[STAT-16] 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_α the critical value associated to right tail. More explicitly, $\mathrm{p}\big[F > F_\alpha\big] = \alpha$.

Distributions derived from the normal

Together the χ^2 and the t distributions, the F distributions are presented in textbooks as **distributions 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

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

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**

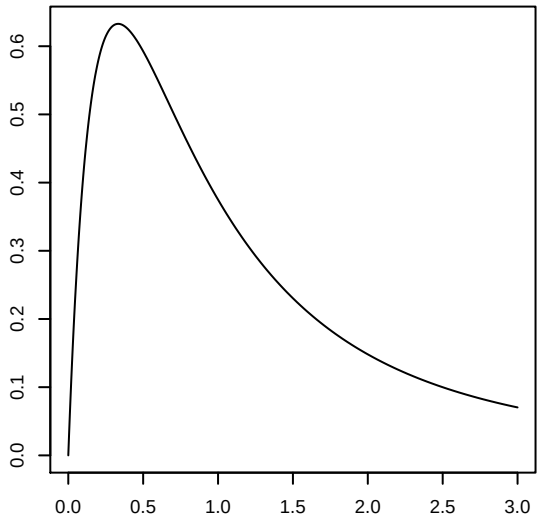

Figure 1. $F(4,4)$ **density curve**

(ANOVA), based on decomposing sum of squares. This lecture is covers only the analysis of variance of one factor, or **one-way ANOVA**.

In the two-sample t test, the data set was composed by two independent groups. As the square of a variable with a $t(n-2)$ distribution has an $F(1,n-2)$ distribution, I denote the squared t statistic by F. It is easy to check that

$$\frac{n_1 n_2}{n}\big(\bar{x}_1 - \bar{x}_2\big)^2 = n_1\big(\bar{x}_1 - \bar{x}\big)^2 + n_2\big(\bar{x}_2 - \bar{x}\big)^2$$

and, therefore, the squared t statistic can be written as

$$F = \big(n-2\big)\frac{n_1\big(\bar{x}_1 - \bar{x}\big)^2 + n_2\big(\bar{x}_2 - \bar{x}\big)^2}{\displaystyle\sum_{j=1}^{n_1}\big(x_{1j} - \bar{x}_1\big)^2 + \sum_{j=1}^{n_2}\big(x_{2j} - \bar{x}_2\big)^2}\,.$$

Note that the numerator of this fraction is a sum of squares related to **between-group variation**, whereas the denominator is related to **within-group variation**. The number of summands is the same, n. The advantage of this expression is that it can be easily generalized to the case of k samples, leading to the **one-way ANOVA** F **test**.

Suppose now k independent samples, of sizes $n_1, \ldots, n_k$, 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.

The denominator in the above formula is equal to the sum of squares within sample 1, plus the sum of squares within sample 2. To generalize this, we define the **within-group sum of squares** as

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

The numerator can also be regarded as a sum of squares, repeated so that the number of summands is the same as in the denominator. The generalization to k groups is straightforward, leading to the **between-groups sum of squares**,

$$\mathrm{SSB} = n_1\big(\bar{x}_1 - \bar{x}\big)^2 + \cdots + n_k\big(\bar{x}_k - \bar{x}\big)^2.$$

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.

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 general one-way ANOVA F statistic is defined as

$$F = \frac{n-k}{k-1}\,\frac{\text{SSB}}{\text{SSW}} = \frac{n-k}{k-1}\,\frac{n_1\big(\bar{x}_1-\bar{x}\big)^2+\cdots+n_k\big(\bar{x}_k-\bar{x}\big)^2}{s^2}\,.$$

Note that, for $k=2$, the factor $k-1$ can be omitted, leading to the formula given above. Under the null, this F statistic has an $F(k-1, n-k)$ distribution, which can be used to calculate a p-value.

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.

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 pooled variance is

$$s^2 = \frac{120\times 0.902^2 + 110\times 0.865^2 + 190\times 0.814^2}{420} = 0.728,$$

and the F statistic,

$$F = \frac{420}{2}\,\frac{121\big(4.158-4.227\big)^2+111\big(4.413-4.227\big)^2+191\big(4.162-4.227\big)^2}{0.728} = 3.58.$$

With $(2, 429)$ degrees of freedom, this leads to $P = 0.029$ and, therefore, to the rejection of the null. We can conclude, then, 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 $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW}$, which on the left side has the **total sum of squares**,

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

A number of degrees of freedom is assigned to each sum of squares. Roughly speaking, it is the number of independent terms in the sum. Since $\bar{y}$ is the mean of all the observations y_{ij}, SST has $n-1$ degrees of freedom. In SSB, there are k different terms, but the sum of the k deviations $\bar{y}_i - \bar{y}$ is zero, so that SSB has $k-1$ degrees of freedom. Finally, the contribution of the ith group to the within-group sum of squares has $n_i - 1$ degrees of freedom, and, hence, the number of degrees of freedom for SSW is $(n_1 - 1) + \cdots + (n_k - 1) = n - k$. Thus, the $n-1$ degrees of freedom are splitted, $k-1$ going to SSB and $n-k$ to SSW.

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	$N-1$			

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

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

Note that the within-group mean square MSW is equal to the pooled variance s^2. 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 $\mathcal{N}(\mu_i, \sigma^2)$ 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 in depth in the Econometrics 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		

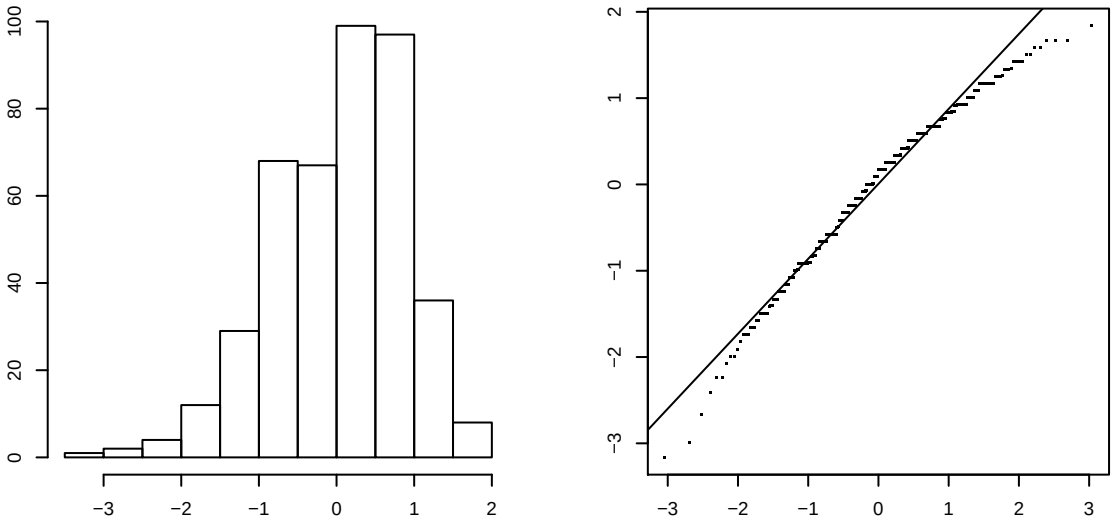

Figure 2. Distribution of the residuals (Example 1)

Homework

A. Draw 250 independent random samples of size 5 from $\mathcal{N}(0,1)$ and calculate the sample variance for each sample. The same for $\mathcal{N}(1,1)$. Divide the first by the second, getting 250 F statistics and plot a histogram. Compare this histogram with Figure 1.

B. 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.