

Lesson 15 ANOVA (analysis of variance)

Outline

Variability

- between group variability
- within group variability
- total variability
- F-ratio

Computation

- sums of squares (between/within/total)
- degrees of freedom (between/within/total)
- mean square (between/within)
- F (ratio of between to within)

Example Problem

Note: The formulas detailed here vary a great deal from the text. I suggest using the notation I have outlined here since it will coincide more with what we have already done, but you might look at the text version as well. Use whatever method you find easiest to understand.

Variability

Please read about this topic on the web page by locating the ANOVA demonstration or you can click here: <http://faculty.uncfsu.edu/dwallace/sanova.html>

Between group variability and within group variability are both components of the total variability in the combined distributions. What we are doing when we compute between and within variability is to partition the total variability into the between and within components. So: Between variability + within variability = total variability

Hypothesis Testing

Again, with ANOVA we are testing hypotheses that involve comparisons of two or more populations. The overall test, however, will indicate a difference between any of the groups. Thus, the test will not specify which two, or if some or all of the groups differ. Instead, we will conduct a separate test to determine which specific means differ. Because of this fact the research hypothesis will state simply that “at least two of the means differ.” The null will still state that there are no significant differences between any of the groups (insert as many “mu’s” as you have groups). $H_0: \mu_1 = \mu_2 = \mu_3$

Critical values are found using the F-table in your book. The table is discussed in the example below.

Computation

How do we measure variability in a distribution? That is, how do we measure how different scores are in the distribution from one another? You should know that we use variance as a measure of variability. With ANOVA or analysis of variance, we compute

a ratio of variances: between to within variance. Recall that variance is the average square deviation of scores about the mean. We will compute the same value here, but as the definition suggests, it will be called the “mean square” for the computations.

So, we are computing variance. Recall that when we compute variance we first find the sum of the square deviations, and then divide by the sample size (n -1 or degrees of freedom for a sample).

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}$$

\leftarrow Sums of Squares
 \leftarrow degrees of freedom

When we compute the Mean Square (variance) in order to form the F-ratio, we will do the exact same thing: compute the sums of squares and divide by degrees of freedom. Don't let the formulas intimidate you. Keep in mind that all we are doing is finding the variance for our between factor and dividing that by the variance for the within factor. These two variances will be computed by finding each sums of squares and dividing those sums of squares by their respective degrees of freedom.

Sums of Squares

We will use the same basic formula for sums of squares that we used with variance. While we will only use the between variance and within variance to compute the F-ratio, we will still compute the sums of squares total (all values) for completeness.

Total Sums of Squares

$$SS_{TOT} = \sum X_{TOT}^2 - \frac{(\sum X_{TOT})^2}{N_{TOT}}$$

\leftarrow Note that it is the same formula we have been using. The

subscript (tot) stands for the total. It indicates that you perform the operation for ALL values in your distribution (all subjects in all groups).

Within Sums of Squares

$$SS_{WITHIN} = \left(\sum X_1^2 - \frac{(\sum X_1)^2}{n_1} \right) + \left(\sum X_2^2 - \frac{(\sum X_2)^2}{n_2} \right) + \dots + \left(\sum X_k^2 - \frac{(\sum X_k)^2}{n_k} \right)$$

Notice that each segment is the same formula for sums of squares we used in the formula for variance and for the total sums of squares above. What is different here is that you consider each group separately. So, the first segment with the subscript “1” means you compute the sum of squares for the first group. Group two is labeled with a “2”, but notice that after that we have group “k” instead of a number. This notation indicates that you continue to find the sums of squares as you did for the first two groups for however

many groups you have in the problem. So, “k” could be the third group, or if you have four groups then you would do the same sums of squares computation for the third and fourth group.

Between Sums of Squares

$$SS_{BETWEEN} = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \dots + \frac{(\sum X_k)^2}{n_k} - \frac{(\sum X_{TOT})^2}{N_{TOT}}$$

We have the same “k” notation here. Again, you perform the same operation for each separate group in your problem. However, with this formula once we compute the value for each group we must subtract an operation at the final step. This operation is half the sums of squares we computed for the sums of squares total.

Degrees of Freedom

Again, we will first compute the sums of squares for each source of variance, divide the values by degrees of freedom in order to get the two mean square values we need to form the F-ratio. Degrees of freedom, however, is different for each source of variability.

Total Degrees of freedom

$N - 1$ ← this N value is the total number of values in all groups

Within Degrees of freedom

$df_{within} = N - K$ ← K is the number of categories or groups, N is still the total N

Between Degrees of freedom

$df_{Between} = K - 1$

We will also use degrees of freedom to locate the critical value on the F-table (see page A-29 for alpha .05 and A-30 for alpha .01). The numerator of the F-ratio is the between factor, so we will use the degrees of freedom between along the top of your table. The denominator of the F-ratio is the within subjects factor, so will use degrees of freedom within along the left margin of the table.

Mean Square

Now we divide each sums of squares by the respective mean square. Don’t let the formula’s intimidate you. All we are doing is matching up degrees of freedom with the Sums of squares to get the mean square (variance)

Within Mean Square

$$MS_{Between} = \frac{SS_{Between}}{df_{Between}}$$

Between Mean Square

$$MS_{WITHIN} = \frac{SS_{Within}}{df_{Within}}$$

F-ratio

The final step is to divide our between by within variance to see if the effect (between) is large compared to the error (within).

$$F = \frac{MS_{Between}}{MS_{Within}}$$

Example

A therapist wants to examine the effectiveness of 3 therapy techniques on phobias. Subjects are randomly assigned to one of three treatment groups. Below are the rated fear of spiders after therapy. Test for a difference at $\alpha = .05$

Therapy A	Therapy B	Therapy C
5	3	1
2	3	0
5	0	1
4	2	2
2	2	1
$\Sigma x_1 = 18$	$\Sigma x_2 = 10$	$\Sigma x_3 = 5$
$\Sigma x_1^2 = 74$	$\Sigma x_2^2 = 26$	$\Sigma x_3^2 = 7$

STEP 1: State the null and alternative hypotheses.

H_1 at least one mean differs

$H_0: \mu_1 = \mu_2 = \mu_3$

STEP 2: Set up the criteria for making a decision. That is, find the critical value.

You might do this step after Step 3 since that is where you compute the critical value.

$$df_{Between} = K - 1 = 3 - 1 = 2$$

$$df_{within} = N - K = 15 - 3 = 12$$

$$F_{critical} = 3.88$$

STEP 3: Compute the appropriate test-statistic.

Although in this example I have given the summary values, for some problems you might have to compute the sum of x, and sum of squared x's yourself.

$$SS_{TOT} = \sum X_{TOT}^2 - \frac{(\sum X_{TOT})^2}{N_{TOT}} = SS_{TOT} = 107 - \frac{(33)^2}{15} = 107 - \frac{1089}{15} = 107 - 72.6 = 34.4$$

$$SS_{WITHIN} = \left(\sum X_1^2 - \frac{(\sum X_1)^2}{n_1} \right) + \left(\sum X_2^2 - \frac{(\sum X_2)^2}{n_2} \right) + \dots + \left(\sum X_k^2 - \frac{(\sum X_k)^2}{n_k} \right)$$

$$SS_{WITHIN} = \left(74 - \frac{(18)^2}{5} \right) + \left(26 - \frac{(10)^2}{5} \right) + \left(7 - \frac{(5)^2}{5} \right)$$

$$SS_{WITHIN} = \left(74 - \frac{324}{5} \right) + \left(26 - \frac{100}{5} \right) + \left(7 - \frac{25}{5} \right)$$

$$SS_{WITHIN} = (74 - 64.8) + (26 - 20) + (7 - 5)$$

$$SS_{WITHIN} = 9.2 + 6 + 2 = 17.2$$

$$SS_{BETWEEN} = \frac{(\sum X_1)^2}{n_1} + \frac{(\sum X_2)^2}{n_2} + \dots + \frac{(\sum X_k)^2}{n_k} - \frac{(\sum X_{TOT})^2}{N_{TOT}}$$

$$SS_{BETWEEN} = \frac{(18)^2}{5} + \frac{(10)^2}{5} + \frac{(5)^2}{5} - \frac{(33)^2}{15}$$

$$SS_{BETWEEN} = \frac{324}{5} + \frac{100}{5} + \frac{25}{5} - \frac{1089}{15}$$

$$SS_{BETWEEN} = 64.8 + 20 + 5 - 72.6 = 17.2$$

Note that anytime you compute two of the Sums of Squares you can derive the third one without computation because Between + Within = Total

$$df_{tot} = N - 1 \quad df_{Between} = K - 1 \quad df_{within} = N - K$$

$$df_{tot} = 15 - 1 = 14 \quad df_{Between} = 3 - 1 = 2 \quad df_{within} = 15 - 3 = 12$$

$$MS_{Between} = \frac{SS_{Between}}{df_{Between}} \quad MS_{WITHIN} = \frac{SS_{Within}}{df_{Within}} \quad F = \frac{MS_{Between}}{MS_{Within}}$$

$$MS_{Between} = \frac{17.2}{2} = 8.6 \quad MS_{WITHIN} = \frac{17.2}{12} = 1.43 \quad F = \frac{8.6}{1.43} = 6$$

Once we have computed all the values, very often we place them in a source table (below). Putting the values in a table like this one may make it easier to think about the statistic. Notice that once we get the Sums of Squares on the table, we will divide those values by the df in the next column. Once we get the two mean squares we divide those to get F.

Source	SS	df	MS	F
Between (group)	<u>17.2</u>	<u>2</u>	<u>8.6</u>	<u>6</u>
Within (error)	<u>17.2</u>	<u>12</u>	<u>1.43</u>	
Total	<u>34.4</u>	<u>14</u>		

STEP 4: Evaluate the null hypothesis (based on your answers to the above steps).

Reject the null

STEP 5: Based on your evaluation of the null hypothesis, what is your conclusion?

There is at least one group that is different from at least one other group.