# [STAT-17] One-way ANOVA

**Miguel-Angel Canela**
**Associate Professor, IESE Business School**

## The $F$ distribution

Some popular tests, used in the analysis of variance and linear regression, are based on the ratio of two sample variances. They are based on a well known model for the distribution of these ratios, the $F$ **distribution**.

Take two independent samples with distributions $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$. Then, the ratio of the sample variances,

$$F = \frac{S_1^2}{S_2^2}\,,$$

has an $F$ distribution with $(n_1, n_2)$ degrees of freedom, in short $F(N_1, n_2)$. The general formula for the density (see Figure 1 for a graphical example) is

$$f(x) = \frac{\Gamma\big(n_1/2 + n_2/2\big)\, n_1^{n_1/2}\, n_2^{n_2/2}}{\Gamma\big(n_1/2\big)\,\Gamma\big(n_2/2\big)}\,\frac{x^{n_1/2-1}}{\big(n_2 + n_1 x\big)^{(n_1+n_2)/2}}, \quad x > 0.$$

The first factor is a normalization constant. $n_1$ and $n_2$ are the parameters of this model. When it is used as a model for a ratio of two sample variances, $n_1$ is associated to the numerator and $n_2$ to the denominator. We denote by $F_\alpha$ the critical value associated to right tail. More explicitly, $\mathrm{p}\big[F > F_\alpha\big] = \alpha$.

## Distributions derived from the normal

As the $\chi^2$ and the $t$ distributions, the $F$ distributions is presented in textbooks as a **distribution derived from the normal**. The $F$ distribution can be related to the other two in two ways:

- Assuming independence, $X_1 \sim \chi^2(n_1)$ and $X_2 \sim \chi^2(n_2)$ imply $\dfrac{X_1}{X_2} \sim F(n_1, n_2)$.
- If $X \sim t(n)$, then $X^2 \sim F(1, n)$.

It is important to keep in mind this second property, which implies that any $t$ test can be seen as an $F$ test. The only thing lost in taking the squares is the sign. A consequence of this relationship is that

$$|t(n)| > c \iff F(1, n) > c^2$$

or, equivalently, that the respective critical values satisfy $t_{\alpha/2}(n)^2 = F_\alpha(1, n)$.

## The one-way ANOVA $F$ test

I present in this section the extension of the two-sample $t$ test to $k$ (independent) samples. It applies to a null $H_0:\ \mu_1 = \cdots = \mu_k$, and it is one of the many variants of **analysis of variance** (ANOVA), based on decomposing sum of squares. This lecture is covers only the analysis of variance of one factor, or **one-way ANOVA**.

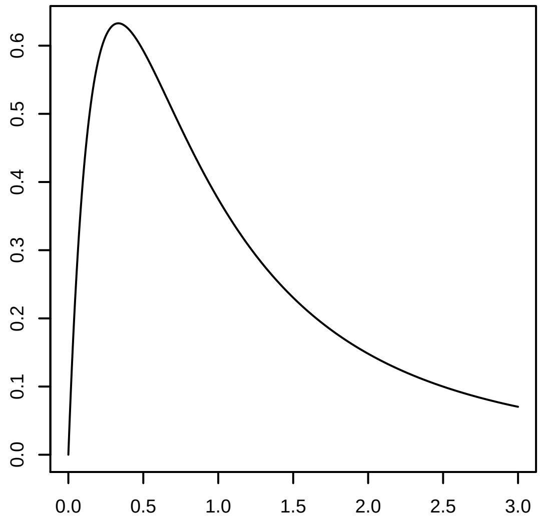

**Figure 1.** $F(4,4)$ **density curve**

Suppose now $k$ independent samples, of sizes $n_1, \dots, n_k$, and let $n = n_1 + \cdots + n_k$ be the total sample size. The data can be arranged as in Table 1, where each group takes a column (the columns can have different lengths) and the last row contains the group means.

**TABLE 1. Data for a one-way ANOVA test**

| Group 1 | Group 2 | $\cdots$ | Group $k$ |
|---|---|---|---|
| $x_{11}$ | $x_{21}$ | $\cdots$ | $x_{k1}$ |
| $x_{12}$ | $x_{22}$ | $\cdots$ | $x_{k2}$ |
| $\vdots$ | $\vdots$ | | $\vdots$ |
| $x_{1n_1}$ | $x_{2n_2}$ | $\cdots$ | $x_{kn_k}$ |
| $\bar{x}_1$ | $\bar{x}_2$ | $\cdots$ | $\bar{x}_k$ |

The **within-group** and the **between-groups sum of squares** are defined as

$$\mathrm{SSW} = \sum_{j=1}^{n_1} \big(x_{1j} - \bar{x}_1\big)^2 + \cdots + \sum_{j=1}^{n_k} \big(x_{kj} - \bar{x}_k\big)^2, \qquad \mathrm{SSB} = n_1\big(\bar{x}_1 - \bar{x}\big)^2 + \cdots + n_k\big(\bar{x}_k - \bar{x}\big)^2.$$

Presenting the data as in Table 1 helps to understand these sums of squares. We can consider two different sources of variability in this table. Vertically, we see the variability within the groups, measured by SSW. Horizontally, in the means of the last row, we see the variability between the groups, measured by SSB.

Since the grand mean $\bar{x}$ is the weighted average of the group means $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$, SSB has $k-1$ independent terms. We say that it has $k-1$ degrees of freedom. SSW is composed of $k$ blocks. The first block has $n_1 - 1$ independent terms, so $n_1 - 1$ degrees of freedom. Counting degrees of freedom in the same way for the other groups, the number of degrees of freedom of SSW is

$$\big(n_1 - 1\big) + \big(n_2 - 1\big) + \cdots + \big(n_k - 1\big) = n - k.$$

Now, the **one-way ANOVA** $F$ **statistic** is defined as

$$F = \frac{\mathrm{SSB}/(k-1)}{\mathrm{SSW}/(n-k)}\,.$$

It can be proved that, under the null, this statistic has an $F(k-1, n-k)$ distribution, so the (right) tail area provides a $p$-vlue for testing the null. This is the **one-way ANOVA $F$ test**.

**Example 1.** The `jobsat1` data set contains data on job satisfaction (average of a 12-item Likert scale) from three countries, Chile (CH), Mexico (ME) and Spain (SP). The group statistics are reported in Table 2. Here,

$$\mathrm{SSB} = 121\big(4.158 - 4.227\big)^2 + 111\big(4.413 - 4.227\big)^2 + 191\big(4.162 - 4.227\big)^2 = 5.217,$$

$$\mathrm{SSW} = 120 \times 0.902^2 + 110 \times 0.865^2 + 190 \times 0.814^2 = 305.6.$$

**TABLE 2. Group statistics (Example 1)**

| Statistic | Chile | Mexico | Spain | Total |
|---|---|---|---|---|
| Size | 121 | 111 | 191 | 423 |
| Mean | 4.158 | 4.413 | 4.162 | 4.227 |
| Stdev | 0.902 | 0.865 | 0.814 | 0.858 |

The $F$ statistic is

$$F = \frac{5.217/2}{305.6/420} = 3.58 \quad (P = 0.029).$$

So, we reject the null, concluding that are differences between countries.

## The ANOVA table and the analysis of residuals

The ANOVA table (Table 3) is a classical presentation of the $F$ test, which illustrates the steps to be followed in order to obtain the $F$ value. It is based on the decomposition SST = SSB + SSW, which on the left side has the **total sum of squares**,

$$\mathrm{SST} = \sum_{i,j} \big(x_{ij} - \bar{x}\big)^2.$$

**TABLE 3. 1-way ANOVA table**

| Source | Sum of squares | Degrees of freedom | Mean square | $F$ statistic | P-value |
|---|---|---|---|---|---|
| Between-groups | SSB | $k-1$ | MSB | $F$ | $P$ |
| Within-groups | SSW | $N-k$ | MSW | | |
| Total | SST | $N-1$ | | | |

A number of degrees of freedom is assigned to each sum of squares, with the logic used in the preceding section. Next, a **mean square** (MS) is calculated, dividing the sums of squares by their respective numbers of degrees of freedom. The $F$ statistic is the ratio

$$F = \frac{\mathrm{MSB}}{\mathrm{MSW}}\,.$$

In one-way ANOVA, the conditions for the validity of the $F$ test are the same as in the two-sample $t$ test. The data set is partitioned into $k$ groups, which are assumed to be independent samples of

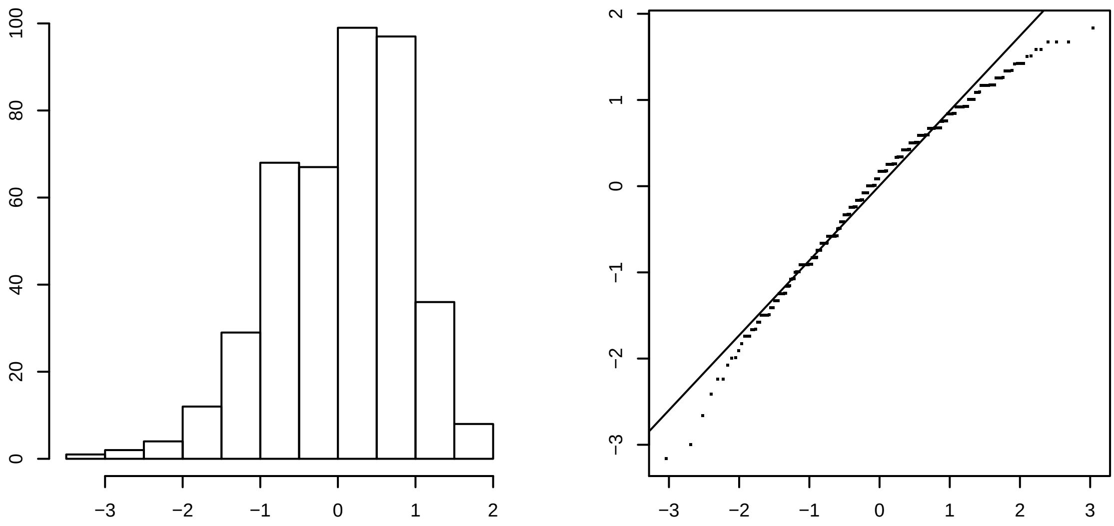


**Figure 2. Distribution of the residuals (Example 1)**

$\mathcal{N}(\mu_i, \sigma^2)$ distributions $i = 1, \ldots, k$. Whether these assumptions are acceptable is usually checked through the **residuals**. In one-way ANOVA, the residuals are the deviations with respect to the group means, $e_{ij} = x_{ij} - \bar{x}_i$. If the one-way ANOVA assumptions were valid, the residuals should look as a random sample of the $\mathcal{N}(0, \sigma^2)$ distribution, and this is what we check in practice. The assumption that the variance is the same for all samples is called **homoskedasticity**. This assumption will be discussed later in this course.

**Example 1 (continuation).** The ANOVA table corresponding to Example 1 is Table 4. You can see the histogram and the normal probability plot of the residuals in Figure 2. The skewness is –0.5. So far, the normality assumption is not clear at all, but you should not worry about the validity of the conclusion, since, with such samples sizes, the $F$ test is safe enough.

**TABLE 4. ANOVA table (Example 1)**

| Source | Sum of squares | Degrees of freedom | Mean square | $F$ value | Significance level |
|---|---|---|---|---|---|
| Between-groups | 5.217 | 2 | 2.609 | 3.58 | 0.029 |
| Within-groups | 305.6 | 420 | 0.728 | | |
| Total | 310.8 | 422 | 0.736 | | |

## Homework

**A.** The `jobsat2` data set comes from the same study as Example 1, but includes data from nine countries. Test the differences among countries using the methods of this lecture. The same for genders.