Kozak, M. and Powers, S. J. 2017. If not multiple comparisons, then what? *Annals of Applied Biology.* 171 (3), pp. 277-280.

Annals of Applied Biology ISSN 0003-4746

EDITORIAL

If not multiple comparisons, then what?

M. Kozak[1] & S.J. Powers[2]

[1] Department of Botany, Faculty of Agriculture and Biology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland
[2] Department of Computational and Analytical Sciences, Rothamsted Research, Harpenden, UK

Correspondence
M. Kozak, Department of Botany, Faculty of
Agriculture and Biology, Warsaw University of
Life Sciences – SGGW, 02-776 Warsaw, Poland.
Email: nyggus@gmail.com; and S.J. Powers,
Department of Computational and Analytical
Sciences, Rothamsted Research, Harpenden
AL5 2JQ, UK. Email:
stephen.powers@rothamsted.ac.uk

doi:10.1111/aab.12379

Suppose that you apply analysis of variance (ANOVA) for a designed experiment. The F-test rejects the general hypothesis that the treatment means are the same. You check the model by graphing residuals; perhaps you need to transform the data, but eventually everything seems all right. What do you do now?

Many researchers will apply a multiple comparison procedure, such as Bonferroni's, Duncan's, Tukey's, Student–Newman–Keuls', Sidak's, Dunnett's, Scheffé's, or their simpler alternative, the least significant difference (LSD). However, before deciding to use any of these procedures, you should first ask the following question: *Should I make all pairwise comparisons of treatments at all*? Or, even better: *Why would I want to compare all pairs of treatments*?

We hope that, after reading this editorial, you will answer the questions like this: *No, I do not have to compare all pairs of treatments, and so I do not have to make all pairwise comparisons.* And then, instead of comparing all pairs of treatments, you will compare only those treatments you indeed wish to compare.

Multiple testing in pairwise comparisons and how to deal with it

Here is what we are often told about comparing treatments: when making multiple comparisons, remember that you are doing many pairwise statistical tests at the same time. Forgetting about this, you inflate the chance of rejecting a null hypothesis that is true (such rejection is called the type I error). To solve this problem, you should adjust the significance level for the individual tests. Such adjustment guards against being overly optimistic about the results of the many tests done: in other words, when a 5% ($\alpha = 0.05$) significance level is used for many tests, then this 5% level should not be kept for each of them independently. If it is kept – that is, if you use the 5% significance level for each test – then you should expect 5% of these tests to give significance just by chance.

To illustrate this problem of multiple testing, Carter (2010) escapes from statistics to a commonplace context. He considers manufacturing bicycles in which a corporate goal is to ensure that no more than 1 in 20 (5%) of the bicycles can be defective, defective meaning here that at least one component of a bicycle is not functioning properly. And now Carter asks: 'Is it reasonable to expect that no more than 1 in 20 bicycles will be defective if all of the components have a 1/20 chance of being defective?' This illustration indeed explains why one should adjust the significance level for multiple testing *when all tests are important*.

When *all tests are important*, different multiple comparison procedures adjust the significance level for the individual tests in different ways, but they all aim to assure that the overall significance level is still 5%. In the Bonferroni procedure, the individual tests are at level:

$$\frac{0.05}{c\,(c-1)\,/2},$$

where c is the number of means to compare; thus, the overall denominator is the number of pairwise tests between all means. For example, if $c = 10$, there are 45 tests, and the level for each test is 0.0011 (0.11%). Yes, as low as that! So, imagine that you are especially interested in comparing one pair of means among these 45. If you use the Bonferroni procedure, you use the 0.0011 significance level to compare these two means at the 0.05

significance level. Clearly, you should not be happy about doing this!

Working with small significance levels for individual tests, the Bonferroni procedure gives conservative testing. Thus, many other procedures have been invented for pairwise comparison of many means. Although they try to be less severe than the Bonferroni approach in decreasing the significance level for the individual tests, they all share the aim, which is to control the overall significance level.

Now the question is: *Should we always compare all pairs of means?* A simple answer is: no, not always. In fact, almost never should we compare all pairs of means. Instead, we should focus only on those pairs which we are interested in comparing. For example, consider a two-way experiment in which every level of one factor occurs with every level of the other, and we use ANOVA to test their main effects and interaction (such an experiment uses a fully crossed two-way design). If the interaction is significant, it makes no biological sense to compare some treatments in a two-way table of means. To illustrate, let us suppose that the two crossed factors are cultivar with three levels (C1, C2 and C3) and growth promoter with three levels (GP1, GP2 and GP3), and that we aim to find the combination giving the greatest yield. Why, therefore, would we wish to test the difference between the mean for cultivar C1 using growth promoter GP1 and the mean for cultivar C2 using growth promoter GP2 when neither of these combinations gives the greatest yield, which happens to be provided by cultivar C3 with growth promoter GP3?

Here is the point, then: in most situations in which ANOVA is applied to data from designed experiments, multiple pairwise comparisons *should not be made whatsoever*! Although some of you might find this recommendation restrictive, we show below that it is rational from both statistical and biological points of view.

Any known structure among the treatments can and should be assessed as part of ANOVA, using nested or crossed treatment factors and/or contrasts. Here, ‘nested’ means that levels of one factor occur within levels of another: for example, potato cultivars can be nested in cooking type (Muttucumaru *et al.*, 2017). Contrasts combine particular levels of factors to give further comparisons of interest. For example, in a field trial of wheat varieties, we can use contrasts to compare groups of varieties with high and low disease resistance. Any treatment structure is known before (*a priori*) the analysis and may already represent all the relevant hypotheses (tested using the F-test) required to answer biological questions posed. So, whenever possible, we should formulate ANOVA to give the answers to these *a priori* biological questions.

A special group of contrasts is *orthogonal contrasts* (see, e.g. Quinn & Keough, 2002). They are independent to one another, which means that they account for separate portions of the variance due to treatments in ANOVA. In other words, if you add the sums of squares for such contrasts, you will obtain the overall sum of squares for treatments. Hence, orthogonal contrasts offer an alternative representation or a further partitioning of an already known treatment structure.

Whether using orthogonal contrasts or, more broadly, linear contrasts, such a clear-cut formulation of a full set of treatments is not always possible. What is more, even when it is possible, quite often after the analysis we get interested in comparing means we did not plan to compare before the analysis. Such comparisons of means that we come up with after the general ANOVA are called *post hoc* comparisons of means. It might be tempting to make these comparisons with the multiple comparison procedures we discussed before. But here is the point again: *Why should we compare all pairs of means when we are not interested in comparing some, or even most, of them whatsoever?*

We can make *post hoc* comparisons of most interest using the LSD value, calculated from the relevant standard error of the difference (SED) from ANOVA. In biology, we normally use the 5% significance level; a lower level (e.g. 1%) can allay fears about the type I error. Admittedly, in a study with a single treatment factor comprising a small number of treatments (typically three or four), all pairwise comparisons may be acceptable. However, note that even if there are only three treatments in an experiment, comparing just two of them may suffice for *post hoc* statistical assessment. Also, note that if the treatments are levels of a factor on a quantitative scale, then, more often than not, we should use regression modelling. We might then want to model how the dependent trait responds to the changing quantitative level of the factor. If after careful attempts to model the relationship we fail to explain its shape in any sensible form, we can still choose to treat the quantitative factor as a qualitative one using ANOVA. Do not, however, treat this approach as a regular one but as a last resort.

What about unbalanced experiments, in which treatments have unequal numbers of replications? Each comparison may then have a different SED value. In this case, we still should make only *post hoc* comparisons of most interest, but we should provide SEDs and LSDs for each of them.

Good practice: example

Let us consider an example of how to avoid all pairwise comparisons. Muttucumaru *et al.* (2017) combined data from identical potato trials at two different sites. At each site, 20 varieties were grown in a randomised block design with three blocks. Tubers from each plot were analysed for content of the contaminant acrylamide, amino acids and sugars at 2 months or 6 months after harvest. What is more, each variety was a member of (nested in) one of three types (boiling, crisping or French fry), giving an *a priori* treatment structure of interest. So, for ANOVA, Muttucumaru *et al.* (2017) used the following treatment structure:

$$\begin{aligned}&\text{Site (2 levels)}\times\\&[\text{variety (20 levels) nested in type (3 levels)}]\times\\&\text{storage (2 levels)}.\end{aligned}$$

When the full interaction between these factors was significant ($P < 0.05$ for the F-test), there were as many as ${}^{80}C_2$ (80 choose 2) = 3160 possible pairwise comparisons of means that could have been made. Can you imagine a table representing all these comparisons? Do not: clearly, most of them were unimportant. So why make them? Instead, the authors used the LSD at the 5% significance level to compare pairs of means of interest. Because the most important response of interest in the study was acrylamide – a probable carcinogenic processing contaminant (Halford *et al.,* 2011) – the authors were interested in (a) varieties with the highest or lowest means of this chemical, at either site, and (b) in the effect of storage for these varieties. So, instead of comparing 3160 pairs of means, the authors conducted a few (10, to be precise) *post hoc* comparisons of means *without adjusting for multiple testing*, a sensible approach to interpret so complex a treatment structure.

What do others say?

If you review applied biology literature, you will find that this approach of making only comparisons of interest instead of all of them has rarely been followed. Readers have to struggle through tables with means supported by letters 'abcdefg…'. Seeing such a long chain of letters, does anyone feel like interpreting them? Would such interpretation lead to anything constructive? Let us see what other statisticians say about the issue.

The eminent statistician John Nelder FRS was nicely blunt about the matter: 'In my view, multiple comparison methods have no place at all in the interpretation of data' (Nelder, 1971). Webster (2006) is also clear in his opinion: 'Do not use experiment-wise multiple comparison tests'. He refers to an argument between those who advise using LSDs and those who criticise LSDs in favour of multiple comparison procedures. 'Investigators who compare every pair of means by one of the above-mentioned tests', he says, 'seem not to appreciate the difference between a whole experiment, for which these techniques have been developed, and individual comparisons of interest… But in applying such stringent tests they fail to detect differences that they should identify and penalize themselves for the efficiency of their experimental designs. The experiment per se has not been the object of study, and so they should not apply an experiment-wise test'.

Welham & Clark (2006) refer to authors who criticise multiple comparison procedures for obscuring conclusions of experiments and producing contradictory results. 'These authors all agree', they state, 'that multiple comparison tests are inappropriate for analysis of experiments with a factorial treatment structure, and that interpretation of the patterns in the main effects/interactions found to be significant is more informative'. Of the authors referenced by Welham and Clark, Perry (1986) argues that even when the structure among the treatments cannot be accounted for in ANOVA, multiple comparison tests are still unhelpful. Instead, he says, it is better to compare all means by graphical methods to make biologically sensible groups of them.

More on comparisons of means: *P*-values and the effect size

Above, we focused more on statistics than biology. So let us turn to biology and ask how we can use statistics to assess the *biological effect* of a treatment.

When we compare two means (not *all* pairs of means), we should beware of mixing up statistical and biological significance. Instead of focusing on a P-value for the difference, we should rather focus on the importance of the size of the difference, the so-called *effect size*. To judge treatment effects, more important than hypothesis testing is estimation of means for treatments and of the differences between these means. When interpreting the size of a difference between means, one can use the confidence interval around that difference. Readers will often find the lower (upper) limit of such a confidence interval for a positive (negative) difference useful: *How far away from zero is it?* is the question relevant to concluding about the treatment effect. Is this metalled road not better than the stony track of making many unrequired pairwise comparisons without biological meaning? Webster (2006) would agree with us that it is: he said, '… remember that the mean values are the most important outcomes of almost all investigations. Emphasize them and draw readers' attention to the magnitude of their differences. Statistical significance is of secondary importance'.

Annals of Applied Biology and multiple comparisons

Annals of Applied Biology does not accept multiple comparison procedures. Instead, the authors are advised to present the SED or LSD, calculated directly from the appropriate stratum of ANOVA. In this journal, McNicol (2013) already explained these requirements. He also suggested that, in some scenarios, ranking treatments by their means would suffice to find the best treatments (i.e. the ones with the highest or lowest means) or to group treatments with similar means of the analysed trait.

What about more complex designs than those with only one treatment factor? ‘The process of uncovering the nature of the differences among the levels of a treatment [factor]’, McNicol says, ‘is more interesting in the case of significant interaction among two or more treatments [factors]. The principles are the same but the use of simple two-way plots of means, trellis plots or biplots greatly assists in the detection of any patterns. And it is general patterns or trends rather than specific pairwise comparisons, or worse all possible pairwise comparisons, which should be the focus’.

We are not going to claim that multiple comparisons should never be made. Instead, we wish to emphasise that, when making pairwise comparisons using the LSD, adjusting the significance level to account for multiple testing makes sense only when the analyst is certain that more than just a few biologically meaningful *post hoc* comparisons are required. *And such a situation is, frankly, very rare.* But what do we mean by ‘a few’? The answer is simply the number of ‘biologically meaningful comparisons’. Our experience suggests that multiple comparisons in agricultural sciences seldom call for adjustment for multiple testing.

References

Carter R.E. (2010) A simple illustration for the need of multiple comparison procedures. *Teaching Statistics*, **32**, 90–91.

Halford N.G., Curtis T.Y., Muttucumaru N., Postles J., Mottram D.S. (2011) Sugars in crop plants. *Annals of Applied Biology*, **158**, 1–25.

McNicol J. (2013) Significance testing after analysis of variance. *Annals of Applied Biology*, **162**, 288–289.

Muttucumaru N., Powers S.J., Elmore J.S., Dodson A., Briddon A., Mottram D.S., Halford N.G. (2017) Acrylamide-forming potential of potatoes grown at different locations, and the ratio of free asparagine to reducing sugars at which free asparagine becomes a limiting factor for acrylamide formation. *Food Chemistry*, **220**, 76–86.

Nelder J. (1971) Discussion on papers by Wynn, Bloomfield, O’Neill and Wetherill. *Journal of the Royal Statistical Society, B*, **33**, 244–246.

Perry J.N. (1986) Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology*, **79**, 1149–1155.

Quinn G.P., Keough M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge, UK: Cambridge University Press.

Webster R. (2006) Analysis of variance, inference, multiple comparisons and sampling effects in soil research. *European Journal of Soil Science*, **58**, 74–82.

Welham S.J., Clark S.J. (2006) Issues concerning the design and analysis of comparative field experiments. *Outlooks on Pest Management*, **17**, 175–180.