MULTIPLE COMPARISON PROCEDURES

D.J. Saville
Biometrics Unit, Ministry of Agriculture and Fisheries, Lincoln

ABSTRACT

Procedures such as Duncan's multiple range test, Fisher's restricted least significant difference test and the Waller and Duncan test are used by agronomists for the analysis of data from experiments in which the treatments have no easily defined structure. These and similar procedures, referred to generally as multiple comparison procedures, have been the subject of controversy in the statistical world for over half a century.

Different procedures are known to give markedly different results in some cases. Examples can easily be found in which one procedure declares a difference nonsignificant and another procedure declares the same difference 1% significant. What is not well known is that all but one of the procedures are "inconsistent" within themselves. This paper produces examples in which a particular multiple comparison procedure declares a given difference nonsignificant in one experiment but declares the same difference (with the same standard error) 1% significant in a second experiment, with the only change being in the number of treatments in the two experiments or in the values of the other treatment means.

The only "consistent" procedure is the unrestricted least significant difference procedure. This is also the simplest and most powerful of the procedures. For these and other reasons it is the procedure the author would recommend to agronomists.

Additional Key Words: Duncan's test, LSD test, Waller and Duncan test, Tukey's test, inconsistency, type I error rate, power.

INTRODUCTION

Multiple comparison procedures have been widely used and misused by agronomists over the last few decades. Their valid use is in the analysis of data from studies in which the experimental treatments have no definable structure. However, they have also been misused in the analysis of data from studies in which the treatments have a clearly defined structure (e.g. treatments follow a 2 x 2 factorial design, or consist of 5 rates of seeding). This has been pointed out by Little (1978) and other writers in the applied science journals.

Multiple comparison procedures have probably generated more confusion and controversy among applied researchers and statisticians than any other statistical tool. The fact that new procedures are continually being invented is evidence of the inadequacy of the existing procedures. It is also a reflection upon the ill defined nature of the search for a "good" multiple comparison procedure. This paper will reflect upon the nature of this search and point out the inconsistency inherent in most of the resulting procedures.

REVIEW

The simplest multiple comparison procedure consists of t tests comparing all possible pairs of treatments using a pooled variance estimate, s^2. This procedure will be referred to as the "unrestricted LSD" procedure. The word "LSD" is used since a least significant difference (LSD) can be calculated to serve as a critical value for the difference between any pair of treatment means. The word "unrestricted" is used to distinguish the procedure from Fisher's restricted LSD procedure in which the t tests are carried out only if a preliminary overall F test is significant.

Fisher suggested the preliminary overall F test as a means of adding a greater level of conservatism to the procedure since he was unhappy with the idea of carrying out a multitude of unplanned tests of poorly formulated hypotheses. His ad hoc modification to the natural procedure started the search for the "perfect" multiple comparison procedure. Unfortunately this search had no clear objectives and has consequently reached no satisfactory conclusion. The only real guideline in this search is the requirement that any resulting procedure should be "more conservative" than the unrestricted LSD procedure. This is such a loose requirement that an infinity of solutions are possible. To date several scores of solutions have been proposed in the statistical literature and new solutions are proposed each year.

To reduce the number of possible solutions, we must consider what requirements we should demand from a multiple comparison procedure. One possible requirement is that it be "true to label" in terms of comparisonwise type I error rate. That is, a 5% level procedure should spuriously declare nonexistent differences to be real at the rate of 5% of null comparisons. This requirement cuts the infinity of solutions down to just one solution: the unrestricted LSD procedure. Another possible requirement is that the procedure should have a 5% experimentwise error rate, i.e. a nonexistent difference is declared real in just 5% of experiments. This reduces the possibilities down to just Tukey's procedure. However, this procedure is widely regarded as being far too conservative. The requirement forces the comparisonwise type I error rate to decrease with increasing number of treatments: this in turn forces the type II error rate to increase, resulting in a decrease in the power of the test.

CONSISTENCY

An intuitively reasonable requirement is that a given procedure should be "consistent" in the decisions it produces about whether two particular treatments are truly different. More precisely, the decision should depend only on the magnitude of the difference between the two treatment means, the standard error, and the error degrees of freedom. It should not depend on the number of treatments included in the experiment nor the observed means for the other treatments.

To illustrate this view we shall borrow the terminology of Carmer and Walker (1982) and consider the case of a statistician, Goldilocks, who has three clients, Baby Bear, Mama Bear and Papa Bear. The Bears are all avid porridge eaters so are keen to breed new varieties of porridge with high yields. The three Bears each carry out an experiment which includes four porridge varieties. By chance there are 2 varieties, A and B, which are common to all three experiments. The design for each experiment is completely randomised, with five replicates per treatment, allowing 16 degrees of freedom for error. Figure 1 gives the mean yield, in thousands of breakfast portions per hectare, for each treatment for (a) Baby Bear, (b) Mama Bear and (c) Papa Bear. For the two common varieties, similar data is observed in all three experiments so that the observed means for varieties A and B are 20.0 and 24.3 in all three experiments. The pooled standard error of the mean also turns out to be the same, at 1.00, in all three experiments.

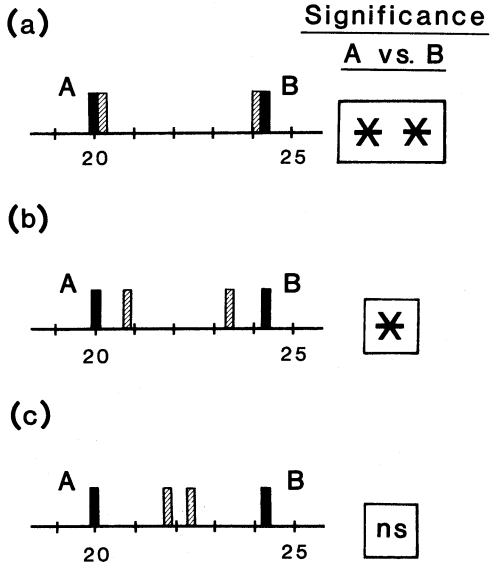

Figure 1: Significance of the difference between populations A and B in 3 experiments statistically analysed using Fisher's restricted LSD procedure. (Each vertical bar represents a treatment mean).

Goldilocks analysed the data provided by each Bear using Fisher's restricted LSD procedure. One morning over porridge the Bears got into an argument about whether they should grow porridge B in preference to porridge A. Baby Bear said porridge B was definitely better than porridge A since in his experiment the difference between the two varieties was significant at the 1% level. Mama Bear agreed with Baby Bear since in her experiment the difference was 5% significant. However, Papa Bear disputed whether there was a real difference since in his experiment the observed difference was not significant. As the argument progressed the three Bears realised something quite peculiar. They had all observed the same difference (4.3), they all had the same level of variability (se(mean) = 1) and yet the level of significance for this difference varied from not significant to 1% significant. Clearly Goldilocks had made a mistake during the analysis of their datasets.

That afternoon the Bears visited Goldilocks to point out the inconsistency between their results. No, there was no mistake, said Goldilocks. Baby Bear's overall F value of 5.62 was 1% significant, so he was entitled to declare a 1% significant result for the t test comparing variety A with B. Mama Bear's overall F value of 4.21 was only 5% significant, so she only deserved a 5% significant result for the same t test. Papa Bear had failed to achieve significance with his overall F value of 3.14, so he was not entitled to any significant differences.

Papa Bear was outraged. Why should he be penalised just because he was unlucky in his choice of treatments, he raged. Papa Bear, being very quick witted, had realised that the problem was that in his experiment the 2 other varieties had observed means about halfway between the means for varieties A and B whereas Baby and Mama Bear had observed means which were more spread out. He figured that in future he would have to include one of the old varieties of porridge (with a mean yield of about 15) to ensure that he passed Goldilocks' preliminary F test. However, this made him very annoyed. Why should he have to do extra field work just to convince Goldilocks of something that Papa Bear already knew: that the varieties were not all identical. Though the Bears accepted Goldilocks' view of what was acceptable to statisticians, their opinion of the statistical community was not improved by their discussion.

The inconsistency illustrated by the above occurs with all procedures except the unrestricted LSD procedure. Examples such as depicted in Figure 1 can be readily constructed for most procedures. The author has constructed such a figure for Tukey's procedure by varying the number of treatments from 2 to 4 to 8. For Waller and Duncan's k-ratio LSD test, he has constructed a similar figure using three experiments each with 7 treatments.

As one would expect, the level of inconsistency varies with procedure. Tukey's procedure can be relied upon to be inconsistent, since its critical LSD value depends simply on the number of treatments in the experiment. For Tukey's and Fisher's LSD procedures it is easy to construct examples in which the significance of A versus B varies from not significant to 0.1% significant. For Waller and

Duncan's procedure it requires a little thought to construct examples in which the significance of A versus B varies from not significant to 1% significant (k-ratio = 500). For Duncan's multiple range test (unrestricted) it is only possible to construct examples in which the significance of A versus B varies from not significant to 5% significant. In summary, Tukey's procedure is consistently at the inconsistent end of the spectrum and Duncan's multiple range test is close to the consistent end of the spectrum. However, the unrestricted LSD is the only totally consistent procedure.

DISCUSSION

Interestingly enough it is Duncan's multiple range test, the most consistent of the alternatives to the unrestricted LSD, which has enjoyed the greatest acceptance among agronomists. Duncan's test is also the least conservative of the alternative procedures and the procedure which is most similar to the unrestricted LSD procedure. In fact, one can speculate that the t test (unrestricted LSD) is subconsciously accepted by agronomists as the "standard" so that any procedure which is too different is unacceptable.

When Duncan's multiple range test was first introduced in New Zealand, "Duncan's letters" were introduced as a way of reporting differences between treatments: those with a letter in common had been judged not different and those without a letter in common had been judged different. This sales gimmick proved very popular and was part of the reason for the widespread acceptance of the test. The letters can of course be used in conjunction with any multiple comparison procedure, including the unrestricted LSD procedure.

Usage of Duncan's multiple range test was a compromise satisfactory to both statisticians and agronomists. It did something to ease the disquiet among the statisticians, while the agronomists retained most of the power of the unrestricted LSD test. If subjected to cost/benefit analysis however, the usage of Duncan's test cannot be justified. On theoretical grounds it is inferior to the unrestricted LSD procedure in that it is less powerful, has a variable type I error rate and is inconsistent, albeit only to a small extent. On practical grounds it is computationally much more complicated than the unrestricted LSD and has undoubtedly led to an increase in the number of wrong conclusions based on undetected computational errors. In recent years many field scientists in New Zealand have wisely reverted to using the unrestricted LSD procedure.

Another procedure which has enjoyed some usage in New Zealand is the Waller and Duncan's k-ratio LSD test. When used with k-ratios of 100 and 500, this test often produces results similar to those obtained from 5% and 1% unrestricted LSD tests. It is a suitable compromise procedure since it reduces the disquiet among statisticians while retaining the power of the unrestricted LSD. However, a cost/benefit analysis would also find it sorely lacking.

Comparative studies by Carmer and Swanson (1971, 1973) come out in favour of Fisher's restricted LSD procedure and Waller and Duncan's k-ratio LSD procedure. The unrestricted LSD, although superior in their simulations in terms of type II error rate, is rejected because it "unduly deemphasises protection against type I errors". In a later paper, Carmer and Walker (1982) suggest, on the grounds of simplicity, that Fisher's restricted LSD is the procedure which is most suitable for agronomists. The evidence presented in these papers, however, also supports the view that the unrestricted LSD is the most suitable procedure.

More recently O'Brien (1983) questioned the appropriateness of multiple comparison procedures, raising among other issues that of inconsistency. His conclusion was that simple t tests were most appropriate, although he took the more extreme view that variances should not be pooled between treatments.

The author's view is that the unrestricted LSD procedure (consisting of pairwise t tests using a pooled variance estimate) is the most appropriate of the multiple comparison procedures. It is the natural extension of the 2 population case, has the virtue of simplicity and has a constant comparisonwise type I error rate (e.g. 5% for a "5% level" procedure). The calculated LSD can also be used to derive a confidence interval for the difference between two population means, viz. observed difference ± LSD.

"Data-dredging" is a term often used in connection with multiple comparison procedures; this means "looking for large differences in the data then testing them for significance". Fear of data-dredging is one of the main reasons why statisticians would like to be conservative in the multiple comparison case. However, an alternative to being conservative is to accept the dubiety of forming and testing an hypothesis using the same dataset.

To elaborate on this point, we digress to consider the hypothesis testing scenario which is acceptable to statisticians. This is the scenario of a well designed study in which orthogonal contrasts are prespecified to correspond to a "vision of reality" which will hopefully be supported by the data. If the original vision is not supported by the data, however, it is sometimes found that another set of orthogonal contrasts provides a good description of the data. This second set would normally be used in the data presentation, however, an honest researcher would make it clear that the new vision of reality had been formed from the data and still required confirmation.

When a multiple comparison procedure is used for data analysis it is clear that the researcher is using the nonorthogonal pairwise comparison contrasts for "vision formulation", not "vision testing". For example, with the unrestricted LSD procedure the calculated 5% and 1% level LSD's are used simply as a yardstick to give some indication as to which differences are likely to be real differences. Researchers using this procedure realise that if in fact there are no differences among their populations, a

5% level LSD will in the long run produce "false significances" at the rate of 5%. This knowledge is informally integrated with subject matter knowledge in the vision reformulation process. Subsequent experimentation is then necessary for the vision testing.

In agronomic research conclusions are normally based on evidence from a range of sources, including perhaps several field trials. Most researchers would in fact be reluctant to make any firm statements using the results from just a single trial. In this context it is not necessary to treat the data from each individual trial as though it was the only data available in the world.

CONCLUSION

This paper has attempted to show that the search for a perfect multiple comparison procedure is like the search for the gold at the end of the rainbow. In fact, the natural procedure, the unrestricted LSD, has more good characteristics than any of its competitors. This writer would therefore suggest that the unrestricted LSD procedure should be used when usage of a multiple comparison procedure is appropriate.

ACKNOWLEDGEMENTS

Dr F. Jackson Hills, extension agronomist and biometrician, University of California, Davis is thanked for stimulating the work on this paper. Dr. H.V. Henderson, biometrician, Ruakura Agricultural Research Centre, Hamilton; Dr. G.R. Wood, Mathematics Department, University of Canterbury, Christchurch and Mr. M.P. Ryan, Department of Statistics, Christchurch are thanked for helpful discussions. Mr G. Arnold, Department of Mathematics and Statistics, Massey University, Palmerston North is also thanked for his constructive suggestions.

REFERENCES

Carmer, S.G., Swanson, M.R. 1971. Detection of differences between means: A Monte Carlo study of five pairwise multiple comparison procedures. *Agronomy Journal 63:* 940-945.

Carmer, S.G., Swanson, M.R. 1973. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association 68:* 66-74.

Carmer, S.G., Walker, W.M. 1982. Baby Bear's dilemma: A statistical tale. *Agronomy Journal 74:* 122-124.

Little, T.M. 1978. If Galileo published in HortScience. *HortScience 13:* 504-506.

O'Brien, P.C. 1983. The appropriateness of analysis of variance and multiple comparison procedures. *Biometrics 39:* 787-794.