j \frac{\sum_j (\bar{y}_j - \bar{y})^2}{(k-1)}$$

n_j can be incorporated within the sum to obtain the following estimate of the common variance:

$$s^2_y = n_j s^2_{\bar{y}} = \frac{\sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2}{(k-1)} = \frac{\text{ASS}}{k-1} = \text{MS}_\text{A}$$

So, MS_A is an estimate of the common variance σ^2 if H_0 is true.

In summary: if H_0 is true and the groups of observations are drawn from the same statistical population, or from populations having the same mean μ and same variance σ^2, then MS_R and MS_A are both estimates of σ^2. These estimates should be nearly equal.

5 - Comparison test

• If H_0 is true (H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$), MS_R and MS_A represent two estimates of σ^2. *Their ratio is expected to be near 1*.

• In all cases, MS_R is an estimate of σ^2. One had to check the equality of the variances of the populations from which the *k* groups have been drawn (condition of homogeneity of the variances or *homoscedasticity*: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$) before the ANOVA.

• If H_1 is true, the among-group variance MS_A is not an estimate of σ^2. Indeed, in that case, the distribution of the means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ does not represent the sampling distribution of a common mean μ.

⇒ In that case, the distribution of the means $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$ is wider and more flat than the sampling distribution of the common mean μ. MS_A is then necessarily larger than MS_R.

• MS_R and MS_A are two independent components of the total variance since TSS = RSS + ASS. If H_0 is true, their ratio (which is near 1) is a test-statistic distributed like the Fisher-Snedecor *F* distribution:

$$F_A = \frac{MS_A}{MS_R} \qquad (13\text{-}30)$$

The degrees of freedom of the numerator and denominator are respectively: $\nu_1 = k - 1$ and $\nu_2 = n - k$. Since MS_A is the largest of the two mean squares if H_1 is true, that value is placed in the numerator in order to obtain a value of $F_A > 1$ in that case.

• The test is one-tailed in all cases. The reason is the following:

If H_0 is true, $MS_A \approx MS_R$ and hence $F_A \approx 1$;

if H_1 is true, $MS_A > MS_R$ and hence $F_A > 1$.

Decision rules

• Do not reject H_0 if $F_A < F_\alpha$ where F_α is the critical value of F at significance level α (e.g. 5%), or if the associated p-value > α.

• Reject H_0 if $F_A \geq F_\alpha$, or if the associated p-value $\leq \alpha$.

Example revisited:

Here is the complete analysis of variance table.

	d.f.	**Sums of squares**	**Mean squares**	***F***	**p-value**
Brand	2	ASS = 17.7333	$MS_A = 8.8667$	9.5	0.00336
Residual	12	RSS = 11.2000	$MS_R = 0.9333$		
Total	14	TSS = 28.9333	$Vat_{Tot} = 2.0667$		

R language: functions *aov* and *summary.* The classification criterion must be declared as a factor by the command: x = as.factor(x).

6 - Conditions of application of the *F*-test in ANOVA

- Response variable quantitative (to be able to compute $\bar{y}$ and s_y).

- Independence of the observations (not autocorrelated).

- The population from which each group is drawn must be normal.

- Equality of the variances (homoscedasticity) of the groups. Reason: as in the two-group case, the *F*-test is simultaneously testing two different null hypotheses: the equality of means **and** the equality of variances (*Behrens-Fisher problem*). Tests for equality of the variances in the R language: Bartlett test (*bartlett.test*: more power to detect small differences), Levene test (*levene.test*: good compromise between power and robustness when the data are not quite normal).

• Violation of the conditions of application:

- Homoscedasticity: the ANOVA *F*-test is robust in the presence of a certain amount of heteroscedasticity. The results thus remain valid if the within-group variances are not quite homogeneous.

- Normality: the ANOVA *F*-test is also robust in the presence of a certain amount of skewness (asymmetry) and kurtosis of the distributions. For skewness (α_3), the following criterion can be used: the *F*-test can be used is $n_j \geq 25\,(\alpha_3^2)_j$.

- Strong violation of the normality condition:

1. Transform the data before analysis.
2. Test F_A by permutations. R functions are available on the Web page http://www.bio.umontreal.ca/legendre/indexEn.html .
3. Use the nonparametric Kruskal-Wallis test (R language: function *kruskal.test*).

- If the observations are spatially or temporally correlated, the test is either too liberal (rejecting H_0 too often) or too conservative (the opposite), depending on the type of spatial dependence among the observations. See Legendre et al. (2004), *Ecology* 85: 3202-3214. One must control for the spatial structure in a modified form of the test.

7 - Alternative computation methods

If ANOVA is used to test the difference between the means of two groups, the result of the *F*-test is the same as that of a two-tailed *t*-test without Welch correction.

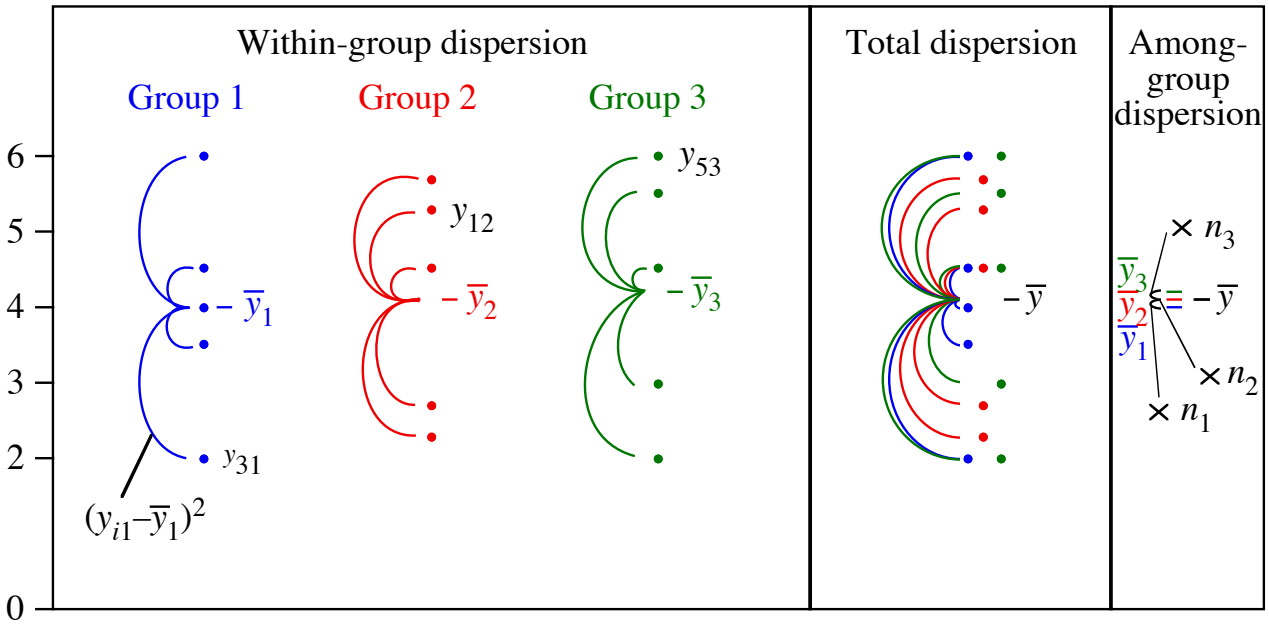
An ANOVA *F*-test is equivalent to the *F*-test of the coefficient of determination in a multiple regression between the response variable and a series of dummy variables representing the groups, shown in section 2 (right-hand representation of **x**). See *Practicals*.

One-way ANOVA. Example 1: $\mathbf{H_0}$ is true

Classification criterion →

Observations {

Group 1	Group 2	Group 3
4.0	5.3	2.0
6.0	2.7	3.0
2.0	4.5	4.5
4.5	2.3	5.5
3.5	5.7	6.0



$n_1 = 5$ $T_1 = 20.0$ $\bar{y}_1 = 4.0$ $\Sigma(y_{i1}-\bar{y}_1)^2 = 8.50$

$n_2 = 5$ $T_2 = 20.5$ $\bar{y}_2 = 4.1$ $\Sigma(y_{i2}-\bar{y}_2)^2 = 9.36$

$n_3 = 5$ $T_3 = 21.0$ $\bar{y}_3 = 4.2$ $\Sigma(y_{i3}-\bar{y}_3)^2 = 11.30$

$n = 15$ $T = 61.5$ $\bar{y} = 4.1$ $\Sigma(y_{ij}-\bar{y})^2 = 29.26$ → **TSS**

$\Sigma n_j(\bar{y}_j-\bar{y})^2 = 0.10$ → **ASS**

RSS = 29.16

TSS = ASS + RSS

Sources of variation	Dispersions (SS)	Degrees of freedom	Mean squares (variances)
Total	TSS = 29.26	15 – 1 = 14	29.26 / 14 = 2.09
Among-group	ASS = 0.10	3 – 1 = 2	MS_A = 0.10 / 2 = 0.05
Within-group	RSS = 29.16	15 – 3 = 12	MS_R = 29.16 / 12 = 2.43

$F = MS_A/MS_R = 0.0206$ $P = 0.9797$

$F_{(0.05,\,2,\,12)} = 3.89$

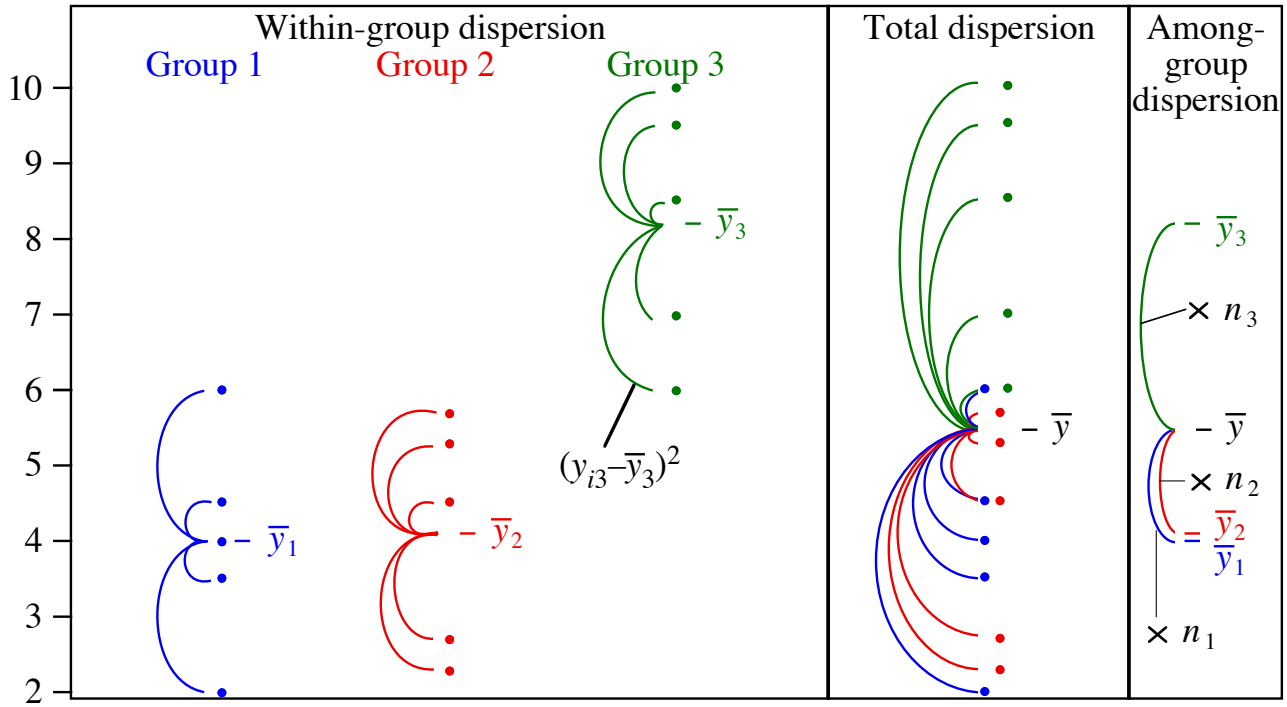
Daniel Borcard, 2004

One-way ANOVA. Example 2: $\mathbf{H_0}$ is false

Critère de classification →

Observations

Group 1	Group 2	Group 3
4.0	5.3	6.0
6.0	2.7	7.0
2.0	4.5	8.5
4.5	2.3	9.5
3.5	5.7	10.0



$n_1 = 5$ $T_1 = 20.0$ $\bar{y}_1 = 4.0$ $\Sigma(y_{i1}-\bar{y}_1)^2 = 8.50$

$n_2 = 5$ $T_2 = 20.5$ $\bar{y}_2 = 4.1$ $\Sigma(y_{i2}-\bar{y}_2)^2 = 9.36$

$n_3 = 5$ $T_3 = 41.0$ $\bar{y}_3 = 8.2$ $\Sigma(y_{i3}-\bar{y}_3)^2 = 11.30$

$n = 15$ $T = 61.5$ $\bar{y} = 5.43$ $\Sigma(y_{ij}-\bar{y})^2 = 86.59$ → **TSS**

$\Sigma n_j(\bar{y}_j-\bar{y})^2 = 57.43$ → **ASS**

RSS = 29.16

TSS = ASS + RSS

Sources of variation	Dispersions (SS)	Degrees of freedom	Mean squares (variances)
Total	TSS = 86.59	15 – 1 = 14	86.59 / 14 = 6.19
Among-group	ASS = 57.43	3 – 1 = 2	MS_A = 57.43 / 2 = 28.72
Within-group	RSS = 29.16	15 – 3 = 12	MS_R = 29.16 / 12 = 2.43

$F = MS_A/MS_R = 11.82$ $P = 0.0015$

$F_{(0.05,\,2,\,12)} = 3.89$ $F_{(0.01,\,2,\,12)} = 6.93$

Daniel Borcard, 2004