Behavioral Ecology
doi:10.1093/beheco/arn020
Advance Access publication 24 February 2008

Forum

Time for some a priori thinking about post hoc testing

Graeme D. Ruxton[a] **and Guy Beauchamp**[b]
[a]Institute of Biomedical Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK and [b]Faculty of Veterinary Medicine, University of Montréal, PO Box 5000, St-Hyacinthe, Québec, Canada J2S 7C6

Researchers are commonly in a situation, often after an experiment, where they want to compare the central tendency of some measure across a number of groups. If the number of groups is simply 2, then there is little controversy as to the appropriate analysis, with normally a *t*-test or a nonparametric equivalent being adopted. If the number of groups is greater than 2, most elementary statistical textbooks suggest performing an analysis of variance (ANOVA) to test the null hypothesis that all the groups are the same and, if this null hypothesis is rejected, implementing some post hoc testing to identify which groups are significantly different from which other groups.

However, as readers and reviewers of scientific papers in behavioral science, we have noted a great diversity of approaches when comparing more than 2 groups often with little or no justification for the adoption of a specific approach. Hence, our aim in this note is to briefly survey current practice in this regard and to provide clear guidance on how such testing might most appropriately be carried out in different instances.

WHAT DO RESEARCHERS CURRENTLY DO?

We surveyed the 12 issues making up the last complete volumes of *Behavioral Ecology* (volume 18) and *Animal Behaviour* (volume 73). We found 70 papers where the authors compared central tendencies across more than 2 groups. Two of these papers presented a set of planned comparisons that did not involve testing for homogeneity across the whole set of groups; however, such papers were highly exceptional. In the remaining 68 cases, analysis involved testing for homogeneity across all groups, using an ANOVA or a nonparametric equivalent (generally the Kruskal–Wallis test). In these papers, comparison between specific groups never occurred if the null hypothesis of homogeneity across all groups could not be rejected. However, such comparisons invariably followed if the null hypothesis of homogeneity across groups was rejected. These comparisons were done using a very wide diversity of different tests (see Table 1).

Table 1 illustrates that there is a great deal of variation in current practice. It is not clear how much of this variation is driven by careful matching of tests to variation in research questions and approaches and how much of it is driven by the adoption of suboptimal practices, perhaps driven by options available from commonly used software packages. In order to examine this situation, we must first state how we consider that such analysis might most effectively be performed. This is presented in the next section.

ASPECTS OF GOOD PRACTICE

Researchers should strive for planned comparisons

When an experiment is carried out, we presume that the researchers often have specific hypotheses in mind and can write down before collecting the data the specific comparisons between groups or between combinations of groups that they are interested in so as to test these specific hypotheses. In observational studies, it is also often possible to use prior knowledge to formulate hypotheses. In some cases, the goal in observational studies is to explore differences between groups with no firm expectation as to the strength and direction of effects. In this case, all pairwise comparisons will be of interest, and we discuss appropriate treatment of this case (which we call "unplanned comparisons") in a following section.

Nevertheless, in many cases, researchers will often be able to produce a set of "planned comparisons." Wherever possible, we recommend that researchers do this. In general, the set of planned comparisons will be a subset of the set of all possible comparisons, and so the researchers can save themselves from the risk of inflating type I error rate through making uninteresting comparisons. If the set of planned comparisons encompasses all possible comparisons, we recommend treatment as if carrying out unplanned comparisons (see following section). However, we expect such cases to be the exception (especially if the number of groups is greater than 3), and there are several advantages to carrying out a set of planned comparisons:

i) By only testing the comparisons they are interested in, researchers simplify their analysis and reduce the risk of type I errors. If a researcher adopts the convention of 95% confidence, then they accept a 5% risk of a type I error in each comparison they make. As the number of comparisons involved in the analysis of an experiment goes up, so does the likelihood of type I errors. There are procedures (such as Tukey testing) that can control the experimentwise type I error rate (EER), but such control comes at a cost in statistical power, and the more comparisons are involved the more power is lost to maintain a given level of control. Thus, removing comparisons that are not of scientific interest should allow a more attractive trade-off between type I and type II errors.
ii) If every planned comparison derives from a specific hypothesis, then interpretation of each comparison that is performed should be straightforward.
iii) Complex comparisons (e.g., comparing treatment A against the aggregate grouping of treatments B, C, and D) can easily be accommodated.
iv) The appropriate analysis for a set of planned comparisons is simple conceptually and practically as shown below.
v) If each planned comparison tests a different specific hypothesis, then many influential texts consider that no formal control of EER is required (e.g., Kirk 1995; Sokal and Rohlf 1995; Quinn and Keough 2002).

Because phrases like "EERs" are commonly used in the literature, but with a confusing diversity of definitions, we should be careful to define what we mean in this paper. By our definition, if we set EER to α, then this means that we constrain the probability of at least one of our set of comparisons relating to a specific experiment yielding a type I error to be α.

Table 1

Frequency of use of different tests following rejection of a test of homogeneity of central tendancy across a number of groups greater than 2

Test	Tukey[a]	Duncan's multiple range	Fisher's LSD	SNK[b]	Undefined "post hoc test"	*F*-test or *t*-test	Dunnett	Dunn's	By eye	*U*-test[c]
#	20	1	12	2	12	8	1	2	4	6

[a] Tukey test, Tukey–Kramer, or other derivatives.

[b] SNK (sometimes called simply Neuman–Keuls).

[c] Wilcoxon–Mann–Whitney.

Implementing a set of planned comparisons

If a set of planned comparisons can be justified on the basis of theory or previous results, then this provides a more attractive means of producing the greatest power while still controlling EER.

Because each of these comparisons tests a unique hypothesis, all comparisons should be performed no matter the outcome of the ANOVA. Each planned comparison is carried out by partitioning the sum of squares from the ANOVA or by a *t*-test. Detailed instructions can be found in Sokal and Rohlf (1995) and Quinn and Keough (2002). In fact, the 2 methods are functionally equivalent and always give identical results. We present the *t*-test because it has a more compact presentation. Simply, in comparing groups A and B, the *t*-test is performed in the usual way except that the standard error in the comparison is given by

$$\sqrt{\left(\frac{1}{n_{\mathrm{A}}}+\frac{1}{n_{\mathrm{B}}}\right)\mathrm{MS}_{\mathrm{residual}}},$$

where n_{A} and n_{B} are the sample sizes and $\mathrm{MS}_{\mathrm{residual}}$ is the residual mean square from an ANOVA calculation.

As already discussed, some influential texts suggest that no formal control of EER may be required. However, this is an area of some controversy, and we suggest a more cautious line: that formal control of EER is not required if the set of contrasts is "orthogonal."

Orthogonality is a very strict set of conditions on the independence of the contrasts. One of its conditions is that the same term never appears in 2 different contrasts. Consider the case where we have 3 groups: A, B, and C. If there are k groups, then there are $k - 1$ orthogonal contrasts, so in this case there are 2. If one of our contrasts is "A versus B," then the only orthogonal contrast is "C versus the aggregate of A and B." "A versus B" and "B versus C" do not form a set of orthogonal contrasts because the same term "B" appears in 2 contrasts.

Our advice is that researchers should write down the list of comparisons that they are interested in, prior to collecting the data. Let us assume that the experiment involves k groups.

a) In the special case, where this list encompasses all possible pairwise comparisons between groups (with k groups, this involves $k(k - 1)/2$ comparisons), they should control EER by performing a set of unplanned comparisons (see below).
b) If the number of comparisons is greater than $k - 1$, then this must be a nonorthogonal set, and we recommend a planned comparison with control of EER.
c) If the number of comparisons is exactly $k - 1$, then the researchers should check to see whether these comparisons form an orthogonal set (for clear instructions on this, see Sokal and Rohlf 1995; Crawley 2005). Only if they are orthogonal is no formal control of EER required.
d) If the number of comparisons is less than $k - 1$, then they should be tested to see if further hypotheses can be added to the list to create an orthogonal set. If this is possible, then no formal control of EER is required, otherwise it is. Note that there is no need to actually perform the added comparisons because they are presumably not of interest.

EER can be controlled using, for example, the sequential Bonferroni technique. Such error control in the context of planned comparisons is discussed in depth by Castaneda et al. (1993). An alternative method is the Dunn–Sidak method, which is fully described by Sokal and Rohlf (1995). These methods control type I error rate at the cost of a loss of power. Recent methods that strike a possibly more attractive compromise between the 2 types of errors are also available (for a discussion of this in the context of ecological studies, see Waite and Campbell 2006).

Implementing a set of unplanned comparisons

A set of planned comparisons that encompasses all pairwise comparisons should be treated in the same way as a set of unplanned comparisons, as emphasized by Sokal and Rohlf (1995). In addition, if there are no prior results or underlying theoretical framework to guide analysis and the purpose of analysis is to explore the data, then unplanned comparisons are appropriate, and we consider the Scheffe procedure to be the most appropriate response to a significant ANOVA.

The Scheffe procedure is the only multiple comparison procedure that is entirely coherent with ANOVA results. That is, with the Scheffe procedure, the researcher is guaranteed that if the ANOVA suggested a difference between groups then at least one of the Scheffe comparisons will be significant at the same level; whereas if the ANOVA is not significant, then neither will be any of the Scheffe comparisons. This is not true for any other test of unplanned comparisons (Castaneda et al. 1993). The Scheffe procedure also has the desirable properties of 1) having similar robustness properties to assumptions of normality and homogeneity of variance as ANOVA, 2) allowing different sample sizes in each group, 3) allowing great flexibility in comparisons including comparisons involving more than 2 groups (e.g., comparing group A with the aggregate of groups B and C), and 4) allowing control of EER.

However, the Scheffe procedure is more conservative than any of the alternative procedures for unplanned comparisons. This occurs because the Scheffe critical value is derived from an unlimited number of pairwise and complex comparisons of the means. Because, in a research context, all the possible comparisons cannot be defined or interpreted, this is an unattractive characteristic of the Scheffe procedure. This may cause the researcher to turn to alternatives, but these alternatives do not have the property of coherence with the ANOVA. Thus, we consider the commonly adopted policy of first performing an ANOVA and only investigating post hoc comparisons if the ANOVA is significant to be logically flawed. Rather, we recommend that when an ANOVA is carried out and produces a significant result, then it should be followed by the Scheffe procedure. If any procedure other than Scheffe's is used (see directly below), then it should be implemented regardless of the outcome of the ANOVA.

There are many alternatives to the Scheffe procedure for performing unplanned comparisons (for a comprehensive review, see Day and Quinn 1989). We do not recommend Fisher's protected least significant difference (LSD) test, Duncan's multiple range test, or the Student–Neuman–Keuls (SNK) test, none of which can be relied on to keep EER at or below the nominal level (Day and Quinn 1989; Quinn and Keough 2002).

The most commonly used appropriate alternative is the Tukey honestly significantly differenced test (sometimes called the *T*-test). Ryan's test (sometimes called the Ryan–Einot–Gabriel–Welsch test) is much more challenging to calculate than the Tukey test but is a little more powerful. For a very thorough comparison of alternative procedures that reaches the same conclusions, see Toothaker (1993). With unequal sample sizes, there are a number of modifications to the Tukey test; the most common of these is the Tukey–Kramer test, but the T′ and GT2 tests are also appropriate. Sokal and Rohlf (1995) recommend implementing all 3 and then selecting the one with the lowest minimum significant difference (the product of critical value of the test statistic and the standard error used in the comparisons).

Dunnett's test is recommended for tests in a situation where a specified group (generally a control) is compared with each of the other groups. Although this is a set of planned contrasts and thus could be evaluated in that way, the contrasts are clearly nonorthogonal, and the Dunnett's test is preferable in terms of a combination of statistical power and convenience, relative to any means of controlling EER in a set of planned comparisons.

The nonparametric case

The Kruskal–Wallis test is the nonparametric equivalent of 1-way ANOVA. Pairwise multiple comparison procedures based on all possible rank comparisons are available (including Dunnett's T3, Dunnett's C, and Games–Howell tests; Kirk 1995); Toothaker (1993) recommends the last of these and provided a recipe for implementation. Sokal and Rohlf (1995) recommend a simultaneous test procedure, but this requires equal sample sizes in each group. Probably the most commonly used method for pairwise multiple comparisons without making assumptions about normality is the Dunn procedure, as laid out in Zar (1999). Thus, for performing all pairwise comparisons, we would recommend either the Games–Howell procedure (as described by Toothaker 1993) or the Dunn procedure from Zar (1999). For a small set of planned comparisons, we would recommend Mann–Whitney *U*-tests with control of EER using one of the methods discussed above for parametric comparisons.

THE CASE FOR A CHANGE IN BEHAVIOR AMONG RESEARCHERS

We can now ask how the behavior of current researchers (based on our survey of recent issues of *Behavioral Ecology* and *Animal Behaviour*) squares against aspects of our suggested best practice.

First, we suggested that researchers should use planned comparisons wherever appropriate. In our survey of 70 papers, only 11 used planned comparisons. Of the remaining 59, we suggest that many would have adopted unplanned comparisons (at the cost of reduced statistical power) needlessly. In many of the relevant papers, we found that clear hypotheses, which could have led to a set of planned comparisons, could be identified from reading the introduction to the paper. In some cases, this set of comparisons may have encompassed all possible pairwise comparisons, and so performing an unplanned set of comparisons was not inappropriate; but in many cases, we were not convinced that all the possible pairwise comparisons carried out by their unplanned analysis were of interest to the researchers, and so adoption of a set of planned comparisons would often have given them a better balance between EER and statistical power.

Next, we suggest that in planned comparisons, the omnibus test of homogeneity across all groups should only be done if this is an explicit planned comparison. Only 2 of the 11 papers presenting a set of planned comparisons justified the appropriateness of the omnibus test, yet 9 of them performed such a test. Only 2 of the 9 papers implemented EER control for their planned comparisons.

For unplanned comparisons, we suggest either the Sheffe procedure be implemented after the omnibus test suggests a significant effect or an alternative comparison procedure be adopted no matter the result of the omnibus test. None of the 59 papers presenting unplanned comparisons adopted this approach. None used the Scheffe procedure, and all only ever made comparisons on subsets of the groups if (and only if) the omnibus test rejected the null hypothesis of homogeneity.

Instead of Scheffe's procedure for unplanned comparisons, we recommend the Tukey test or the Ryan test. The minority adopted these tests (20 adopting the Tukey or its derivatives and none adopting the Ryan). Of the others, Fisher's LSD was popular, but we do not recommend this because of its unreliability in controlling type I error. The same problem is true for the less commonly used Duncan and SNK tests. Clearly, evaluation of differences between groups on the basis of visual inspection of the means is more subjective than is necessary. In many of the cases involving a set of *F*-tests or *t*-tests and in all the cases using Wilcoxon–Mann–Whitney tests, there was no attempt to EER, control that is an integral part of the Tukey procedure.

We strongly recommend that researchers adopt planned comparisons whenever this is practical. This approach encourages focussing on biologically meaningful results. In this regard, it is important for researchers to also consider what constitutes a biologically interesting effect (as distinct from a statistically significant one). Although we have focussed in this paper on hypothesis testing, we do believe that effect sizes and confidence intervals should be given more prominence as part of a discussion of the biological relevance of observed results (for further discussion, see Colegrave and Ruxton 2003; Nakagawa and Cuthill 2007). This is because effect size estimates provide information about the magnitude of a "result" and the precision of the estimate of the effect. This is particularly relevant for nonsignificant pairwise results that can arise for a number of reasons including the statistical method used, the lack of an effect, or sample size. Planned comparisons may result in a greater inspection of biologically interesting effects.

Although we have couched our discussion of multiple comparisons in terms of simple comparisons between groups, as would be performed using a 1-way ANOVA or Kruskal–Wallis test, our main issue has wider applicability. In comparing between groups (e.g., in a survival analysis), one should avoid test procedures that make all pairwise comparisons if only a small subset of these comparisons are of interest. Sometimes a smaller number of planned comparisons combined with control of EER might give a more attractive trade-off between type I and type II errors. One particular situation that may prove problematic from the point of view of planned comparisons involves repeated-measures designs in which differences between the same groups are tested repeatedly over several time periods. With a large number of time periods, the number of comparisons between groups can increase very rapidly resulting in a loss of power to detect any differences between groups at any particular time. If it is not clear at which time period differences between groups will arise, then we suggest implementing a strict control of EER using all relevant contrasts as explained above. However, if differences are expected at some times but not at others, one could use EER control at each time period separately.

CONCLUSION

Statistical testing of comparisons among a number of groups remains a common occurrence in behavioral science. However, our survey suggests that common practice is highly variable and almost always suboptimal, with researchers suffering from lower power and/or higher type I error rates than are necessary. We hope that this paper presents a template for carrying out much more efficient analyses of this kind.

There is a need for researchers to take this type of analysis more seriously, 12 of our survey of 70 papers failed even to state the test they used, but rather described the procedure as "post hoc testing." The overwhelming majority of cases among the 70 do not provide calculated test values where appropriate, but simply give *P* values. Happily, the template for more effective testing laid out above is not really significantly more complex or time consuming that many of the suboptimal practices currently adopted, and small changes in practice should lead to significantly more powerful and reliable analysis in a commonly encountered situation.

We thank 2 anonymous referees for insightful comments.

Address correspondence to G.D. Ruxton. E-mail: g.ruxton@bio.gla.ac.uk.

Received 2 December 2007; revised 9 January 2008; accepted 10 January 2008.

REFERENCES

Castaneda MB, Levin JR, Dunham RB. 1993. Using planned comparisons in management research: a case for the Bonferroni procedure. J Manag. 19:707–724.

Colegrave N, Ruxton GD. 2003. Confidence intervals are a more useful compliment to nonsignificant tests than are power calculations. Behav Ecol. 14:446–447.

Crawley MJ. 2005. Statistics: an introduction using R. New York: Wiley.

Day RW, Quinn GP. 1989. Comparisons of treatments after analysis of variance in ecology. Ecol Monogr. 59:433–463.

Kirk RE. 1995. Experimental design. Pacific Grove (CA): Brooks/Cole.

Nakagawa S, Cuthill IC. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev. 82:591–605.

Quinn GP, Keough MJ. 2002. Experimental design and data analysis for biologists. Cambridge (UK): Cambridge University Press.

Sokal RR, Rohlf FJ. 1995. Biometry. 3rd ed. New York: WH Freeman.

Toothaker LE. 1993. Multiple comparison procedures. London: Sage Publications.

Waite TA, Campbell LG. 2006. Controlling false discovery rate and increasing statistical power in ecological studies. Ecoscience. 13:439–442.

Zar JH. 1999. Biostatistical analysis. 4th ed. New York: Prentice Hall.