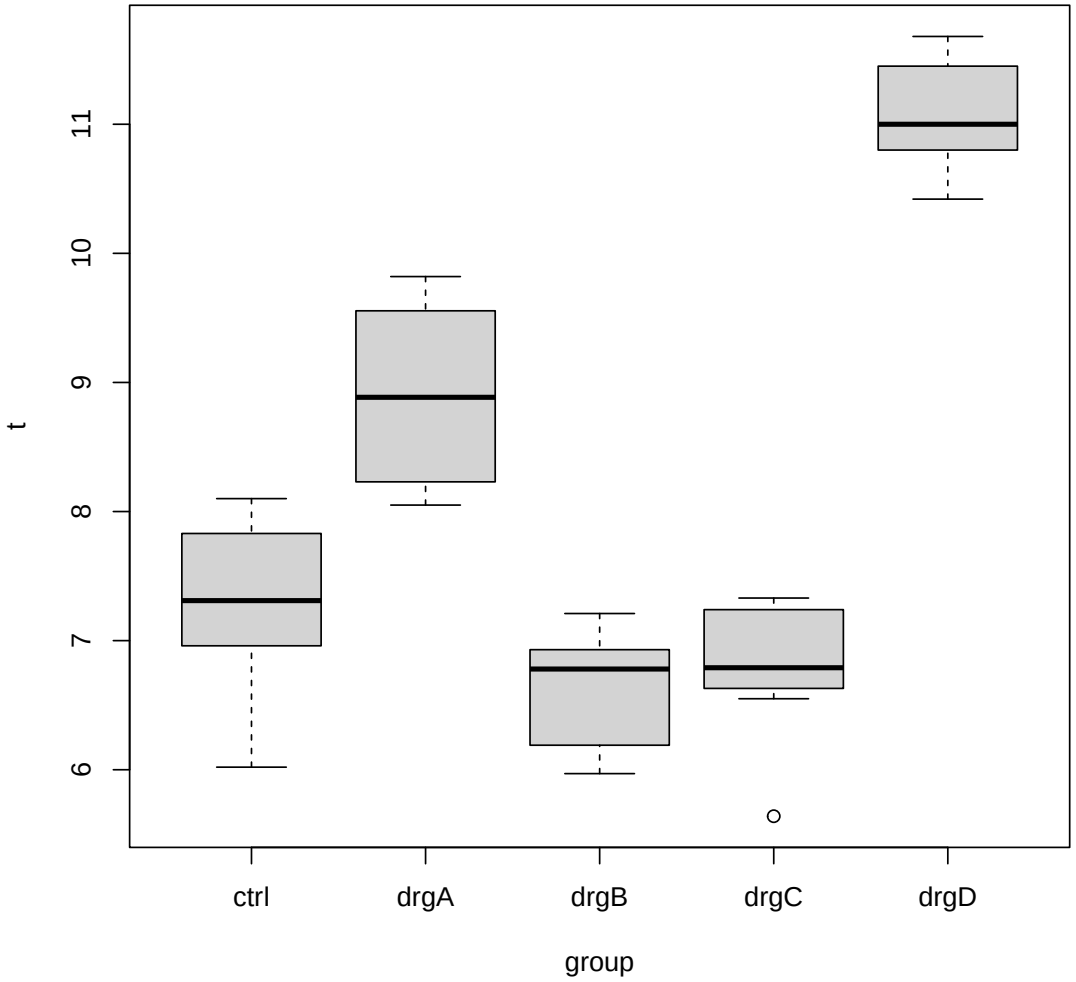
Quantitative Understanding in Biology Short Course
Session 8
Working with Categorical Data; ANOVA

Jason Banfelder
Luce Skrabanek

May 6th, 2025

1 The One Way ANOVA Omnibus Test as a Model Comparison

In all of the model-fitting examples that we've looked at so far, we have only considered cases where the explanatory variables were continuous. But in many cases, we work with systems where some (or all) of the explanatory variables are categorical. For example, we may want to investigate if any of a number of drugs have an effect on cognitive ability, as measured by how long it takes to solve a puzzle. We'll use the dataset provided in the supplementary file `puzzletime.txt` to explore these concepts. These data are found in the Appendix, and plotted below.



Although you might analyze these data with a series of t-tests effecting various comparisons among the groups (with the appropriate multiple-hypothesis testing corrections, of course!), a more appropriate approach is to begin with an onmibus test, in this case a one way ANOVA. In general terms, when performing an omnibus test, we analyze all the data at once, producing a single p-value that tells us if there is anything particularly interesting or surprising in the data.

In the case of a one-way ANOVA, the null hypothesis is that the means of all of the groups

(control, and each of the treated groups) are the same, and the alternate hypothesis is that the groups have distinct means. There are two equivalent ways to think about this.

The first approach (which is how ANOVA is typically introduced), partitions the variance in the data into variance between groups and variance within groups, with the idea being that if all the observed data were really samples from the same distribution, then the between and within group variances wouldn't be all that different. This approach gives ANOVA its name: Analysis of Variance.

A mathematically equivalent line of reasoning, and the one we'll prefer here, is to view the problem in the context of what we've recently learned about model fitting. We consider two models that might explain the data, and use an F-test to decide which model is preferred. The only thing that's really new in this approach is that the model uses a categorical variable (which group an observation belongs to), rather than a continuous variable, as the predictor.

The simple model is the one where the groups don't matter. In that case, the mathematical model to explain the data is:

$$y = \beta + \epsilon \tag{1}$$

In this model, β is our best guess for any data value, and we acknowledge that there is some random variance (due to biological diversity and/or measurement noise) around that value. Unsurprisingly, the value of β that minimizes the sum-of-squares of the residuals is the average of the observations: $\beta = \overline{x}$.

The alternative model is one where each group has its own, distinct average. For our dataset, such a model could be written as:

$$y = \beta + \beta_A x_A + \beta_B x_B + \beta_C x_C + \beta_D x_D + \epsilon \tag{2}$$

In this case, the values of the parameters that minimize the SSQ of the residuals work out such that β is the mean of the control group, and β_i is the difference in the means of the group treated with drug i and the control group. In other words, β_i is the estimate of the effect of drug i relative to the control. The x_is are "dummy" variables that are zero if the measurement isn't in the group, and one if they are.

If we adopt this line of reasoning, we have two models that are candidates for explaining the data, and one is a special case of the other (equation 2 reduces to equation 1 when all of the $\beta_i = 0$). This is begging for an F-test, and that, in fact, is all a one way ANOVA is.

In our example dataset, there are 42 data points. For the simple model, there is just one parameter, so DF=41. For the more complex model, there are 5 parameters, leaving 37 degrees of freedom. The F-test will answer the question of whether the introduction of

the additional four parameters improves the SSQ enough to justify the additional model complexity.

You can import our sample data into Prism (use a datasheet based on Column data), and choose "One-way ANOVA" from the "Column analyses" section. You should see an F-value of 83.1, and a strong p-value. This indicates that the more complex model with a distinct mean for each group better explains the data.

2 ANOVA Post-tests

You should be aware that a one-way ANOVA with just two groups is exactly equivalent to a t-test. What is a bit odd about an ANOVA with three or more groups is that the result of the omnibus test doesn't say anything about which groups differ from each other. To explore that, we perform so-called post-tests (or "followup tests" in Prism), but only if the omnibus test yields a statistically significant result.

For a one-way ANOVA, there are two commonly used post-test procedures. The first, Dunnett's method, compares each group to a single control group. In Prism, this can be accessed via the Multiple Comparisons tab and selecting "Compare the mean of each column with the mean of a control column". The second, Tukey's method (also called Tukey's Honest Significant Difference, or HSD, method), compares every group with every other group. Tukey's method will compare drug A with drug C, while Dunnett's will not.

While we'll use both methods here for pedagogical purposes, it is important to keep in mind that this is almost never a good idea in practice. You should choose the method you're going to use based on the scientific question you are asking, and stick to that one. Additionally, you should have decided which procedure you were going to use before you collected your data. If you looked at the plot above and decided you didn't want to compare drug B with drug C, recognize that you did a (shortcut) post-test in your head. The table below may help:

t-test	Does methylphenidate affect cognitive ability?
Omnibus test	Do dopamine reuptake inhibitors affect cognitive ability?
Dunnett's procedure	Which dopamine reuptake inhibitors affect cognitive ability?
Tukey's procedure	Which dopamine reuptake inhibitor(s) affect cognitive ability the most?

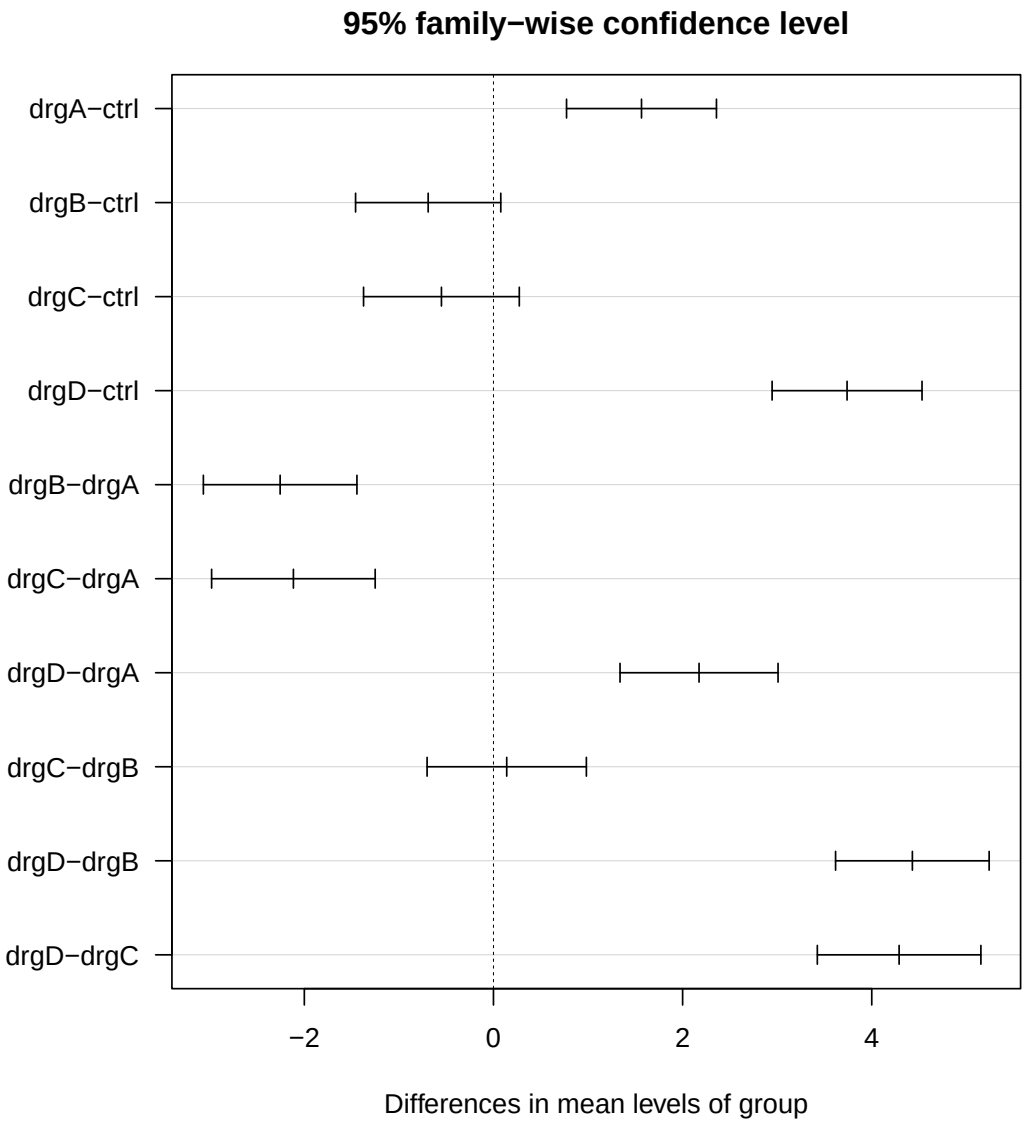
Without going into the gory details, one of the advantages of ANOVA with post-tests vs. a serialized t-test approach is that when using ANOVA, you use all of the collected variance information, even when comparing two groups.

Another weird consequence of ANOVA is that it is sometimes possible to obtain a significant p-value for the omnibus test, but not observe any significant results for the post-tests! Part of the reasoning is that the post-tests (even the ostensibly exhaustive Tukey procedure)

don't explore all possible comparisons. For example, if the reality is that drugs B and C don't have an effect on cognitive ability, while drugs A and D have the same effect, then the "correct" model is the two parameter model:

$$y = \beta_{ctrl,B,C} + \beta_{A,D} \cdot x_{A,D} + \epsilon \tag{3}$$

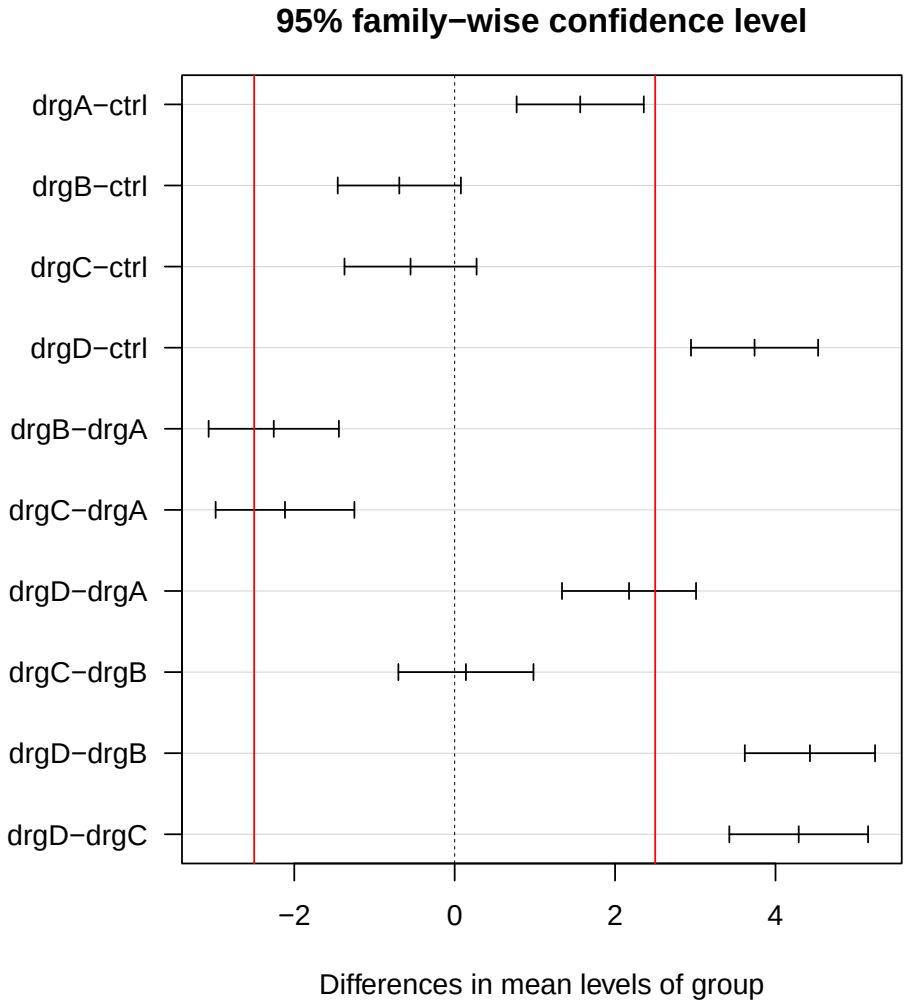
One of the most useful outputs of post-tests is a plot of confidence intervals of the difference between the means for the selected contrasts. Be sure to ask for this on the Options tab in Prism. Below is a plot you'd get for Tukey's method.



As you've grown to appreciate by now, CIs that include zero correspond to non-significant p-values. If you choose to go with Dunnett's procedure, you'd find that drug B would have been found to have a significant effect relative to control. This shouldn't surprise you, as both Tukey's and Dunnett's methods incorporate multiple hypothesis testing corrections, and since Dunnett's procedure involves testing fewer hypotheses, the correction is less aggressive. Since this particular contrast is borderline, this makes a difference.

It is important to remember that, so far, these plots and analyses only address *statistical*

significance. If, for example, you believe that differences of less than 2.5 for the time to complete the cognitive task are not biologically significant, then you might adorn the plot as follows:



Your interpretation might then be that although drug A has a statistically significant effect relative to the control condition, it is not plausible that the difference is biologically significant. Drug D, however, clearly has biologically significant activity relative to control. Similarly, while we can see that while drug D has statistically significantly different activity than drug A, we can't tell if the difference is large enough to be considered biologically significant.

3 Some cautions around ANOVA

There are three assumptions that go into an ANOVA analysis:

1. The data within each group is sampled from a normal distribution
2. The SDs of the groups are the same
3. The data are independent

Like many statistical tests we've covered, ANOVA is pretty robust to the first assumption. As we've done before, unless your data are pathologically non-normal, you're probably OK on the first point.

Data independence is more a function of study design and sampling methods than of statistical analysis. Other than to re-iterate that it is important to ensure that data collected are indeed independent, we won't discuss this further here.

The second assumption, that of equal SDs, is important because ANOVA is not particularly robust to this assumption. As a rule of thumb, if the SD of one group is three or more times that of another, you very well might have something to worry about. The effects are magnified when the groups have substantially different sample sizes, and especially so when the ns are small (say less than 5 or 6 per group). The effect is manifested as a substantial departure of the true Type I error rate from the α cutoff used. Depending on the particulars, this could be an increase or a decrease in the deviation. If you do need to worry about different SDs (or non-normality), Prism offers riffs on the classical one-way ANOVA such as the Brown-Forsythe procedure, and the non-parametric Kruskal-Wallis test based on ranks.

4 Homework

Using the data provided here, assess the effect of each drug relative to control using a series of t-tests, both with and without appropriate hypothesis testing correction. Compare your conclusions using these two methods to those you would make with the results from a one-way ANOVA with an appropriate post-test.

5 Appendix

5.1 Puzzle Time data.

	ctrl	drgA	drgB	drgC	drgD
1	7.16	9.62	6.19	5.64	10.42
2	6.85	8.05	6.57	6.55	10.85
3	8.10	8.27	7.21	6.79	10.75
4	7.40	9.33	5.97	7.24	10.86
5	7.22	9.49	6.93	7.33	11.61
6	7.83	8.44	6.09	7.24	11.68
7	7.76	9.82	6.91	6.71	11.14
8	6.02	8.19	6.78		11.29
9	8.06		7.16		
10	6.96				

