13.3: The F Distribution and the F-Ratio

The distribution used for the hypothesis test is a new one. It is called the F distribution, named after Sir Ronald Fisher, an English statistician. The F statistic is a ratio (a fraction). There are two sets of degrees of freedom; one for the numerator and one for the denominator.

For example, if F follows an F distribution and the number of degrees of freedom for the numerator is four, and the number of degrees of freedom for the denominator is ten, then $F \sim F_{4,10}$.

> The F distribution is derived from the Student's t-distribution. The values of the F distribution are squares of the corresponding values of the t-distribution. One-Way $ANOVA$ expands the t-test for comparing more than two groups. The scope of that derivation is beyond the level of this course.

To calculate the F ratio, two estimates of the variance are made.

a. **Variance between samples:** An estimate of σ^2 that is the variance of the sample means multiplied by n (when the sample sizes are the same.). If the samples are different sizes, the variance between samples is weighted to account for the different sample sizes. The variance is also called **variation due to treatment or explained variation.**
b. **Variance within samples:** An estimate of σ^2 that is the average of the sample variances (also known as a pooled variance). When the sample sizes are different, the variance within samples is weighted. The variance is also called the **variation due to error or unexplained variation.**

- $SS_{\text{between}} =$ the sum of squares that represents the variation among the different samples
- $SS_{\text{within}} =$ the sum of squares that represents the variation within samples that is due to chance .

To find a "sum of squares" means to add together squared quantities that, in some cases, may be weighted. We used sum of squares to calculate the sample variance and the sample standard deviation in discussed previously.

MS means "mean square." MS_{between} is the variance between groups, and MS_{within} is the variance within groups.

Calculation of Sum of Squares and Mean Square

- $k =$ the number of different groups
- $n_j =$ the size of the j^{th} group}
- $s_j =$ the sum of the values in the j^{th} group
- $n =$ total number of all the values combined (total sample size):

$$n = \sum n_j \tag{13.3.1}$$

- $x =$ one value:

$$\sum x = \sum s_j \tag{13.3.2}$$

- Sum of squares of all values from every group combined:

$$\sum x^2 \tag{13.3.3}$$

- Between group variability:

$$SS_{\text{total}} = \sum x^2 - \frac{\left(\sum x^2\right)}{n} \tag{13.3.4}$$

- Total sum of squares:

$$\sum x^2 - \frac{\left(\sum x\right)^2}{n} \tag{13.3.5}$$

- Explained variation: sum of squares representing variation among the different samples:

$$SS_{\text{between}} = \sum \left[\frac{(s_j)^2}{n_j}\right] - \frac{\left(\sum s_j\right)^2}{n} \tag{13.3.6}$$

- Unexplained variation: sum of squares representing variation within samples due to chance:

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}} \tag{13.3.7}$$

- df's for different groups (df's for the numerator):

$$df = k - 1 \tag{13.3.8}$$

- Equation for errors within samples (df's for the denominator):

$$df_{\text{within}} = n - k \tag{13.3.9}$$

- Mean square (variance estimate) explained by the different groups:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} \tag{13.3.10}$$

- Mean square (variance estimate) that is due to chance (unexplained):

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} \tag{13.3.11}$$

MS_{between} and MS_{within} can be written as follows:

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}} = \frac{SS_{\text{between}}}{k-1} \tag{13.3.12}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}} = \frac{SS_{\text{within}}}{n-k} \tag{13.3.13}$$

The one-way $ANOVA$ test depends on the fact that MS_{between} can be influenced by population differences among means of the several groups. Since MS_{within} compares values of each group to its own group mean, the fact that group means might be different does not affect MS_{within}.

The null hypothesis says that all groups are samples from populations having the same normal distribution. The alternate hypothesis says that at least two of the sample groups come from populations with different normal distributions. If the null hypothesis is true, MS_{between} and MS_{within} should both estimate the same value.

The null hypothesis says that all the group population means are equal. The hypothesis of equal means implies that the populations have the same normal distribution, because it is assumed that the populations are normal and that they have equal variances.

F-Ratio or F Statistic

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{13.3.14}$$

If MS_{between} and MS_{within} estimate the same value (following the belief that H_0 is true), then the F-ratio should be approximately equal to one. Mostly, just sampling errors would contribute to variations away from one. As it turns out, MS_{between} consists of the population variance plus a variance produced from the differences between the samples. MS_{within} is an estimate of the population variance. Since variances are always positive, if the null hypothesis is false, MS_{between} will generally be larger than MS_{within}.Then the F-ratio will be larger than one. However, if the population effect is small, it is not unlikely that MS_{within} will be larger in a given sample.

The foregoing calculations were done with groups of different sizes. If the groups are the same size, the calculations simplify somewhat and the F-ratio can be written as:

F-Ratio Formula when the groups are the same size

$$F = \frac{n \cdot s_{\bar{x}}^2}{s_{\text{pooled}}^2} \tag{13.3.15}$$

where ...

- $n =$ the sample size
- $df_{\text{numerator}} = k-1$
- $df_{\text{denominator}} = n-k$
- $s_{\text{pooled}}^2 =$ the mean of the sample variances (pooled variance)
- $s_{\bar{x}^2} =$ the variance of the sample means

Data are typically put into a table for easy viewing. One-Way $ANOVA$ results are often displayed in this manner by computer software.

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	$SS(\text{Factor})$	$k-1$	$MS(\text{Factor}) = \frac{SS(\text{Factor})}{(k-1)}$	$F = \frac{MS(\text{Factor})}{MS(\text{Error})}$
Error (Within)	$SS(\text{Error})$	$n-k$	$MS(\text{Error}) = \frac{SS(\text{Error})}{(n-k)}$	
Total	$SS(\text{Total})$	$n-1$		

Example 13.3.1

Three different diet plans are to be tested for mean weight loss. The entries in the table are the weight losses for the different plans. The one-way ANOVA results are shown in Table.

Plan 1: $n_1 = 4$	Plan 2: $n_2 = 3$	Plan 3: $n_3 = 3$
5	3.5	8
4.5	7	4
4		3.5
3	4.5	

$$s_1 = 16.5, s_2 = 15, s_3 = 15.7 \tag{13.3.16}$$

Following are the calculations needed to fill in the one-way ANOVA table. The table is used to conduct a hypothesis test.

$$SS(\text{between}) = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{\left(\sum s_j\right)^2}{n} \tag{13.3.17}$$

$$= \frac{s_1^2}{4} + \frac{s_2^2}{3} + \frac{s_3^2}{3} + \frac{(s_1 + s_2 + s_3)^2}{10} \tag{13.3.18}$$

where $n_1 = 4, n_2 = 3, n_3 = 3$ and $n = n_1 + n_2 + n_3 = 10$ so

$$SS(\text{between}) = \frac{(16.5)^2}{4} + \frac{(15)^2}{3} + \frac{(5.5)^2}{3} = \frac{(16.5 + 15 + 15.5)^2}{10} \tag{13.3.19}$$

$$= 2.2458 \tag{13.3.20}$$

$$S(\text{total}) = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \tag{13.3.21}$$

$$= (5^2 + 4.5^2 + 4^2 + 3^2 + 3.5^2 + 7^2 + 4.5^2 + 8^2 + 4^2 + 3.5^2) \tag{13.3.22}$$

$$- \frac{(5 + 4.5 + 4 + 3 + 3.5 + 7 + 4.5 + 8 + 4 + 3.5)^2}{10} \tag{13.3.23}$$

$$= 244 - \frac{47^2}{10} = 244 - 220.9 \tag{13.3.24}$$

$$= 23.1 \tag{13.3.25}$$

$$SS(\text{within}) = SS(\text{total}) - SS(\text{between}) \tag{13.3.26}$$

$$= 23.1 - 2.2458 \tag{13.3.27}$$

$$= 20.8542 \tag{13.3.28}$$

One-Way $ANOVA$ Table: The formulas for $SS(\text{Total})$, $SS(\text{Factor}) = SS(\text{Between})$ and $SS(\text{Error}) = SS(\text{Within})$ as shown previously. The same information is provided by the TI calculator hypothesis test function $ANOVA$ in STAT TESTS (syntax is $ANOVA(L1, L2, L3)$ where $L1, L2, L3$ have the data from Plan 1, Plan 2, Plan 3 respectively).

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)
Factor (Between)	$SS(\text{Factor}) = SS(\text{Between}) = 2.2458$	$k -$
Error (Within)	$SS(\text{Error}) = SS(\text{Within}) = 20.8542$	$n -$
Total	$SS(\text{Total}) = 2.2458 + 20.8542 = 23.1$	$n -$

Mean Square (MS)	F
$MS(\text{Factor}) = \frac{SS(\text{Factor})}{(k-1)} = \frac{2.2458}{2} = 1.1229$	$F = \frac{MS(\text{Factor})}{MS(\text{Error})} = \frac{1.1229}{2.9792} = 0.3769$
$MS(\text{Error})) = \frac{SS(\text{Error})}{(n-k)} = \frac{20.8542}{7} = 2.9792$	

Exercise 13.3.1

As part of an experiment to see how different types of soil cover would affect slicing tomato production, Marist College students grew tomato plants under different soil cover conditions. Groups of three plants each had one of the following treatments

- bare soil
- a commercial ground cover
- black plastic
- straw
- compost

All plants grew under the same conditions and were the same variety. Students recorded the weight (in grams) of tomatoes produced by each of the $n = 15$ plants:

Bare: $n_1 = 3$	Ground Cover: $n_2 = 3$	Plastic: $n_3 = 3$	Straw: $n_4 = 3$	Compost: $n_5 = 3$
2,625	5,348	6,583	7,285	6,277
2,997	5,682	8,560	6,897	7,818
4,915	5,482	3,830	9,230	8,677

Create the one-way ANOVA table.

Answer

One-Way ANOVA table

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F
Factor (Between)	36,648,561	$5 - 1 = 4$	$\frac{36,648,561}{4} = 9,162,140$	$\frac{9,162,140}{2,044,672.6} = 4.4810$
Error (Within)	20,446,726	$15 - 5 = 10$	$\frac{20,446,726}{10} = 2,044,672.6$	
Total	57,095,287	$15 - 1 = 14$		

The one-way ANOVA hypothesis test is always right-tailed because larger F-values are way out in the right tail of the F-distribution curve and tend to make us reject H_0.

Notation

The notation for the F distribution is $F \sim Fdf(\text{num}), df(\text{denom})$

where $df(\text{num}) = df_{between}$ $and df(\text{denom}) = df_{within}$

The mean for the F distribution is $\mu = \frac{df(\text{num})}{df(\text{denom}) - 1}$

References

1. Tomato Data, Marist College School of Science (unpublished student research)

Chapter Review

Analysis of variance compares the means of a response variable for several groups. $ANOVA$ compares the variation within each group to the variation of the mean of each group. The ratio of these two is the F statistic from an F distribution with (number of groups – 1) as the numerator degrees of freedom and (number of observations – number of groups) as the denominator degrees of freedom. These statistics are summarized in the $ANOVA$ table.

Formula Review

$$SS_{between} = \sum \left[\frac{(s_j)^2}{n_j} \right] - \frac{(\sum s_j)^2}{n}$$

$$SS_{\text{total}} = \sum x^2 - \frac{(\sum x)^2}{n}$$

$$SS_{\text{within}} = SS_{\text{total}} - SS_{\text{between}}$$

$$df_{\text{between}} = df(\text{num}) = k - 1$$

$$df_{\text{within}} = df(\text{denom}) = n - k$$

$$MS_{\text{between}} = \frac{SS_{\text{between}}}{df_{\text{between}}}$$

$$MS_{\text{within}} = \frac{SS_{\text{within}}}{df_{\text{within}}}$$

$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$

F ratio when the groups are the same size: $F = \frac{ns_x^2}{s_{\text{pooled}}^2}$

Mean of the F distribution: $\mu = \frac{df(\text{num})}{df(\text{denom}) - 1}$

where:

- $k =$ the number of groups
- $n_j =$ the size of the j^{th} group
- $s_j =$ the sum of the values in the j^{th} group
- $n =$ the total number of all values (observations) combined
- $x =$ one value (one observation) from the data
- $s_x^2 =$ the variance of the sample means
- $s_{\text{pooled}}^2 =$ the mean of the sample variances (pooled variance)

Contributors and Attributions

Barbara Illowsky and Susan Dean (De Anza College) with many other contributing authors. Content produced by OpenStax College is licensed under a Creative Commons Attribution License 4.0 license. Download for free at http://cnx.org/contents/30189442-699...b91b9de@18.114.