Stat Note 6

In the sixth of a series of articles about statistics for biologists, **Anthony Hilton** and **Richard Armstrong** discuss:

post hoc ANOVA tests

IN A PREVIOUS article in *Microbiologist* (Armstrong & Hilton, 2004), we described the application of analysis of variance (ANOVA) to various experimental designs in Microbiology.

ANOVA is a data analysis method of great elegance, utility and flexibility and is the most effective method available for analysing experimental data in which several treatments or factors are represented. In the simplest case of a one-way ANOVA, in which the experiment consists of a number of independent treatments or groups, the first stage of the analysis is to carry out a variance ratio test (F-test) to determine whether all group means are the same. If treatment groups are few, say three or four, a non-significant F-test would indicate no meaningful differences among the means and no further analysis would be required. However, a significant F-test suggests real differences among the treatment means and the next stage of the analysis would involve a more detailed examination of these differences.

There are various options available depending on the objectives of the experiment. Specific comparisons may have been planned before the experiment was carried out, decided after the data have been collected, or comparisons between all possible combinations of the treatment means may be envisaged. This Statnote provides a more detailed discussion of these questions than was possible in our original article (Armstrong & Hilton, 2004).

The scenario

An experiment was designed to investigate the efficacy of two commercial plasmid-prep kits compared to a standard alkaline-SDS lysis protocol. A 5 ml overnight recombinant *E. coli* culture containing a high copy plasmid was harvested by centrifugation and the pellet resuspended in 100 μl of lysis buffer. Plasmid DNA was subsequently extracted from the cell suspension using a standard SDS-lysis protocol or a commercially available kit following the manufacturer's instructions. In total, ten independent cultures were processed using each of the three extraction methods under investigation. Following extraction the purified plasmid DNA pellet was dissolved in 50 μl of water and the concentration determined spectrophotometrically at 260 nm. The yield of plasmid DNA using each preparation method is detailed in Table 1.

Planned comparisons between the means

The experiment may have been designed to test specific ('planned') differences between the treatment means. Planned comparisons are hypotheses specified *before* the analysis commences whereas *'post-hoc'* tests are for further explanation after a significant effect has been found.

How are the tests done?

The basic strategy for planned comparisons is to divide up the treatments sums of squares among the various hypotheses, called ‘contrasts’, which are then analysed separately either by an F-test or a t-test. If this procedure was carried out for all possible comparisons between the means, then the sums of squares for all contrasts would be greater than the treatments sums of squares as a whole since the comparisons overlap and based on the same sources of variance. Strictly, such comparisons cannot be made independently of each other. As a result, comparisons must be constructed so that they are not overlapping, i.e., they have to be *‘orthogonal.’* Essentially, orthogonal comparisons have no common variance and their coefficients sum to zero. Hence, the sums of squares can be calculated for each contrast and a test of significance made on each. The number of possible contrasts is equivalent to the number of degrees of freedom (DF) of the treatment groups in the experiment. Hence, if an experiment employs three groups, as in our scenario, then two contrasts can be validly tested. This approach has two advantages. First, there is no problem as to the validity of the individual comparisons, a problem present to some extent with all conventional *post-hoc* tests. Second, the comparisons provide direct tests of the hypotheses of interest. Most commercially available software will allow for valid contrasts to be tested for a range of experimental designs.

Table 1. Comparison of two commercial plasmid-prep kits (plasmid yield mg) compared to a standard alkaline-SDS lysis protocol using planned comparisons and *post-hoc* tests.

Culture	Alkaline-SDS lysis	Commercial kit A	Commercial kit B
1	1.7	3.1	4.7
2	2	2.2	3.5
3	1.2	2.8	2.6
4	0.5	4.8	4.3
5	0.9	5	3.8
6	1	1.9	4.5
7	1.4	2	4
8	2.7	3.6	1.9
9	3.2	4.1	2.8
10	0.7	4.7	4.6

ANOVA

Source of variation	Sums of squares	DF	Mean square	F
Treatments	27.3807	2	13.690	13.28
(P<0.001)				
Error	27.998	27	1.0370	

Planned comparisons

Contrast	Estimate	Std. error (SE)	‘t’
1. Std. v (Kit A + Kit B)/2	4.03	0.79	5.109
(P<0.001)			
2.Kit A v Kit B)	0.25	0.45	0.54
(P>0.05)			

***Post-hoc* tests**

Test	Std. v Kit A	Std. v Kit B	Kit A v Kit B
Fisher PLSD	P<0.001	P<0.001	P<0.05
Tukey-Kramer HSD	P<0.001	P<0.001	P<0.05
SNK	P<0.001	P<0.001	P<0.05
Scheffé	P<0.001	P<0.001	P<0.05

An illustrative example

An example of this approach is shown in Table 1. In our scenario, we compared two commercial plasmid-prep kits with a standard alkaline-SDS lysis protocol. Two valid contrasts are possible using this experimental design. First, a comparison of the mean of the two-commercial prep kits with the standard method, *viz.*, do the commercial kits on average improve plasmid yield (contrast 1)? Second, a comparison of the two commercial prep kits themselves (contrast 2). Contrast 1 is highly significant ($t = 5.11$, $P < 0.001$) indicating the superiority of the commercial kits over the standard method but contrast 2 is not significant ($t = 0.54$, $P > 0.05$) showing that the two commercial kits did not differ in their efficacy.

Post-hoc tests

There may be circumstances in which tests between the treatment means are carried out *post hoc* or where multiple comparisons between the treatment means may be required. A variety of methods exist for making post-hoc tests. The most common tests included in commercially available statistical software are listed in Table 2 (Abacus Concepts, 1993; Armstrong *et al.*, 2001). These tests determine the critical differences that have to be exceeded by a pair of treatment means to be significant. However, the individual tests vary in how effectively they address a particular statistical problem and their sensitivity to violations of the assumptions of ANOVA. The most critical problem is the possibility of making a Type 1 error, i.e., rejecting the null hypothesis when it is true. By contrast, a Type 2 error is accepting the null hypothesis when a real difference is present. The *post-hoc* tests listed in Table 2 give varying degrees of protection against making a Type 1 error.

Discussion of the tests

Fisher’s protected least significant difference (Fisher’s PLSD) is the most ‘liberal’ of the methods discussed and therefore the most likely to result in a Type 1 error. All possible pairwise comparisons are evaluated and the method uses Student’s ‘t’ to determine the critical value to be exceeded for any pair of means based on the maximum number of steps between the smallest and largest mean. The Tukey-Kramer honestly significant difference (Tukey-Kramer HSD) is similar to the Fisher PLSD but is less liable to result in a Type 1 error. In addition, the method uses the more conservative ‘Studentised range’ rather than Student’s ‘t’ to determine a single critical value that all comparisons must exceed for significance. This method can be used for experiments that have equal numbers of observations (N) in each group or in cases where ‘N’ varies significantly between groups. However, with modest variations in N, the Spjotvoll-Stoline modification of the above method can be used. The Student-Newman-Keuls (SNK) method makes all pairwise comparisons of the means ordered from the smallest to the largest using a stepwise procedure. First, the means furthest apart, i.e., ‘a’ steps apart in the range, are tested. If this mean difference is significant, the means a-2, a-3, etc., steps apart are tested until a test produces a non-significant mean difference, after which the analysis is terminated. The SNK test is more liable to make a Type 2 rather than a Type 1 error.

By contrast, the Tukey compromise method employs the average of the HSD and SNK critical values. Duncan's multiple range test is very similar to the SNK method, but is more liberal than SNK, the probability of making a Type 1 error increasing with the number of means analysed. One of the most popular methods is Scheffé's 'S' test. This method makes all pairwise comparisons between the means and is a very robust procedure to violations of the assumptions associated with ANOVA (Armstrong & Hilton, 2004). It is also the most conservative of the methods discussed giving maximum protection against making a Type 1 error. The Games-Howell method is one of the most robust of the newer methods. It can be used in circumstances where 'N' varies between groups, with heterogeneous variances (see Statnote 5), and when normality cannot be assumed. This method defines a different critical value for each pairwise comparison and this is determined by the variances and numbers of observations in each group under comparison. Dunnett's test is used when several treatment means are each compared to a control mean. Equal or unequal 'N' can be analysed and the method is not sensitive to heterogenous variances. An alternative to this test is the Bonferroni/Dunn method that can also be employed to test multiple comparisons between treatment means especially when a large number of treatments is present.

Which test to use?

In many circumstances, different *post-hoc* tests may lead to the same conclusions and which of the above tests is actually used is often a matter of fashion or personal taste. However, each test addresses the statistical problems in a unique way. A good way of deciding which test to use is to consider the purpose of the experimental investigation. If the purpose is to decide which of a group of treatments is likely to have an effect, then it is better to use a more liberal test such as Fisher's PLSD. In this scenario it is better not to miss a possible effect. By contrast, if the objective is to be as certain as possible that a particular treatment does have an effect then a more conservative test such as the Scheffé's test would be appropriate. Tukey's HSD and the compromise method fall between the two extremes and the Student-Newman-Keuls (SNK) method is also a good choice. We would also recommend the use of Dunnett's method when several treatments are being compared with a control mean. However, none of these methods is an effective substitute for an experiment designed specifically to make planned comparisons between the treatment means.

An illustrative example

As an example, we analysed data from our scenario using four different post-hoc tests, viz., Fishers PLSD, Tukey-Kramer HSD, the SNK procedure and by Scheffé's test (Table 1). In this example, the results are clear cut and all four tests lead to the same conclusion, i.e., both commercial kits are superior to the standard method but there is no difference between commercial kits A and B thus confirming the results of the planned comparisons.

Conclusion

If data are analysed using ANOVA, and a significant F value obtained, a more detailed analysis of the differences between the treatment means will be required. The best option is to plan specific comparisons among the treatment means before the experiment is carried out and test them using 'contrasts'. In some circumstances, *post-hoc* tests may be necessary and experimenters should think carefully which of the many tests available should be used. Different tests can lead to different conclusions and careful consideration as to the appropriate test should be given in each circumstance.

Table 2. Features of the most commonly used post-hoc tests (modified from Abacus Concepts 1993 and Armstrong *et al*., 2000)

Method	Equal N F	Normality	Use	Error control	Protection
Fisher PLSD	Yes	Yes	Yes	All	Most sensitive to Type 1
Tukey-Kramer HSD	No	Yes	Yes	All	Less sensitive to Type 1 than Fisher PLSD
Spjotvoll-Stoline	No	Yes	Yes	All	As Tukey-Kramer
Student-Newman Keuls (SNK)	Yes	Yes	Yes	All	Sensitive to Type 2
Tukey-Compromise	No	Yes	Yes	All	Average of Tukey and SNK
Duncan's Multiple Range	No	Yes	Yes	All	More sensitive to Type 1 than SNK
Scheffé's S	Yes	No	No	All	Most conservative
Games/Howell	Yes	No	No	All	More conservative than majority
Dunnett's test	No	No	No	T/C	More conservative than majority
Bonferroni	No	Yes	Yes	All, TC	Conservative

Abbreviations: PLSD = Protected least significant difference, HSD = Honestly significant difference.
T = treatment groups, C = Control group, Column 2 indicates whether equal numbers of replicates (N) in each treatment group are required or whether the method can be applied to cases with unequal 'N'. Column 3 indicates whether a significant between treatments F ratio is required before post-hoc tests can be applied and columns 4 and 5 whether the method assumes equal variances in the different treatments and normality of errors respectively. The final column indicates the relative degree of protection against type 1 and type 2 errors.

References

- Abacus Concepts (1993) SuperANOVA. Abacus Concepts Inc., Berkeley CA 94704, USA.
- Armstrong R A, Slade S V & Eperjesi F (2000) An introduction to analysis of variance (ANOVA) with special reference to clinical experiments in optometry. *Ophthal Physiol Opt* **20**: 235-241.
- Armstrong R A & Hilton A (2004) The use of analysis of variance (ANOVA) in applied microbiology. *Microbiologist* vol 5: **No.4** 18-21.
- Hilton A & Armstrong R A (2006) Is one set of data more variable than another? *Microbiologist* Vol. 7: **No.2** 34-36 (June 2006)

Dr Anthony* Hilton and Dr Richard Armstrong**
*Pharmaceutical Sciences and **Vision Sciences, Aston University, Birmingham, UK