RANKING AND SELECTION: WHAT TO DO IF YOUR ANOVA REJECTS THE NULL HYPOTHESIS

H.P. EDWARDS

Department of Mathematics and Statistics, Massey University, Palmerston North

SUMMARY

The analysis of variance (ANOVA) is the most widely used statistical test of homogeneity of treatment means. However, it is of little benefit if a significant result occurs or if the experimenter assumes that some differences exist *a priori*. Thus a more realistic approach to the problem might be to ask: which treatments have the largest (or smallest) treatment means? Problems such as this are called selection problems, and these are introduced and discussed (along with related problems called ranking problems). Procedures to solve these problems are considered, and a computer package designed to implement the procedures is mentioned.

INTRODUCTION

Consider the situation where there are k treatments and we wish to infer something about the treatment means. The usual statistical approach to this situation is the test of homogeneity, i.e. a test to see whether the means differ or not, and the test procedure employed is the well known analysis of variance or ANOVA for short. Whilst the ANOVA is entirely appropriate as a test procedure here, it may fail to provide the sort of information that the experimenter requires. In particular, if the test results in confirmation of significant differences, it does not in any way explain how the means differ. Indeed, there are many situations where the experimenter knows *a priori* that the treatment means are not all the same, and a non-significant result for an ANOVA merely indicates that the test was not sensitive enough to detect the underlying differences. There are of course extensions to the test of homogeneity, such as multiple range tests, but none of these is appropriate if it is assumed from the outset that differences between the means exist. What is required is some means of making an inference about those differences.

DEFINITION OF RANKING AND SELECTION PROBLEMS

Ranking problem: how to rank the treatments from worst to best to be able to say (with some confidence P*) that the ranking is correct.

Selection problem: how to select some subset of the treatments to be able to say (with some confidence P*) that the subset contains the "best" treatments (where "best" may refer to those with the largest means or those with the smallest means).

Both problems may be generalised or even combined, so for simplicity we shall restrict our attention to the selection problem. A full discussion and presentation of statistical ranking and selection may be found in the text by Gibbons *et al* (1977) from which the following example is taken:

Example Four varieties of corn are to be compared on the basis of yield in bushels/acre. There are seven replicates of each treatment. The results are presented in Table 1. We shall further assume that there is only one "best" treatment, i.e. we are only interested in the treatment with the largest mean yield.

There are three types of selection problems, and we shall use the data given in Table 1 to illustrate each type.

INDIFFERENCE ZONE SELECTION

Here it is desired to select the best treatment only; more generally, the selected subset contains all and only the best treatments. Hence a Correct Selection (CS for

TABLE 1: Yields of corn (bushels/acre)

	Lancaster	Clark	Silver King	Osterland
	4.6	5.3	12.1	7.8
	5.2	6.1	15.9	7.5
	3.9	7:2	9.2	8.3
	5.7	4.7	10.5	9.1
	6.3	5.2	11.4	6.9
	6.8	6.3	8.6	7.7
	4.8	8.1	10.5	8.1
Mean	5.33	6.13	11.17	7.91
S.D. of means = 0.655	S.D. of data (σ) = 1.732			

short) occurs if the treatment selected as best has associated with it the largest mean. (Assume henceforth that "best" means "largest").

In order to do this, the experimenter must specify those values of the treatment means for which a correct selection is not important. To do this, it suffices to specify the *smallest significant distance* between the largest and the second largest treatment means. A distance between two treatment means is said to be significant if an error in ranking the two treatments is deemed to be important. In other words, there is some positive value δ^* such that if the largest and second largest treatment means differ by less than δ^* then selection of the "second-best" treatment instead of the "best" is not considered to be important; the experimenter is indifferent to a correct selection here and hence the name indifference zone.

The choice of δ^* may be subjective, or it may be dictated by other (e.g. economic) considerations, or it may be estimated by the data. Once it is chosen, however, it is possible to evaluate the performance of the selection procedure

"select the treatment giving rise to the largest sample mean"

by means of the probability of making a correct selection, or P(CS) for short. Of course this depends on the value of δ^*. The P(CS) is analogous to the power of a statistical test (note that there is no analogy to a significance level in a selection problem because the null hypothesis has already been rejected).

Using the corn example as an illustration, suppose that for economic reasons a value of $\delta^* = 1$ is considered appropriate. If the experiment were to be designed for a P(CS) equal to 0.9, Table A.1 of Gibbons *et al* yields the value of a constant $\tau = 2.4516$ for k = 4 treatments. The common number n of replicates per treatment required is then determined from this constant by the formula

$$n = \left(\frac{\sigma\tau}{\delta^*}\right)^2 = 19 \text{ (to the next largest integer)}$$

Therefore, if 19 replicates of each treatment are made, the sample means are computed, and the treatment giving rise to the largest sample mean is selected as best, then the probability of correct selection will be no less than 0.9 (provided of course that the difference between the largest and second largest treatment means is greater than 1).

Having already carried out the experiment based on seven replicates, however, it is possible to use the data obtained to estimate δ^* and hence the P(CS) of the procedure. Since Silver King gave rise to the largest sample mean, it is selected as the best variety. Using the method described in section 2.3.4 of Gibbons *et al*, the P(CS) of this procedure is estimated to be greater than .999! Therefore, although the number of replicates is small, the resulting data values indicate that the selection is correct with very high probability.

SUBSET SELECTION

In a subset selection problem the goal is to select a subset of the treatments that contains the best treatment (or treatments). Therefore a correct selection occurs if the selected subset contains the best treatment. Note that this does not imply that all the selected treatments are "good", or "better" than all the non-selected treatments.

Subset selection is useful when there is a very large number of treatments under consideration. As experiments or trials may be very expensive and time-consuming it is often desirable to reduce this number by selecting a subset containing the best, using the results of a preliminary experiment, and then performing more extensive trials with the smaller subset of treatments. The other desirable feature of this approach is that no indifference zone specification is required. However, the size of the selected subset is random, and if the number of replicates is very small then the subset may be very large and, at worst, may be the same as the original set of treatments!

The subset selection rule for normal data is very simple, viz:
"select all treatments giving rise to sample means that lie within τ standard errors of the largest sample mean"
where τ is found from the same table of Gibbons *et al* and depends on both the total number of treatments k and the desired P(CS). In the corn example the largest sample mean is 11.17 and the value of τ corresponding to $k=4$ and $P(CS)=0.9$ is 2.4516. Since $11.17-2.4516\times0.655=9.56$ is greater than all the other sample means, it follows that the selected subset contains only one treatment, Silver King, and this is obviously the best. (May you always be so lucky!). In fact, it turns out in this case that the same selection is made for a P(CS) as high as 0.999.

SELECTION OF TREATMENTS BETTER THAN A STANDARD OR CONTROL

This problem is very similar to the subset selection problem in that the goal is to select a subset containing all the "best" treatments. The difference is in how the treatments are defined to be best. In addition to the treatments of interest there is a control treatment whose mean may be known (a standard) or unknown (a control). The "best" treatments are all those whose treatment means are greater than that of the control.

This approach is very useful when it is desired to screen the treatments with respect to one (primary) trait, and then to select the best treatment with respect to some other (seconday) trait. For example, when comparing several herbicides for a particular crop, one may wish to screen out all those herbicides for which the crop yield is no less than a control value, and then select the best herbicide from these according to e.g. effectiveness. Again, it should be noted that the selected subset contains all treatments better than the control, but it may also contain others which are worse.

The selection rule for normal data is similar to the subset selection rule:
"select all treatments giving rise to sample means that lie above or within τ standard errors of the standard or control value"
where τ is the same as before if a control is used (but the number of treatments k is increased by one).

Under this procedure it is possible that no treatments are selected. In this case the conclusion is (with probability P(CS)) that none of the experimental treatments is better than the control.

To illustrate this approach, consider the data in Table 1 again and suppose that it is desired to select all varieties whose true mean yields are no less than 10 bushels/acre. For $k=4$ experimental treatments and a P(CS) of 0.95 the value of τ is 2.23 (using normal tables), and since $10.00-2.23\times0.655=8.54$ is greater than all but the sample mean from the Silver King variety, our selected subset again contains only this one variety, which we conclude (with probability 0.95) is better than the standard. In fact, if a P(CS) of 0.999 were desired, then the selected subset contains both Silver King and Osterland. The procedure for a control sample mean (based on 7 replicates) is similar but the number of treatments k increases to 5.

A COMPUTER PACKAGE FOR RANKING AND SELECTION

All the selection procedures described above may be carried out using an interactive computer package called RANKSEL, at present under development by the author. It is hoped to have a preliminary version of RANKSEL available on the Ministry of Agriculture and Fisheries' PRIME computer at Levin by mid-1982 for testing and evaluation. Anyone interested in obtaining this untested version should write to the author at the above address (RANKSEL is written in FORTRAN 77).

CONCLUSION

Ranking and selecting is an appropriate means of inference about the means of several treatments when it is believed that differences exist, either *a priori* or as the result of a test (e.g. ANOVA). Unlike Duncan's (1955) multiple range test, which groups means according to wheather they differ significantly or not, a ranking procedure ranks the treatments according to the treatment means from worst to best, whereas a selection procedure selects a subset of the treatments containing the "best" (where "best" is defined according to the ranking of the treatment means). The three procedures are designed to solve three different problems, and it is important that the appropriate procedure is used for a given problem.

Ranking and selection procedures are not new. The references for the three problems described are: indifference zone ranking and selection (Bechhofer 1954), subset selection (Gupta 1956, 1965) and selection of treatments better than a standard or control (Gupta and Sobel 1958). Furthermore, a text devoted to the methodology (as opposed to the theory) of statistical ranking and selection is now available (Gibbons *et al* 1977), and a computer package designed to carry out these procedures, described above, is almost completed.

Finally, it is the author's opinion that ranking and selection procedures should at least be considered as options whenever several treatments are to be compared on the basis of their means.

REFERENCES

Bechhofer, R.E., 1954. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics 25:* 16-39.

Duncan, D.B., 1955. Multiple range and multiple F tests. *Biometrics 11:* 1-42.

Gibbons, J.D., Olkin, I. and Sobel, M., 1977. *Selecting and ordering populations.* Wiley, New York.

Gupta, S.S., 1956. On a decision rule for a problem in ranking means Mimeo. Series No. 150, Institute of Statistics, University of North Carolina, Chapel Hill.

Gupta, S.S., 1965. On some multiple decision (selection and ranking) rules. *Technometrics 7:* 225-245.

Gupta, S.S. and Sobel, M., 1958. On selecting a subset that contains all populations better than a standard. *Annals of Mathematical Statistics 29:* 235-244.