5.03 Analysis of Variance: One-way ANOVA post-hoc comparisons

In this video we'll see how to follow up a significant F-test in one-way ANOVA with **pairwise comparisons** of group means using pairwise t-tests or confidence intervals. Pairwise comparisons help us to determine why the overall effect occurred; they tell us which group means differ significantly and in what direction.

Follow-up comparisons are often referred to as **post-hoc** comparisons. The term post-hoc indicates that we're making comparisons 'after the fact', so without having a clear hypothesis about which groups will differ in which direction before collecting the data and performing the analysis. This implies using two-sided tests or confidence intervals.
If we do have a clear expectation about how the individual group means will differ we can perform **planned comparisons**. These are outside the scope of this introduction however.

Suppose we performed an F-test to compare healthiness of three groups of cats that consumed different diets: Raw meat, canned food and dry food. Health was rated on a scale from zero to ten. [1]Suppose we found an F-value of 3.793, with 2 and 46 degrees of freedom and with a p-value of 0.03, indicating a significant difference in the groups. To find out how we should interpret this significant overall effect we'll determine post-hoc confidence intervals.
If we have g groups there are g * (g - 1) / 2 comparisons to be made. In our example we have three groups, so three comparisons (3 * 2 / 2 = 3).
Remember, these comparisons should only be performed if the overall test is significant.

Assumptions

When we perform the comparisons, whether using pairwise t-tests or confidence intervals, the same assumptions should hold as for the F-test: **Independence, normality** and **homogeneity of variances**. Of course you'll have already checked these before performing the overall F-test.

Test statistic

The formulas for the t-test and confidence interval:

$$t = \frac{\bar{y}_j - \bar{y}_k}{\sqrt{\frac{SS_{within}}{n-g}} \cdot \sqrt{\frac{1}{n_j} + \frac{1}{n_k}}} \quad \text{and} \quad CI = \bar{y}_j - \bar{y}_k \pm t_{\alpha/2} \cdot \sqrt{\frac{SS_{within}}{n-g}} \cdot \sqrt{\frac{1}{n_j} + \frac{1}{n_k}}$$

[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
```

are almost the same as for the 'regular' t-test and confidence interval for two independent groups, assuming equal population variances:

$t = \frac{\bar{y}_j - \bar{y}_k}{s\sqrt{\frac{1}{n_j}+\frac{1}{n_k}}}$ and $CI = \bar{y}_j - \bar{y}_k \pm t_{\alpha/2} \cdot s\sqrt{\frac{1}{n_j}+\frac{1}{n_k}}$.

In post-hoc comparisons we use **Fisher's Least Significant Difference** (**LSD**) method, which refers to the use of the **residual standard deviation** - the square root of the within-group variance - instead of the pooled standard deviation to calculate the standard error. So in each pairwise comparison we estimate the standard error based on the variance in *all* groups, including the ones not in the comparison.
Here's the formula for the confidence interval:

$$CI = \bar{y}_j - \bar{y}_k \pm t_{\alpha/2} \cdot \sqrt{\frac{SS_{within}}{n-g}} \cdot \sqrt{\frac{1}{n_j}+\frac{1}{n_k}},$$

it's the difference between the means of group j and k minus and plus the appropriate t-value times the standard error. The t-value is the value associated with half of the significance level and the error degrees of freedom, which equal the total number of observations in all groups minus the total number of groups (df=n-g).
The standard error equals the residual standard deviation - the square root of the within sum of squares divided by the error degrees of freedom - times the square root of one over the size of group j plus one over the size of group k. If the assumption of homogeneity is violated you should use the formula that makes no assumption about the population variances[2].

Since we'll be making multiple comparisons, we should correct for the inflated **Family-Wise Error Rate** (**FWER**) - the probability that *at least* one of the comparisons will result in a false rejection of the null hypothesis. There are many correction methods, often referred to as **multiple comparison** methods. We'll consider two of these.

The **Bonferroni** method involves dividing the desired overall alpha by the number of comparisons and using the resulting corrected alpha for the individual comparisons. With this correction the actual probability of falsely rejecting the null will be smaller than or equal to the desired overall alpha. In many cases the correction is overly conservative, resulting in a smaller alpha and less power.

---

[2] $se = \sqrt{\frac{s_j^2}{n_j}+\frac{s_k^2}{n_k}}$ $\quad df = \frac{\left(\frac{s_j^2}{n_j}+\frac{s_k^2}{n_k}\right)^2}{\frac{\left(\frac{s_j^2}{n_j}\right)^2}{(n_j-1)}+\frac{\left(\frac{s_k^2}{n_k}\right)^2}{(n_k-1)}}$

Tukey's Honestly Significant Difference method is less conservative. The actual probability of falsely rejecting the null is closer to the desired overall alpha. It results in more powerful tests and narrower confidence intervals than the Bonferroni method. Tukey's method uses a test statistic distribution slightly different from the t distribution, so we'll leave the calculation of the test statistic, confidence interval and p-values to software.

In this example we'll use the Bonferroni method and divide the standard alpha level of 0.05 by 3, resulting in a corrected alpha of 0.017. If we use tables to determine significance we'll have to settle for a corrected alpha of 0.010, since 0.017 isn't in the table.

The critical t-value is the value listed at 40 degrees of freedom (rounding down from 46) and half the significance level, so 0.005. We find a critical t-value of 2.7045. The residual standard deviation equals 2.0573. Using the formulas for each of the three comparisons with the appropriate group means, we find confidence intervals ranging from -0.07 and 3.93 for the difference between raw meat and canned food, -0.41 and 3.48 for the difference between raw meat and dry food, and -1.52 and 2.31 for the difference between canned and dry food.

None of the intervals show a significant difference. This is not only because the Bonferroni method has less power, but also because we rounded down our alpha and degrees of freedom by using tables. If we use Tukey's method, with more power, we find that only the interval for the difference between raw meat and canned food does *not* contain zero, so we reject the null hypothesis for this comparison only.

Looking at the mean health scores we can conclude that raw meat results in a significantly higher average health score than canned food. The mean for dry food lies in between these means and does *not* differ significantly from either raw meat or dry food.