ELSEVIER

Infection, Genetics and Evolution 2 (2002) 149–152

Infection, Genetics and Evolution

www.elsevier.com/locate/meegid

Discussion

# Further difficulties with multifactorial analysis of variance: Random and nested factors and independence of data

David A. Morrison*

*Molecular Parasitology Unit, University of Technology Sydney, Westbourne Street, Gore Hill, NSW 2065, Australia*

Received 5 April 2002; received in revised form 18 July 2002; accepted 29 August 2002

## 1. Introduction

The statistical analysis of data sets containing more than one factor can be quite complex, and reviews of the biological literature have made it clear that biologists, in particular, frequently have problems performing them correctly (e.g. Underwood, 1981). Such data sets are most appropriately analysed by a multifactorial analysis of variance (ANOVA), where each factor in the analysis represents one factor in the experiment. Much of the difficulty for biologists seems to lie in the fact that when an analysis of variance involves more than one factor, then the calculations used for the *F*-ratios depend on whether the factors are random or fixed and whether they are nested or orthogonal (as described below). Here, I discuss an analysis of variance from the recent literature to illustrate the potential pitfalls when these characteristics are treated inappropriately, in the hope that such problems can be avoided in the future.

Hughes (2001) describes an investigation of the effects of selection on the long-term evolution of dengue viruses and their relatives, particularly in relation to cytotoxic T cell (CTL) epitopes. A phylogenetic tree was constructed based on the complete polyprotein amino acid sequences of various isolates of dengue virus, West Nile virus and Japanese encephalitis virus, thus allowing the reconstruction of their ancestral amino acid sequences. From this information, predicted CTL epitopes (pCTLE) were identified for each of the seven inferred ancestors and eight descendants on the tree. This allowed the calculation of the number of changes to the pCTLE along each of the 14 branches of the tree.

For those pCTLE that were inferred to have experienced one or more changes in their amino acid sequence along a branch, Hughes was interested in studying whether the pCTLE was lost or not (i.e. whether the pCTLE in the ancestral sequence ceased to constitute a pCTLE in the immediate descendant, as a result of the changes in the amino acid sequence). These data on the number of amino acid changes were "analysed by a nested analysis of variance." The model used for this analysis had three factors: (1) comparison of the viral groups (dengue virus versus its relatives), (2) comparison among the phylogenetic branches within the viral groups, and (3) comparison of whether the pCTLE was lost or not. The second factor was treated as nested within the first factor, and the third factor was treated as nested within the second factor.

There are three points that can be made about this statistical analysis: (i) the model used for the nested analysis treated all three of the factors as fixed, whereas the two nested factors should properly be treated as random, thus producing incorrect probabilities for two of the three factors and leading to some incorrect conclusions; (ii) a more appropriate model for analysing these data would treat the third factor as orthogonal to the other two, rather than nested within them, leading to a different set of probabilities and conclusions; and (iii) it is doubtful that these data are appropriate for statistical analysis at all, because the observations are not independent of each other, thus violating one of the fundamental assumptions of all statistical analyses. These points are discussed in this note.

## 2. Nested analysis of variance

An analysis of variance involves a statistical test of the differences between groups of replicate observations defined by a particular experimental "treatment"—a grouping variable is called a "factor" and the groups are called "levels". The calculations of the analysis produce an *F*-ratio, which is used to decide whether there is a statistically significant difference between the means of two or more of the groups. The *F*-ratio is calculated using the mean-squares, which measure the amount of variability both within and between the groups. When an analysis of variance involves more than

* Tel.: +61-2-9514-4159; fax: +61-2-9514-4003.
*E-mail address:* david.morrison@uts.edu.au (D.A. Morrison).


PII: S1567-1348(02)00096-5

one factor, then the calculations used for the *F*-ratios depend on whether the factors are random or fixed and whether they are nested or orthogonal.

For a factor to be considered fixed, all possible levels of the factor that are of interest for the experimental question must have been sampled for the analysis. For a factor to be considered random, only a sample of the possible levels of the factor will have been included in the analysis. If factors are orthogonal then every possible combination of the levels of the factors is included in the analysis. For nested factors, different levels of one of the factors occur in combination with only one of the levels of the other factors. The main point for our purposes here is that, while the levels of the upper factor in a nested model can be random or fixed, the levels of all of the factors nested within the upper factor must be treated as random (Sokal and Rohlf, 1994).

In contrast, in the ANOVA model used by Hughes all three of the factors were treated as fixed, whereas the two nested factors should have been treated as random. This has the consequence that the calculations for two of the *F*-ratios are incorrect, and thus the probabilities associated with them are also incorrect. Table 1 shows the derivation of the calculations for the *F*-ratios from the analysis presented by Hughes, as well as for the more correct analysis (see Winer et al., 1991; Zar, 1999). Note that the difference between the analyses is that when factors are treated as fixed then the appropriate denominator for the *F*-ratios is the Residual term, whereas this is not true for random factors. For factors with other factors nested within them, the appropriate denominator is actually the nested factor (i.e. if B is nested within A then B is the appropriate denominator for A, and if C is further nested within B then C is the appropriate denominator for B). This difference can have a large impact on the magnitude of the value of the *F*-ratio, and therefore on the associated probability.

Clearly, in this example, the conclusions from the two analyses are very different. Instead of rejecting the null hypothesis for all the three factors, as in the published analysis, in the correct analysis we should only reject the null hypothesis for one of the factors, if we use the conventional Type I error rate of $P = 0.05$. Two of the biological conclusions reached by Hughes are thus incorrect. In particular, there is no significant difference between the two viral groups in the mean number of amino acid changes per epitope.

The use of the Residual as the denominator in the *F*-ratio also affects the calculation of the standard errors (S.E.), as it is the mean-square from the denominator that is used in this calculation as well. In this example, the correct values for the mean ± S.E. number of changes per epitope are $1.69 \pm 0.09$ for the dengue virus group and $1.39 \pm 0.12$ for their relatives (based on the data presented in Figs. 3 and 4 of Hughes' paper), as opposed to $1.69 \pm 0.04$ and $1.39 \pm 0.09$, respectively, which are the values cited by Hughes. This change in the calculations highlights the difference in the interpretation of this factor, because it is clear from the correct values that there can be no statistical difference between the two means.

It is worth pointing out that this confused situation regarding the appropriate denominator for the *F*-ratios in a nested ANOVA, which I am highlighting here, appears to arise quite commonly in biological publications. This presumably is because some statistical analysis computer programs, such as SAS (SAS Institute Inc., 2001) and SYSTAT (SPSS Inc., 2000), do not allow the user to explicitly specify which factors are fixed and which are random in an analysis of variance, although they do allow the user to specify which factors are nested and which are orthogonal. These programs always use the Residual as the denominator in the *F*-ratios, which is inappropriate for random factors, thus leading to results that require subsequent re-calculation by the user in order to arrive at the required solution. This re-calculation is easily performed using the program (or even a hand calculator), but the user needs to know that it is necessary to expressly instruct the program to perform it. Other statistical analysis computer programs, such as Minitab (Minitab Inc., 2000) and S-PLUS (Insightful Corporation, 2001), do allow the explicit specification of random factors in an analysis of variance, and these programs will automatically provide the required solution without any re-calculation. This difference among the computer programs is rarely highlighted in statistical textbooks, Glantz and Slinker (2001) being a notable exception.

## 3. A more appropriate model

In addition to the problem discussed in the previous section, there seems to be another difficulty with the model for the ANOVA used by Hughes. Hughes treated the factor for

Table 1
Nested analyses of variance for the experiment described by Hughes (2001)

| Source of variation | Nesting | d.f. | Mean-square | Analysis of Hughes | | | More correct analysis | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | *F*-ratio[a] | *F* | *P* | *F*-ratio[b] | *F* | *P* |
| Between virus groups | A | 1 | 6.37 | A/Res | 8.74 | 0.003 | A/B | 1.79 | 0.206 |
| Among branches | B (A) | 12 | 3.56 | B/Res | 4.89 | <0.001 | B/C | 1.15 | 0.397 |
| pCTLE lost or not | C (AB) | 14 | 3.10 | C/Res | 4.25 | <0.001 | C/Res | 4.25 | <0.001 |
| Residual | Res (ABC) | 777 | 0.73 | | | | | | |

[a] All three factors are fixed.
[b] Factor A is fixed, while factors B and C are random.

analysing the comparison of whether the pCTLE was lost or not as nested. This implies that the different levels of this factor occur in combination with only *one* of the levels of the other two factors. However, the design of the experiment is such that the levels of this factor actually occur in every possible combination with the levels of the other two factors, in which case this factor should be treated as orthogonal to the other two factors.

This point is made clear in Hughes' Fig. 4. In this figure the open bars represent the group of pCTLE that underwent one or more amino acid changes but were still retained as pCTLE, while the black bars represent pCTLE that were lost as a result of one or more amino acid changes, which are the two levels for this factor. The figure shows both black bars and open bars for every level of both of the other factors. That is, the top part of the figure represents the dengue virus group and the bottom part represent its relatives, which are the two levels for the factor comparing the viral groups, and both black and open bars are shown in the upper and lower parts of the figure (i.e. data are shown for all four combinations of the levels for this pair of factors). A similar comment can be made about the seven phylogenetic branches nested within each virus group (i.e. data are shown for all 14 combinations of the levels for this pair of factors). This being so, then the pCTLE factor is orthogonal to both of the other two factors in the experiment, and this fact must be used in the statistical analysis of the data.

A more appropriate model for the ANOVA is thus a mixed one, where some of the factors are random, some are fixed, some are nested and some of them are orthogonal (cf. Morrison, 2002). The factor for the comparison of the viral groups (dengue virus versus its relatives) is fixed; the factor for the comparison among the phylogenetic branches is random and is nested within the factor for the viral groups; and the factor for the comparison of whether the pCTLE were lost or not is fixed and orthogonal to the other two factors. The model for this more appropriate analysis is shown in Table 2. This table shows the derivation of the calculations for the *F*-ratios in the analysis, as well as those results that can be calculated in the absence of the original data. The procedures for determining the expected mean-squares, and therefore the appropriate *F*-ratios, for mixed analyses of variance are described by, for example, Winer et al. (1991) and Zar (1999), and are available in computer programs such as that of Dallal (1988). In particular, it is worth noting that the factor for whether pCTLE is lost or not has as its denominator the interaction between itself and the nested factor (i.e. among branches), which is a result of the fact that the nested factor is random.

Table 2
Mixed analysis of variance for the experiment described by Hughes (2001)

| Source of variation | Nesting | d.f. | *F*-ratio[a] | *F* | *P* |
|---|---|---|---|---|---|
| Between virus groups | A | 1 | A/B | 1.79 | 0.206 |
| Among branches | B (A) | 12 | B/Res | 4.89 | <0.001 |
| pCTLE lost or not | C | 1 | C/BC | –[b] | – |
| Group × pCTLE interaction | AC | 1 | AC/BC | – | – |
| Branch × pCTLE interaction | BC | 12 | BC/Res | – | – |
| Residual | Res (ABC) | 777 | | | |

[a] Factors A and C are fixed, while factor B is random.
[b] Cannot be calculated in the absence of the original data.

Note that there are two differences between this analysis and the one discussed in the previous section. First, the mean-square for the factor comparing whether the pCTLE were lost or not is treated very differently in the mixed analysis, because here it has been sub-divided to represent three sources of variation rather than one (as used in the nested analysis). The two extra sources of variation are the interactions between this factor and the other two factors, which were not statistically tested at all by the nested analysis.

If either of the two interactions in this new model is statistically significant then this would mean that the evolution of the pCTLE was quite different among the different phylogenetic branches of the different virus groups. This is the main importance of interactions in an analysis of variance—they tell us whether a particular biological phenomenon is "universal" throughout the experiment or whether, instead, the existence of the phenomenon depends on other factors within the experiment. This is a point that could be important to assess, and indeed this seems to have been one of the motivations for doing the experiment in the first place. In particular, if the group × pCTLE interaction is significant then it would mean that we could not make any generalisations about whether the pCTLE is lost or not, but would need to make different statements about the two groups of viruses.

The second difference between the two analyses is that the *F*-ratio for the factor comparing the phylogenetic branches is calculated differently in the mixed model compared to the nested model. In fact, the resulting *F*-value is the same as that reported by the original incorrect nested ANOVA, so that, in this example, this part of the original analysis turns out to have been right but for the wrong reason.

## 4. Assumptions of statistical analyses

All statistical analyses are based on a set of assumptions, without which it is impossible to calculate the final probabilities. The two most fundamentals of these assumptions are that the sampling of the data has been both random and independent. A set of randomly chosen sampling units requires that each potential unit has an equal probability of actually being sampled for the experiment. A set of independently chosen sampling units requires that no one unit influences the probability of any other potential unit being sampled for the experiment. The logic of experimental inference also needs these basic requirements. For example, if the units are independent of each other then they are free to behave in any manner during the experiment. Therefore,

the fact that they *do* behave in a predictable and repeatable manner must be a result of the experimental "treatments". If they are not independent then their behaviour could be a result of their underlying non-independence rather than the experimental treatments, and the experiment will thus demonstrate nothing worthwhile. So, independence is the basic experimental and statistical technique for avoiding the confounding of various sources of variation.

In this regard, there are two separate but related independence problems that can potentially occur when analysing data derived from a phylogenetic tree. First, branches on a phylogenetic tree cannot be independent of each other because of the ancestor–descendant relationships, as discussed in detail by Harvey and Pagel (1991). This is what evolution means, that what the descendants look like is constrained by what the ancestors looked like. Thus, genotypic changes along the branch of a phylogenetic tree cannot be independent of the ancestor; if this was not so, then there would be no genetic component to the evolutionary processes. This problem is clearly relative, in the sense that the more distantly related two branches are in an evolutionary sense then the more independent they are likely to be in a statistical sense. So, evolutionarily unrelated branches on a phylogenetic tree can probably be treated as statistically independent for practical purposes, and analyses performed using computer programs such as that of Purvis and Rambaut (1995). However, this is clearly not the case for the dengue viruses and their relatives, because the branches are very closely related, representing different isolates of sister species. Non-independence of the data in this example is therefore likely to be a serious problem from this source, notwithstanding the assumed rapid rate of evolution in these organisms (if the rate of evolution is faster then closely related branches will be more independent).

Second, a phylogenetic tree is often constructed from the very same data that are to be subsequently analysed. This clearly creates logical non-independence, because the characteristics of the branches cannot be independent of the data used to construct them. It is illogical to use data to test for a particular pattern if that pattern was created from those data in the first place, whether you are testing phylogenetic patterns or any other type of pattern. Unfortunately, for the dengue viruses and their relatives the statistical tests have been applied to precisely the same data that were used to construct the tree. Analysing the data under these circumstances as though they were independent is circular—the branches are likely to have certain patterns because they were constructed from a particular data set, and therefore it cannot surprise us when the analysis of these data reveals those very same patterns. This problem can only be solved by constructing the phylogenetic tree from data that are independent of the data to be analysed. Exactly how this might be done when using complete genome sequences of micro-organisms is not entirely obvious.

This problem of non-independence (from two different sources) is potentially the most serious problem for the analysis of this particular data set, because it means that no standard statistical analysis can be validly applied to these data, and therefore the probabilities derived from any of the ANOVA models discussed here will actually be nonsense. It might be appropriate to use some form of simulation procedure to assess these type of data (i.e. comparing the observed data to simulated data derived from the expectation specified by the null hypothesis), but this approach has not yet been developed sufficiently for use by non-experts.

## References

Dallal, G.E., 1988. DESIGN: A Supplementary Module for SYSTAT and SYGRAPH. Version 2.0. SYSTAT Inc., Evanston, IL.

Glantz, S.A., Slinker, B.K., 2001. Primer of Applied Regression and Analysis of Variance, 2nd ed. McGraw-Hill, New York.

Harvey, P.H., Pagel, M.D., 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, Oxford.

Hughes, A.L., 2001. Evolutionary change of predicted cytotoxic T cell epitopes of dengue virus. Infect. Gen. Evol. 1, 123–130.

Insightful Corporation, 2001. S-PLUS 6 for Windows. Insightful Corporation, Seattle, WA.

Minitab Inc., 2000. Minitab Release 13.1 for Windows. Minitab Inc., State College, PA.

Morrison, D.A., 2002. Inappropriate application of a model for mixed analysis of variance: some comments on Elena et al. (2001). Infect. Gen. Evol. 1, 303–305.

Purvis, A., Rambaut, A., 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Comp. Appl. Biosci. 11, 247–251.

SAS Institute Inc., 2001. SAS Version 8.2 for Windows. SAS Institute Inc., Cary, NC.

Sokal, R.R., Rohlf, F.J., 1994. Biometry: The Principles and Practice of Statistics in Biological Research, 3rd ed. Freeman, San Francisco.

SPSS Inc., 2000. SYSTAT 10 for Windows. SPSS Inc., Chicago, IL.

Underwood, A.J., 1981. Techniques of analysis of variance in experimental marine biology and ecology. Ann. Rev. Oceanogr. Mar. Biol. 19, 513–605.

Winer, B.J., Brown, D.R., Michels, K.M., 1991. Statistical Principles in Experimental Design, 3rd ed. McGraw-Hill, New York.

Zar, J.H., 1999. Biostatistical Analysis, 4th ed. Prentice-Hall, Upper Saddle River, NJ.