5.02 Analysis of Variance: One-way ANOVA assumptions and F-test

In this video we'll discuss the overall F-test in one-way ANOVA. We'll look at the assumptions, the statistical null and alternative hypothesis; we'll see how to calculate the test statistic, determine the p-value and how to interpret the results.
Suppose we want to compare healthiness of three groups of cats that consume different diets: Raw meat, canned food and dry food. A veterinarian rates their health on a scale from zero to ten.

Assumptions

First we need to know and check the assumptions of ANOVA. These are equivalent to the assumptions of multiple linear regression, although they might look a little different. First of all, the observations should be independent of each other. Random selection - or random assignment in experiments - should take care of this. If the observations are dependent by design, for instance with repeated measurements of the same participants, or paired participants, you should use repeated measured ANOVA, which we'll not go into here. In our example we will assume cats were randomly assigned to the conditions.

Secondly, the response variable should be normally distributed in each group. The histograms in our example look acceptable. Even if they didn't, moderate violation of normality isn't problematic as long as the samples are large enough, say at least ten observations in each group. If the samples are small and very skewed you should use a non-parametric test instead.

Thirdly, ANOVA assumes homogeneity of variances, which means the population variance is assumed to be the same for all groups. Remember, this assumption forms the basis for the trick of using a population variance estimate that is always accurate and an estimate that is sometimes not so accurate to detect a difference in the means.
Moderate violation of homogeneity of variances is not problematic if the group sizes are equal. If the group sizes are unequal a rule of thumb is that you can still perform ANOVA as long as the largest standard deviation is no more than twice the size of the smallest standard deviation.

Statistical hypotheses

The null hypothesis states that all population means are equal - H_0: $\mu_1 = \mu_2 = \cdots = \mu_g$. The alternative hypothesis states that *at least* one population mean differs from the rest. This is a non-directional hypothesis just like with the overall test in multiple regression; it specifies that there is *a* difference *somewhere*. It doesn't specify for which groups we expect a difference and in what direction.

Test statistic

The test statistic F equals the between-group variance divided by the within-group variance: $F = \frac{MS_{between}}{MS_{within}}$. If the group sizes are all equal you can calculate the F-value manually very easily using the sample variances:

$$F = \frac{MS_b}{MS_w} = \frac{\frac{n_c \Sigma(\bar{y}_j - \bar{y})^2}{g-1}}{\frac{\Sigma s_j^2}{g}}.$$

If the group sizes are unequal you have to calculate the sum of squares in each group and divide by the appropriate degrees of freedom to obtain the variances:

$$F = \frac{MS_b}{MS_w} = \frac{\frac{SS_b}{g-1}}{\frac{SS_W}{n-g}} = \frac{\frac{\Sigma n_j(\bar{y}_j - \bar{y})^2}{g-1}}{\frac{\Sigma\Sigma(y_{ij} - \bar{y}_j)^2}{n-g}}.$$

The within-group sum of squares will probably seem most familiar. For each group you take the difference between each observation and the group mean, you square this difference and you add the squared differences. That's what the inner summation sign symbolized. The outer summation sign indicates that you have to do this for each group and then add the results of all groups.
You've now calculated the within-group sum of squares, which turns into the within-group variance once you divide by the total number of participants in all groups - n - minus the number of groups - g: $F = MS_w = \frac{SS_W}{n-g} = \frac{\Sigma\Sigma(y_{ij} - \bar{y}_j)^2}{n-g}$.

Manually calculating the between-group variance follows the same logic; you just treat the means as if they were individual observations. For each group you take the difference between the group mean and the grand mean, square the difference and multiply by the number of participants in that group. If you do this for each group and add the results you have the between-group sum of squares. This turn into the between-group variance once you divide by the number of groups - g - minus one:

$$MS_b = \frac{SS_b}{g-1} = \frac{\Sigma n_j(\bar{y}_j - \bar{y})^2}{g-1}.$$

Statistical software reports not just the F-value but also the sums of squares and mean sums of squares. Remember, mean sum of squares is just another word for variance! One thing to look out for is the difference in presentation between statistical software packages. The between-group mean sum of

squares is often referred to by the name of the factor that represents the groups. The within-group mean sum of squares is often referred to as the mean square error, abbreviated by MSE.
We've already calculated the two degrees of freedom associated with the F distribution, they are the *numerator* or *between* degrees of freedom - the number of groups minus one: $df_{between} = g - 1$; and the *denominator*, *within*, or *error* degrees of freedom - the total number of observations minus the number of groups: $df_{within} = n - g$.

Test statistic distribution and p-value

The test statistic follows an F distribution, with two separate degrees of freedom. [1]In our example we find an F of 3.793 with 2 and 46 degrees of freedom. Since the F-test is non-directional we always look in the right tail of the distribution. With the significance level set at 0.05, using a table we find the critical F-value with 2 and 40 degrees of freedom is 3.2317. The observed F-value exceeds this value, so we know the p-value is smaller than 0.05. Software provides an exact p-value of 0.03. We can reject the null hypothesis and conclude that at least one of the diet groups differs from the others in terms of mean health rating.

[1]
```
            Df Sum Sq Mean Sq F value Pr(>F)
diet         2   32.1  16.052   3.793 0.0299 *
Residuals   46  194.7   4.232
```