BJP British Journal of Pharmacology

DOI:10.1111/j.1476-5381.2012.01893.x
www.brjpharmacol.org

EDITORIAL

# Analysis of variance: variably complex

Gordon B Drummond[1] and Sarah L Vowler[2]

[1]*Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, UK, and* [2]*Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK*

**Correspondence**
Dr Gordon B Drummond, Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, 51 Little France Crescent, Edinburgh EH16 4HA, UK. E-mail: g.b.drummond@ed.ac.uk

This article is being published in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*.

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Sarah Vowler is Senior Statistician in the Bioinformatics Core at Cancer Research UK's Cambridge Research Institute.

This article is the 7th in a series of articles on best practice in statistical reporting. All the articles can be found at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1476-5381/homepage/statistical_reporting.htm

## Key points

- Although relatively robust to the following, ANOVA assumes
  - normal distribution of residuals
  - variation in each group is similar
- ANOVA is sensitive to outliers, and transforms of the data may be necessary.
- ANOVA is a test of group differences: do at least two of the means differ from each other?
- Two-way ANOVA tests a number of sources of variability:
  - It is important to know which source of variation is of interest.
  - Do the sources of variation interact?
- Care is needed with repeated measures.
- There may be better tests, particularly if the factor can be expressed as a continuous variable.

To compare two groups, we described how the *t*-test is used (Drummond and Tom, 2011a; 2011b). To compare more than two groups, we would use a different test, ANOVA. We start with a premise very similar to the logic of the *t*-test: is it possible that these groups could have been sampled from a single population? A variety of forms of ANOVA exist, and the test can be used (and misused!) in different ways. Some of these variants are very useful in the analysis of common experimental designs, when more that one intervention is used. To appreciate the different types of test of this sort, we will go back to the jumping frogs that we discussed before. (Drummond and Tom, 2011a) Let's suppose that we have a random sample of 30 frogs from California and also have 30 frogs sampled at random from Texas, and 30 from Ohio. We want to know if the means of the jump distance differ, according to the origin of the frogs.

The origin of the frogs is a categorical variable (frogs in our samples are from one of three states), and represent the *factor* that we believe might affect the jump distance. Each particular state is a different 'level' of this 'factor'. In different descriptions of this test, from book to book, these terms can vary. Other expressions may be used for these concepts of different 'levels' of a factor. Often, the factor being analysed might be an intervention imposed on the samples: then the 'factor' becomes a 'treatment', and the 'levels' would be different categories or types of treatment such as placebo, drug A or drug B. This categorization is important: ANOVA works with categories. If, on the other hand, our treatments were different doses (in other words, if the treatments were graded, or ordered, rather than just different), or if the treatment was a continuously distributed variable (such as fluid intake, for example), then there are probably better methods, such as fitting to a dose–response curve, for data analysis, or other forms of ANOVA. It is not particularly efficient to shoehorn continuous measurements into categories.

The terms we are using are

| | |
|---|---|
| Factor under study | Factor, treatment that could affect outcome |
| Categories within that factor | Level, sample: a class of the factor |

The *result* we are considering is the length of the jumps, which is a continuous variable. (Incidentally, the distance measured in a Calaveras competition is the distance covered by each frog in three jumps, not a single jump. A careful reader worked this out from first principles and wrote to tell us.) In the samples from each origin, the jump lengths vary (Figure 1). This is just what we might expect, and have seen in previous samples of frogs. This random variation can also be called residual variation, or unexplained variation. This last expression is used because when we do the test, some of the variation in overall jump lengths is attributed to the factor we are considering, that is the origin of the frogs. The variation in jump length is estimated and thus *explained* by the factor, origin. The rest of the variation is not attributed to origin, and thus this residual variation is termed '*unexplained*'. Because variation is the critical feature of this analysis, ANOVA is sensitive to outlying values. Inspection of a dot plot may indicate a potential problem, and transforming the data to reduce the effects of outliers is often helpful.

A valuable feature of ANOVA is that as further factors are considered (in this case, possibly random factors, not set by the investigator), these can be incorporated into the analysis. In this way, more of the variation is attributed to factors that are known, there is less residual variation, and the power of the test would increase.

| | | | |
|---|---|---|---|
| Variation can be | Within group | or | Between groups |
| Expressed as | Random | or | Due to factor |
| or | Unexplained | or | Explained by factor |
| or | Residual | | |
| or | Variation due to error | | |

Residual or within-group variation is assumed to follow the normal, Gaussian, distribution. In ANOVA, variation is calculated from the sum of squares (SS), and this value is given in ANOVA test results. Briefly, it is computed from the squares of the difference between each individual value and the mean value of the group that this individual value has come from.

In Figure 1A, we see this variation in a group of frogs, from the scatter of the individual values. The SS around the mean of this group's values will increase if the number of frogs in the group is increased. The more frogs, the more are the values that contribute to the mean value and to the sum. Here, the d.f. of the variation are ($n - 1$). This is because the final value is not free to vary, since if the mean value is given, and 29 of the individual values are known, then the final value is fixed. The SS *within* all three groups is computed, and these values are added. To derive an index of variation, and allow comparisons between groups of different sizes, we have to take into account the size of the groups. To do this, the SS is divided by the d.f., to obtain the mean squares (MS) within the groups. This measure of residual variation is computed within each group, using the mean of that group alone, before the values are added (Figure 1B). In this way, the measure of variation is not affected by any differences *between* the mean values of the three groups: it represents the variation *within* groups only.

We calculate the variation *between* the groups in a similar way, by using the mean value of each group and the mean value of all the samples, from all three groups (Figure 1C). ANOVA then tests if the variation *between* the group means is more than we might expect on the basis of the variation *within* the groups. If frogs originating from California, Texas and Ohio were all samples from a single uniform population, these random samples should have jump distances with a similar mean value. However, the mean values for frogs from the three states will not be exactly the same. We compare the variation between the groups and the variation within the groups by calculating the ratio:

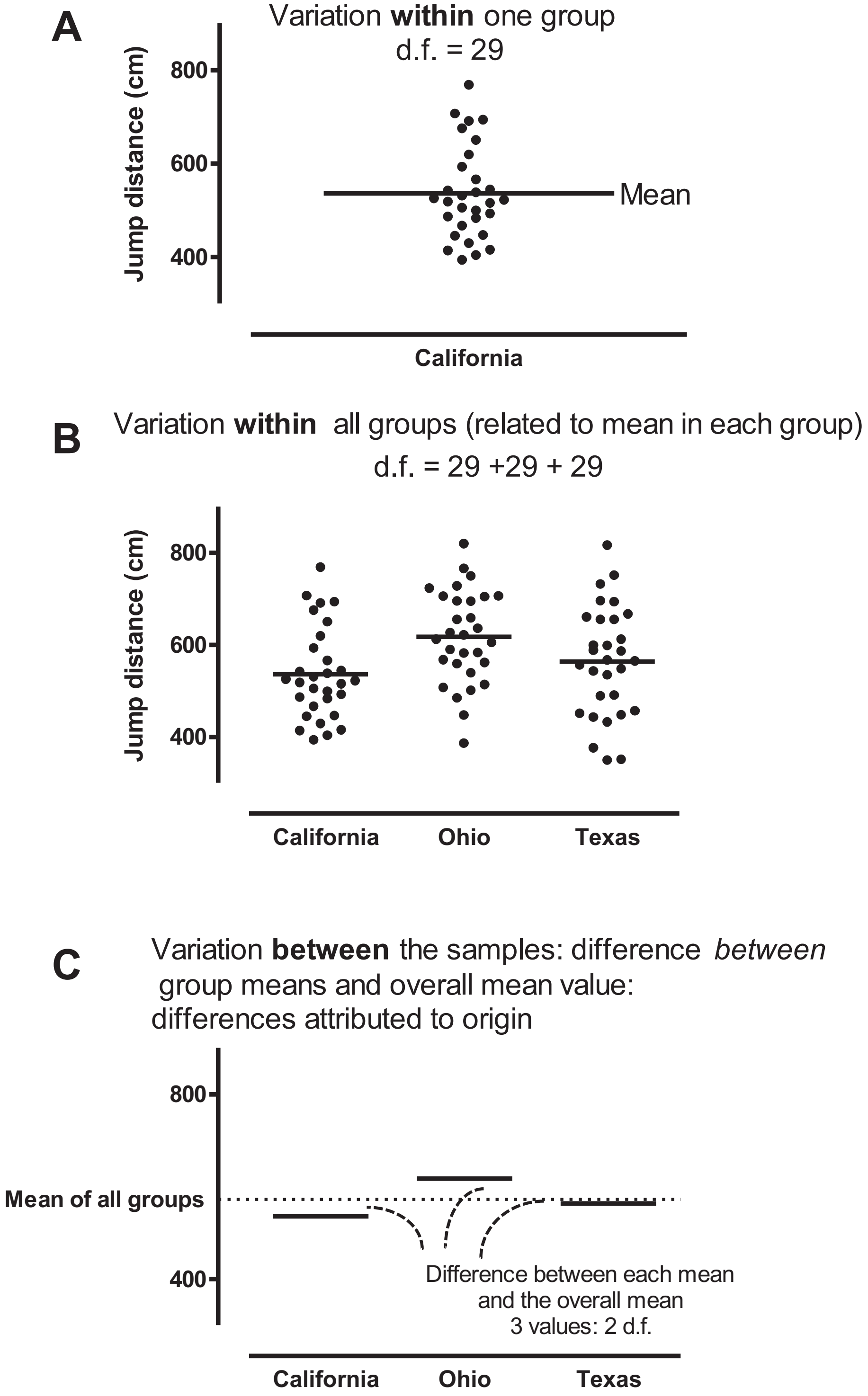
$$F = \text{MS (between groups)/MS (within groups)}$$

If the samples had been taken from a single population, then this ratio would be close to unity, more or less. If the value of $F$ is large, then it is more likely that the factor being studied has been affecting the mean values. In our example, the ratio is sufficiently large to suggest that the variation between the groups is much greater than the variation within the groups. That is unlikely if the groups had all been samples from the same population, and it is possible to estimate how unlikely this might be.

We have to make some assumptions about the population before we do this test: are these important? We need to assume that the variances are sufficiently similar, and that the population is normally distributed within the groups. Broadly speaking, as long as the samples are not too few, the sample sizes are not too different, and the group SD of the samples do not differ by more than twofold, these assumptions are not vital.

The next question: if we do find a difference, where does this difference lie? In this case, there was no reason to expect a specific result. It might be different if we had done an experiment, when there could be results expected in groups that received treatments, and not in a control group, and thus specific comparisons were of particular interest. As we have no reason to expect a specific result, we use Tukey's multiple comparison test, which compares all possible pairs of groups. This suggests that the Ohio frogs jump further than California frogs, but there is no evidence that Texas frogs differ from either other group. If we had one group that could be considered 'normal' or 'baseline' and wished to compare the others to this group, then Dunnett's multiple comparison test could be used. We shall consider multiple comparisons more carefully in a further article.

The consideration of variance attributable to specified factors can be very usefully extended, but interpretation



## Figure 1

Sources of variation. (A) Within a level of a factor (a sample). Variation around the mean in each sample is computed as the SS (individual value – mean of the samples at that level). (B) Variation within all the samples (individual values – mean of the values at that level). (C) Variation between the levels (using each group mean – overall mean).

can become more complex. We first show a simple result with the frogs: we compare males and females of three different species. The mean results are shown in Table 1.

The raw data and the means are also plotted in Figure 2. Plotting the means allows a better understanding of the patterns of the relationships. In Table 1, we can see two patterns. Looking at the *marginal means* (so called because the values

## Table 1

Mean values of samples

| Sex/species | Bullfrog | Leopard frog | Tree frog | Mean for sex |
|---|---|---|---|---|
| Males | 536 | 564 | 618 | 572 |
| Females | 590 | 620 | 679 | 630 |
| Mean for species | 563 | 592 | 649 | |

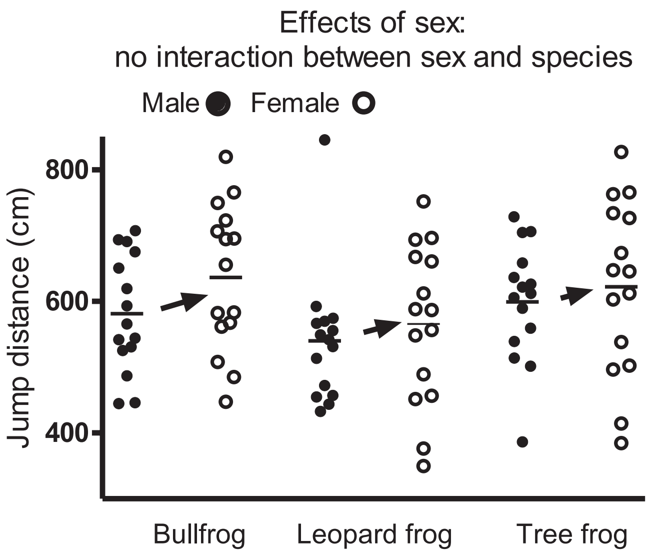
The marginal mean values (row of mean for species, column of mean for sex) indicate the main effects of species and sex.

## Table 2

The results of the analysis of variance

| Source of variation | d.f. | SS | MS | *F* value | *P* value |
|---|---|---|---|---|---|
| Interaction | 2 | 518.5 | 259.2 | 0.02050 | 0.9797 |
| Species | 2 | 228 638 | 114 319 | 9.041 | 0.0002 |
| Sex | 1 | 147 490 | 147 490 | 11.66 | 0.0008 |
| Residual | 174 | 2 200 000 | 12 644 | | |

The mean squares (MS) are calculated from the sum of squares (SS) and the d.f. The d.f. for sex are 1 (there are only two possibilities; if a result is not categorized as one sex, then there is only one further possibility), and for species, the d.f. is 2 (there are three species). *F* is the ratio of the MS between groups/MS within groups, when considering that factor.



## Figure 2

Main effects and interactions. Here, we have two main effects, species and sex, but no interaction between them. Sex has the same absolute effect, irrespective of species.

are at the ends of the rows and columns), there is a *main effect* for species, as the marginal means differ, and also a *main effect* for sex. When ANOVA is applied, the variances that are compared to assess these effects are

For species: $F$ = MS (between species)/MS (within species)

For sex: $F$ = MS (between sexes)/MS (within sex)

The ANOVA table is shown in Table 2.

In Figure 2, the difference between male and female frogs remains just about the same in each species. Although sex does have an effect, this effect is not different in the different species. This is interpreted as *no interaction* between these two factors, species and sex. It is often much easier to see this sort of relationship in a plot of the means.

When used carefully, there may be a place here for more elaborate figures to assess interactions. Let us look at the effect of dietary supplements on the three species (Figure 3). In this case, a combined figure emphasizes the effect. A further feature – interaction – can be seen clearly: a super diet improves performance most in the species that already jumps the furthest. In addition to finding a main effect of species, and a main effect of diet, there is *interaction* between diet and species. Two-way ANOVA quantifies this interaction by analysing the variance attributable to the combination of the factors. It is wise to check for interactions first, in the process of analysis. Multiple ANOVAS do not have to stop with two sources of variation, but graphical presentation of more effects becomes a problem!

ANOVA may not necessarily be the best approach to analysis of several factors. This analysis is based on categories. Even if one of the sources of variation is a graded or continuous factor such as dose, each level is considered independently.

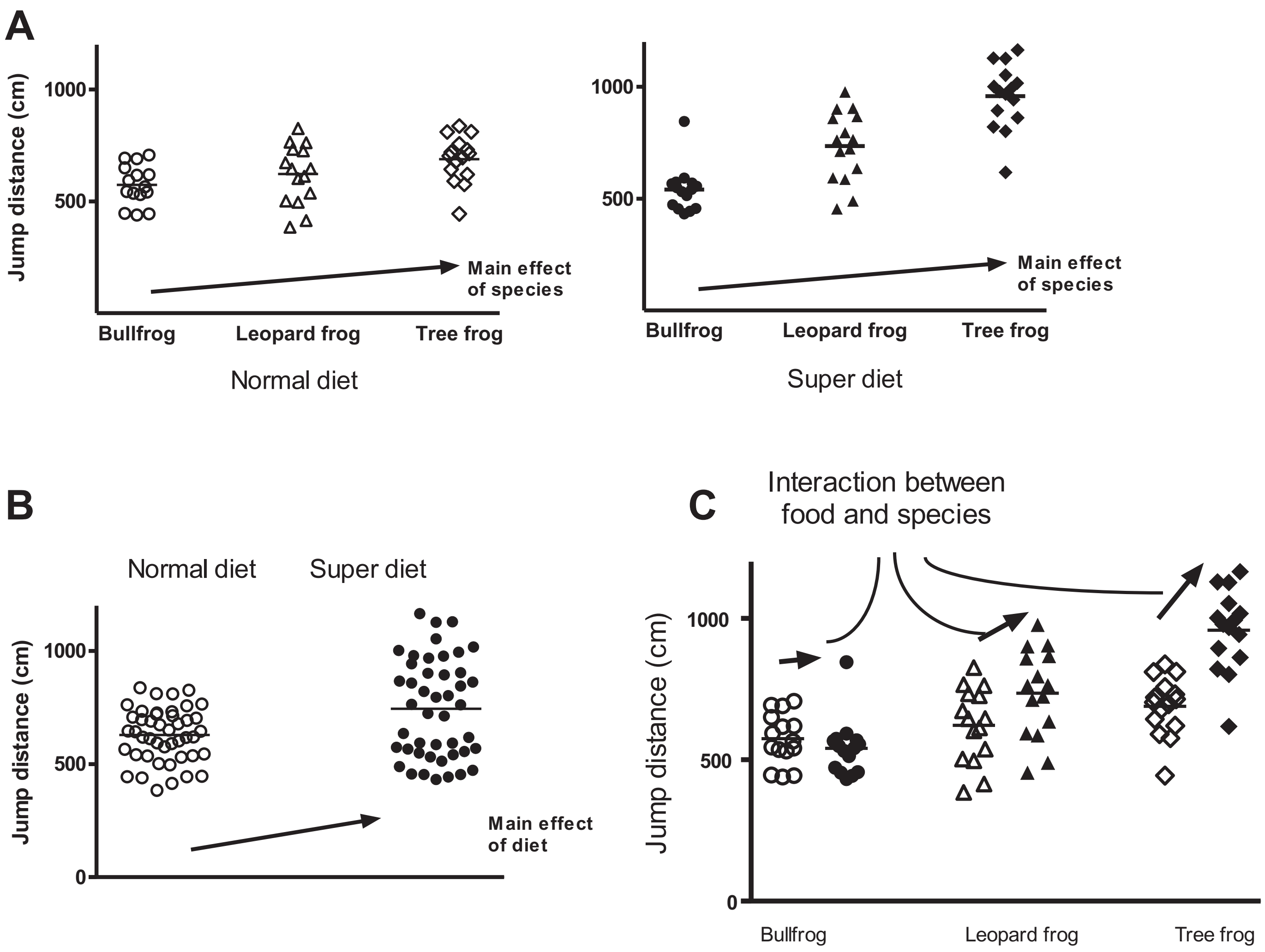


**Figure 3**

Here, there are two main effects evident. (A) There is an effect of species, in both the group fed the normal diet and the group fed the super diet. (B) There is an effect of diet, in all the species. (C) However, the effect of diet is greater in some species. The effects of diet and species interact positively: this is an interaction.

Thus, a small dose, maximal dose and supramaximal dose are considered as independent doses. It is more appropriate to analyse data of this sort using a dose–response relationship, that is by regression, a linear model or a sigmoid dose–response curve. Similarly, lumping continuously distributed factors into categories such as small, medium and large can be wasteful because there could be substantial variation within these groups that is not taken into account.

Another source of confusion is repeated measures. How do frogs grow when we feed them a special diet? If we measure a sample each week, then the small frogs are likely to put on less weight than the big frogs in our sample. In other words, the outcome is related to the preceding value; this phenomenon of *circularity* can influence the results. Often, the problem can be avoided by reducing the number of measurements: why not wait for a longer time and then measure the final weight of the frogs? If the repeated measures are of interest then a repeated measures ANOVA can be used, but this is trickier.

This is a very general account of a widely used test. There are a lot of possible pitfalls in the more complex versions; careful planning of the study and good advice, in equal measure, are important ingredients in the successful use of ANOVA.

## References

Drummond GB, Tom BDM (2011a). How can we tell if frogs jump further? Br J Pharmacol 164: 209–212.

Drummond GB, Tom BDM (2011b). Statistics, probability, significance, likelihood: words mean what we define them to mean. Br J Pharmacol 164: 1573–1576.