[STAT-09] More on testing between-group differences

Miguel-Angel Canela
Associate Professor, IESE Business School

The F distribution

The F **distribution** is a model for the ratio of two sample variances. Some popular tests, used in the analysis of variance and linear regression, are based on such ratios. So I explain briefly the basic facts about the F distributions before introducing the tests.

To get an F distribution, we 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 probability density function 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. The F distribution is a probability model with two parameters n_1 and n_2. 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. Figure 1 show the $F(4,4)$, $F(2,10)$ and $F(10,2)$ density curves.

The notation of the critical values of the F distribution is consistent with the notation used for the critical values of the normal and the Student t. For $0 < \alpha < 1$, we denote by F_α the $(1-\alpha)$-quantile defined by the formula

$$\mathrm{p}\big[F > F_\alpha\big] = \alpha.$$

Distributions derived from the normal

As the χ^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 $\dfrac{X_1/n_1}{X_2/n_2} \sim F(n_1, n_2)$.
- If $X \sim t(n)$, then $X^2 \sim F(1, n)$.

It is useful to keep in mind this second property, which implies that any t test can be seen as an F test. The only thing lost when 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 $t_{\alpha/2}(n)^2 = F_\alpha(1, n)$.

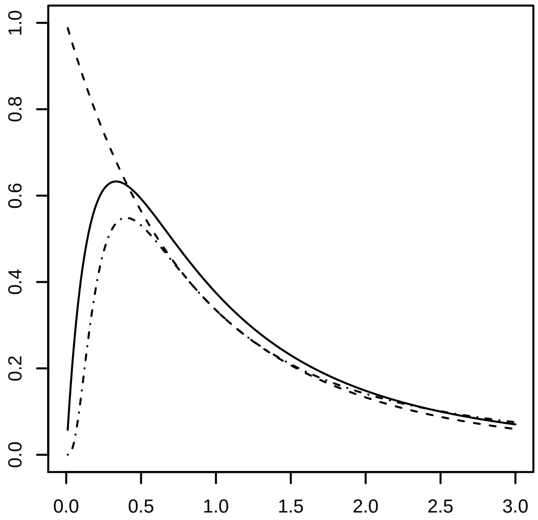

Figure 1. $F(4,4)$, $F(2,10)$ **and** $F(10,2)$ **density curves**

The one-way ANOVA F test

Analysis of variance (ANOVA) is a generic expression which applies to several statistical techniques based on sum of squares. 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, called the **one-way ANOVA**.

Suppose now k independent samples, of sizes n_1, …, 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}
$\bar{x}_1$	$\bar{x}_2$	$\cdots$	$\bar{x}_k$

Presenting the data as in Table 1 helps to understand the sums of squares involved in one-way ANOVA. We can consider two different sources of variability in this table. Vertically, we see the variability within the groups. Horizontally, in the means of the last row, we see the variability between the groups.

The **between-groups sum of squares** is defined as

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

The grand mean $\bar{x}$ is the weighted average of the group means $\bar{x}_1$, $\bar{x}_2$, …, $\bar{x}_k$. So, SSB has $k-1$ independent terms. We say that it has $k-1$ **degrees of freedom**.

The **within-group sum of squares** is defined as

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

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, we define the statistic

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

It can be proved that, under the null, this statistic has an $F(k-1, n-k)$ distribution. We cal it the **one-way ANOVA F statistic**. The (right) tail area of the $F(k-1, n-k)$ distribution provides a p-value for testing the null ($\mu_1 = \cdots = \mu_k$). This is the **one-way ANOVA F test**.

Example 1. The `jobsat1` data set, already used in the preceding lecture, 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

The ANOVA table is a classical presentation of the F test. These tables were designed to illustrate the steps to be followed in order to obtain the F value. Although they have lost their practical interest in the computer age, they still survive in the textbooks and the output of many statistical packages.

Table 3 is a generic one-way ANOVA table. It is based on the decomposition $\mathrm{SST} = \mathrm{SSB} + \mathrm{SSW}$, which on the left side has the **total sum of squares**,

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

TABLE 3. One-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	$N-1$			

In column 2 of the ANOVA table, we put the sum of squares, and in column 3 the number of degrees of freedom, assigned with the logic explained in the preceding section. In column 4, a **mean square** (MS) is calculated, dividing the sums of squares by their respective numbers of degrees of freedom. The F statistic of column 5 is the ratio

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

Analysis of the residuals

In one-way ANOVA testing, the conditions for the validity of the F test are the same as in the two-sample t test, although I do not explain here how to perform the test when equality of variances is not accepted. The data set is partitioned into k groups, which are assumed to be independent samples of $\mathcal{N}(\mu_i, \sigma^2)$ distributions $i = 1, \dots, k$.

In practice, this means two conditions:

- Normality.
- The assumption that the variance is the same for all samples is called **homoskedasticity**.

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, which is what we check in practice.

The normality can be easily explored by means of the usual diagnostic plots, as we see in the example that follows. Homoskedasticity 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 statistic	p-value
Between-groups	5.217	2	2.609	3.58	0.029
Within-groups	305.6	420	0.728		
Total	310.8	422	0.736		

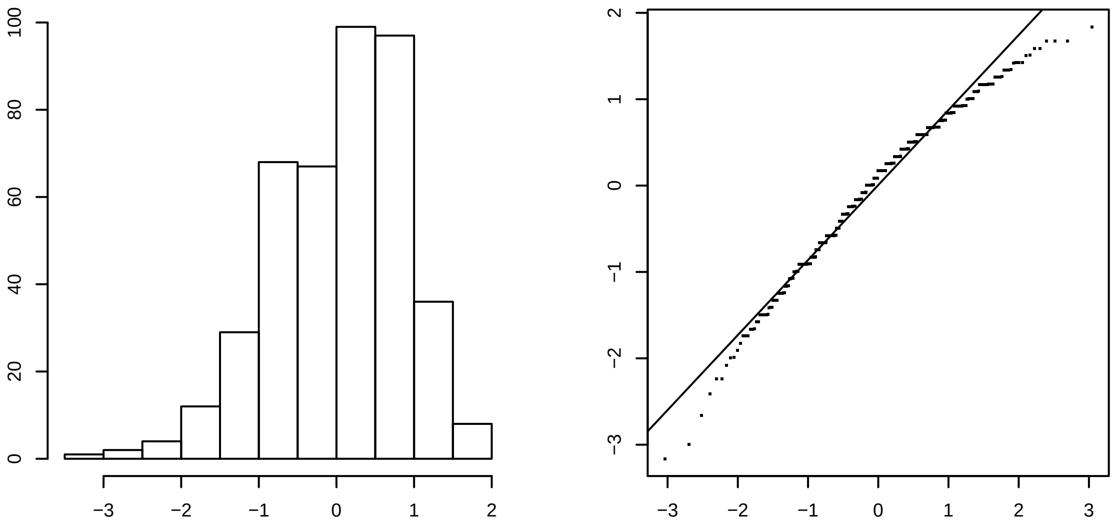

Figure 2. Distribution of the residuals (Example 1)

Nonparametric testing

The tests for mean differences between groups discussed in the last two lectures are valid under normality assumptions and asymptotically valid without them. In general, a **nonparametric test** is one in which it is not assumed that the distribution of the variables involved is of a particular type.

A trivial example of a nonparametric test, the **sign test**, based on the binomial distribution, is an alternative to the one-sample t test (and consequently to the paired data t test), with no distributional assumption. The setting for the test is that of a continuous distribution with median μ and a sample of size n, containing no zeros. The null is $H_0 : \mu = 0$ (for $\mu = \mu_0$, it suffices to subtract μ_0 and perform the test as presented here).

Calling B^+ the number of positive observations and B^- that of negative observations, the test statistic is $B = \max(B^+, B^-)$. The test is based on the fact that, under the null, the probability of a positive result is 0.5. Then B^+ and B^- have a $\mathcal{B}(n, 0.5)$ distribution. The p-value is the double of the probability of the right tail associated to the actual value of B in $\mathcal{B}(n, 0.5)$ distribution. Some stat packages report asymptotic p-values, derived from a normal approximation.

¶ Zero observations are discarded in this test. This is not relevant as far as the continuity assumption, under which repetition is not expected, is tenable.

Example 2. For the `lwages` data set analyzed in the preceding lecture, we find $B = B^+ = 443$ cases, out of $n = 545$, in which the wages have been increased. The two-tail probability, extracted from the $\mathcal{B}(545, 0.5)$ distribution satisfies $P < 0.001$, consistent with the outcome of the t test. An asymptotic p-value can be based on the $\mathcal{N}(272.5, 136.25)$ distribution.

The Wilcoxon signed rank test

The **Wilcoxon signed rank test** is a second alternative to the one-sample t test. The distributional assumptions are the continuity and the symmetry with respect to the mean. I set, as in the preceding section, a null $\mu = 0$. The test statistic is obtained as follows:

- We sort the observations by absolute value. Let me assume first that there are no ties.

- We assign ranks. The first observation gets rank 1, the second one rank 2, etc.
- Calling V^+ and V^- the sum of ranks of the positive and negative observations, respectively, we have $V^+ + V^- = n(n+1)/2$. In the version provided by Stata and R, the test statistic is $V = V^+$, but many textbooks use $V = \max(V^+, V^-)$ to simplify the use of the tables.

Under the null, V_+ has a symmetric (discrete) distribution, with mean and variance given by

$$\mathrm{E}[V_+] = \frac{n\big(n+1\big)}{4}\,, \qquad \mathrm{var}[V_+] = \frac{n\big(n+1\big)\big(2n+1\big)}{24}\,.$$

If there are ties in the absolute values, the tied values get an average rank. The variance must be then corrected. Stat packages usually provide a correction for this case. To get exact p-values for this test and those that follow, one should look at the corresponding tables or use a specialized package (R has an option for exact p-values. What we usually get from generalist stat packages is an **asymptotic p-value**. The difference may be relevant for small sample experimental studies, but not the sample sizes that we usually find in Management Science research. In fact, the tables that we find in textbooks do no go beyond $n = 20$. Asymptotic p-values are based on a normal approximation with the mean and variance given by the above formulas.

Example 2 (continuation). For the `lwages` data set, we get $V = 133,220$ ($z = 15.9$, $P < 0.001$).

The rank-sum test

The **Wilcoxon two-sample rank-sum test** is an alternative to the two-sample t test that only requires that the distributions compared are continuous and of the same type. There are two equivalent versions, that of Wilcoxon, which I present here, and that of Mann and Whitney, sometimes called **Mann-Whitney U test**.

More specifically, it is assumed here that the densities $f_1(x)$ and $f_2(x)$ of the probability distributions compared are related by an equation

$$f_2(x) = f_1(x - \Delta).$$

$\Delta = \mu_1 - \mu_2$ is sometimes called the **treatment effect**. The null is $\Delta = 0$. The test applies to two independent samples of sizes n_1 and n_2, respectively. To simplify the notation, I assume here that $n_1 \le n_2$. The test statistic W is obtained as follows:

- The two samples are merged, and the the resulting sample (size $n_1 + n_2$) is sorted.
- We assign ranks to the observations, averaging ties.
- W is the sum of the ranks of the first sample (`R` subtracts $n_1(n_1+1)/2$).

Under the null, W has a symmetric (discrete) distribution, with

$$\mu = \frac{n_1\big(n_1 + n_2 + 1\big)}{2}\,, \qquad \sigma^2 = \frac{n_1 n_2\big(n_1 + n_2 + 1\big)}{12}\,.$$

As in the signed rank test, exact significance levels are usually extracted from tables, but only for small samples. For $n_2 > 10$, asymptotic levels are accepted.

Example 1 (continuation). In the analysis of the `jobsat1` data set, to compare Chile and Mexico as in the preceding lecture, we get $W = 5,580.5$ ($z = -2.224$, $P = 0.026$), similar to the t test.

The Kruskal-Wallis test

The **Kruskal-Wallis test** is an extension of the rank-sum test to k samples, just as the one-way ANOVA F test is an extension of the two-sample t test. The assumptions are as in the rank-sum tests. The test statistic is obtained as follows:

- The samples are merged and the resulting sample is sorted, assigning ranks.
- The test statistic is
$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1),$$
where R_i is the sum of the ranks of sample i and n is the total sample size ($n = n_1 + \cdots + n_k$).

For $n_i \geq 5$, the distribution of H can be approximated by a $\chi^2(k-1)$.

Example 1 (continuation). To compare the three countries of the `jobsat1` data set, we get $\chi^2(2) = 9.46$ ($P = 0.009$), with more significance than in the one-way ANOVA test.

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.

B. Apply a nonparametric test to the `scapital` data set, and compare the result with that obtained in the preceding lecture.