ISSUES ARISING IN THE APPLICATION OF BONFRERRONI PROCEDURES IN FEDERAL SURVEYS

Susan W. Ahmed
National Center for Education Statistics, Washington, D.C. 20208

1. Introduction

Multiple comparison procedures have been widely discussed, debated, and evaluated, both in the statistical literature and in application journals. Rupert Miller's 1966 book Simultaneous Statistical Inference (1) and his 1977 JASA article (2) summarize the various procedures and comment on their applicability. Hochberg and Tamhane's 1987 book (3) also provides a detailed discussion of various multiple comparison procedures. The Current Index of Linear Models - 1975-1988 (4) has 233 listings under the heading "Multiple Comparison Procedures" and 48 listings under "Bonferroni Inequality". So, what more can there be to say on the topic?

This paper does not present any new multiple comparison techniques, nor does it attempt to provide any new comparisons of existing procedures. Rather, it focuses on issues arising in attempting to apply multiple comparison procedures in the setting of a Federal agency carrying out large scale sample surveys, a setting which makes the problem even more difficult than it is otherwise.

There is a considerable amount of debate at the National Center for Education Statistics (NCES) on the question of whether we should be using multiple comparison procedures and questions arise frequently as to how to apply multiple comparison procedures. This paper arose out of a desire to write guidelines for analysts at the Center. NCHS has tried to write guidelines on this topic which are included in their **Manual for Reviewing Statistical Reports (5).** However, even this manual does not cover many of the situations which we come up against. In the end, I am not sure that I will be any more successful in writing guidelines, but I will try to lay out some of the issues.

2. Background

Multiple comparison procedures are procedures which are used to control the overall level of a Type I error when making simultaneous inferences, i.e. when testing several hypotheses at one time to determine which of several groups are different from each other or which of several contrasts are different from zero. There are really two purposes behind multiple comparison procedures. The first is to adjust for making simultaneous statements; the usual significance level applies only to a single test. If we perform more than one test, the usual alpha level is not applicable to the set of conclusions. Secondly, it applies to preplanned comparisons. If we go searching through the data, data snooping, the original alpha level is no longer in operation.

The following brief derivation of the Bonferroni inequality shows what happens to the alpha level as one performs more than one statistical test.

One test: $\alpha = 0.05$
$P(I_1) = 0.05$

Two tests: P(any Type I error) =
$P(I_1 \text{ or } I_2) =$
$P(I_1) + P(I_2) - P(I_1 \cap I_2)$
$\leq 2\alpha$

Three tests: P(any Type I error) = $P(I_1 \text{ or } I_2 \text{ or } I_3)$
$= P(I_1) + P(I_2) + P(I_3) - P(I_1 \cap I_2) - P(I_1 \cap I_3) -$
$P(I_2 \cap I_3) + P(I_1 \cap I_2 \cap I_3) \leq 3\alpha$

.
.
.

k tests: P(any Type I error) $\leq k\alpha$.

If we perform a single test using a 5% significance level, then our chance of concluding that there is a difference when there is not is 5%. If we perform two tests, each at a 5 % significance level, then our overall chance of making at least one type I error is usually greater than 5% and can be as high as 10% (10% is an upper limit). With k tests, the overall probability of a Type I error is less than or equal to $k\alpha$. This is an **upper** bound; all we know is that the actual probability of a Type I error is somewhere between α and $k\alpha$. If we use a significance level of α/k on each individual test, the overall α level for the **family** of k tests will be controlled at α. This is the Bonferroni multiple comparison procedure, i.e. if we wish to control the overall level of a Type I error for a family of k comparisons at α, use α/k for each individual comparison. Since $k\alpha$ is an upper bound for the overall alpha level, this approach is usually conservative and we are usually not sure how conservative. (If the tests are independent, we can derive the exact probability of a Type I error.) The application of the Bonferroni procedure sounds easy enough. However, when one tries to apply it in a real setting, it becomes much more difficult.

Statistical standards at NCES require that whenever an author comments on a difference in a statistical report, it must be supported by a statistical test. In addition, if the author makes simultaneous statements or multiple comparisons, he must use a procedure which adjusts the alpha level for the multiple comparisons. The procedure most commonly used at the Center is the Bonferroni procedure, somewhat conservative in many situations, but a procedure that is generally applicable and fairly easy to use. The remainder of this paper focuses on several issues which make it difficult to apply these procedures in the setting of a Federal agency.

The first two issues not only have an impact on the application of multiple comparison procedures, but also set the background for our discussion.

3. Issues

Issue 1: Nature of NCES Reports.
Issue 2: Nature of NCES Audiences.

NCES produces a number of types of reports including analytic reports (where this issue is less of a problem), simple collections of tables with statements of major findings (EdTabs), general survey reports which describe some of the basic overall findings on key issues from a survey and which are generally based on univariate analyses (t tests, chi-square), and compendia of results from several surveys across various levels of education.

In addition to the diversity of **types** of reports, NCES writes reports for a number of diverse audiences. If we were speaking simply to researchers, we could describe what we have done, i.e., how we have handled multiple comparisons, how many comparisons we have adjusted for, and let the reader judge whether he agrees with our approach or not. When we speak to someone in Congress or to a teacher or to the general public (people who will not necessarily read the technical details of the methodology, but will take the statements made as truth), however, it becomes more difficult and we need to take more care in the procedures we use, in having them consistent, and in insuring that we are in some agreement that they make some sense.

Issue 3: Policy Implications of Statements that Come from a Federal Agency.

The position of a researcher in a university setting who publishes results from a study is somewhat different from that of an analyst in a Federal agency writing a report on a survey and drawing conclusions such as "Black students are not performing as well as whites on national achievement tests" or "State X performed more poorly than State Y in the recent state survey from the National Assessment of Educational Progress". The policy implications add to the burden of wanting to control the overall α level and not letting it creep up as it will without any adjustment. The following statement from Rupert Miller (1), although used by Miller in distinguishing between the use of multiple comparison procedures in exploratory versus definitive studies, is relevant to the discussion of policy implications of statements from a Federal agency. He says "The statistician does not have to be as conservative for the first type (exploratory) as the second (definitive). But simultaneous techniques are still quite useful for the first in keeping the number of leads that must be traced within reasonable bounds. In the latter type (definitive studies), multiple comparison techniques are very helpful in avoiding public pronouncements of red herrings simply because the investigation was quite large." This latter statement warning against findings which might be statistically significant only because of a large sample size, supports the use of multiple comparison procedures to minimize the possibility of a Federal agency making such public pronouncements.

Issue 4: Desire for Uniform Procedures Throughout the Agency

NCES has a set of statistical standards which is designed to define good statistical practice in the Center and to provide Center staff with guidelines for survey design, monitoring, analysis, and reporting. Center standards require that multiple comparison procedures be used when making simultaneous inferences.

An individual researcher in a university setting or elsewhere, in deciding on how to apply multiple comparison procedures, can choose for himself how he will handle the issue, with his only constraint being the editor of the journal he will send his article to and the reviewers he happens to get. In the case of a federal agency, it is desirable to have uniform procedures throughout the agency. We would like to have the conclusion remain the same whether author A or author B were to write the report. We would like to have uniformity across authors within a Division of the Center and across Divisions of the Center. The most difficult problem in achieving that uniformity with respect to the application of multiple comparison procedures, as anyone who has dealt with the issue knows, is the difficulty in deciding on the family size.

Issue 5: Difficulty in Deciding on Family Size.

Miller (1) discusses the possible range of behavior with respect to multiple comparison procedures, which I have illustrated pictorially in Figure 1. At one extreme, we have the liberal statistician who doesn't believe in multiple comparison procedures at all. He will perform every statistical test he ever does at the same alpha level, e.g. $\alpha=0.05$, and does not worry about the fact that the overall alpha level may accumulate as he performs more and more tests. There are people in the Center who hold this view; they do not wish to ever use multiple comparison procedures. At the other extreme is the ultra conservative who would say that he wishes to have enough control so that for all the comparisons he makes in his entire lifetime, the overall alpha level will be controlled at, say, 0.05. Our standards do not advocate this position. A little to the left is the statistician who would have the entire report

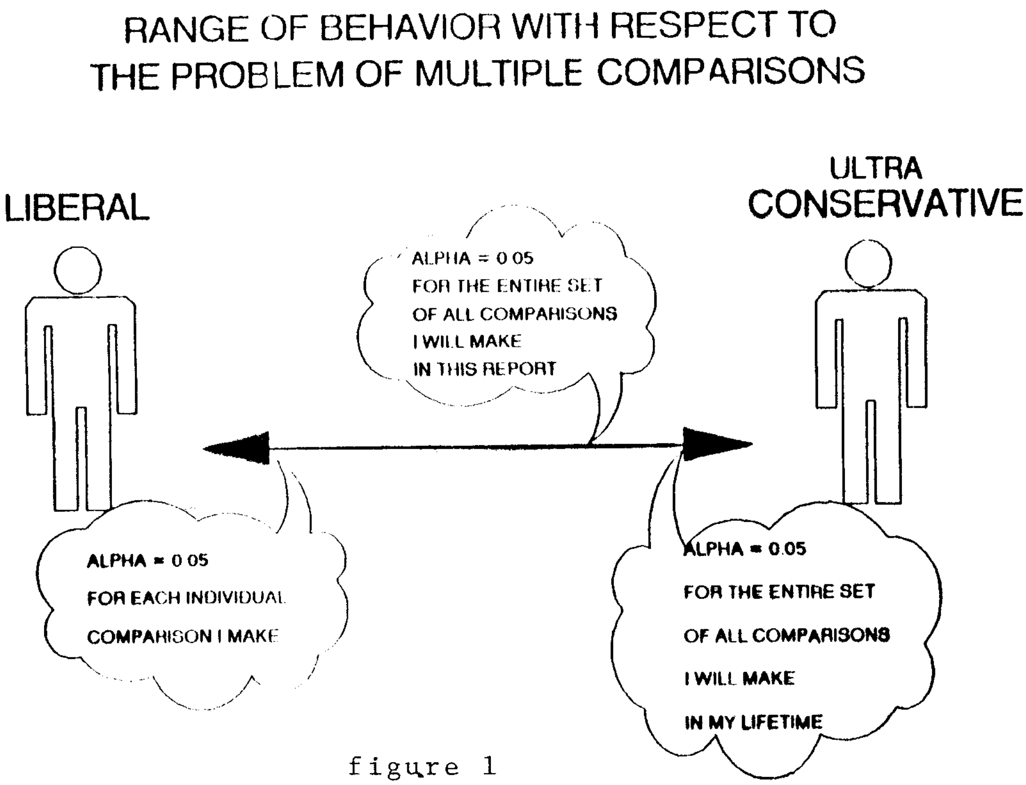

RANGE OF BEHAVIOR WITH RESPECT TO THE PROBLEM OF MULTIPLE COMPARISONS
LIBERAL
ULTRA CONSERVATIVE
ALPHA = 0.05 FOR THE ENTIRE SET OF ALL COMPARISONS I WILL MAKE IN THIS REPORT
ALPHA = 0.05 FOR EACH INDIVIDUAL COMPARISON I MAKE
ALPHA = 0.05 FOR THE ENTIRE SET OF ALL COMPARISONS I WILL MAKE IN MY LIFETIME

figure 1

be controlled at an overall alpha level of 0.05. We do not even advocate this position. So, in some respects, we are really quite liberal in our application of multiple comparison procedures.

In some cases, it is quite easy to see the reason for wanting to control the overall alpha level and easy to determine the number of comparisons to be controlled for, i.e., the family size. Suppose we have five racial/ethnic groups (Asians, Whites, Blacks, Hispanics, and American Indians), and we want to see if Asians are outperforming the other racial/ethnic groups, as has been reported in the literature. Our null hypothesis is H_0: Asian/Pacific Islander eighth graders have the highest math achievement scores; the alternative is H_A: Asian/Pacific Islander eighth graders do not have the highest math achievement scores. There are four groups to be compared to Asian/Pacific Islanders, four t-tests to be performed, and thus the number of comparisons is four. If we make a wrong decision on any one of these tests by concluding that Asians outperform the paricular group being compared when they really don't, then the whole conclusion falls apart. In this case, deciding on the number of comparisons is quite straightforward.

A slightly more complicated example is illustrated in Table 1 from an NCES survey report on eighth graders. This particular table examines the type of science course taken by eighth graders (science course with a lab, science course without a lab, or no science) according to various student characteristics (sex, race/ethnicity, science test performance, amount of science homework, school type). The author wishes to see if the percentage of students taking a science course with a lab differs across levels of each of these characteristics. Our approach to this type of comparison has been to consider each factor by itself when determining family size and thus the family size is determined by the number of categories (c) in the factor. The family size is the number of combinations of c things taken two at a time. Thus, for race, there are 5 categories, 10 combinations of 2, and the comparisonwise alpha level α_c will be $\alpha/10$.

Table 2 illustrates a more complex example. The table examines the gender and race distribution of teachers according to various school characteristics (school level (elementary, secondary, combined), minority enrollment, and school size). Suppose we wish to know if there are more Hispanic teachers in elementary schools, secondary schools, or combined schools. Taking a combination of 3 categories taken 2 at a time, we have 3 comparisons.

Now consider the fact that we wish to make this comparison for public schools and private schools; and within each of these groupings we wish to look at urban, suburban, and rural schools. Do we multiply by 2 and again by 3 to give us a family size of 18.

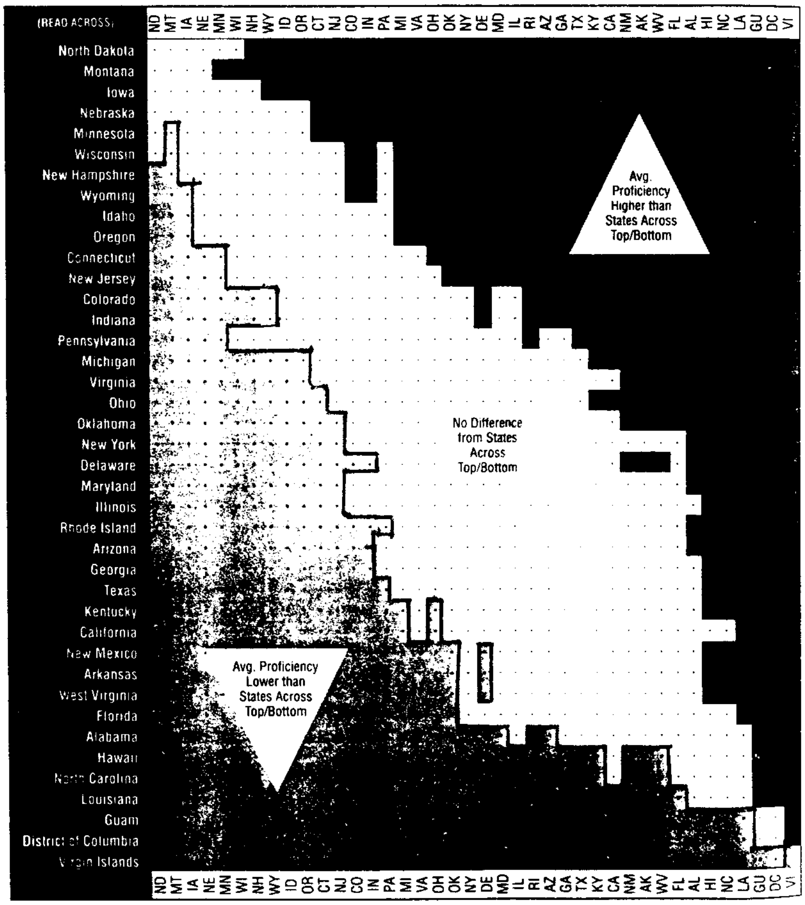
These are only a few of many examples meant to illustrate the complexity of deciding on family size.

Issue 6: Multiple Uses of the Same Results.

This issue is illustrated by the results of our recent study of state by state comparisons of performance of eighth graders in mathematics. Forty states and territories participated in the study which was designed to allow states to compare themselves to others with respect to mathematics achievement. Figure 2 illustrates the results. This display was developed by Gene Johnson from ETS and John Tukey who has been serving as a consultant to NAEP (the National Assessment of Educational Progress). The states are listed down the left side and across the top from highest scoring to lowest scoring. To read the chart, locate a state of interest in the left column. Consider, for example, New York. States in the middle white area, whose names are read at the top, are not significantly different from New York. States in the gray area at the left scored significantly higher than New York and states in the black area at the right scored significantly lower. The determination of significance in this chart was determined using t-tests, Bonferroni adjusted, based on 780 comparisons (all combinations of 40 states taken two at a time).

Suppose, however, that I am in the state of Michigan, and I wish to compare my state to the other 39 states and territories; I am not interested in comparisons among other states, only those involving Michigan. There are 39 such comparisons. This would produce a different chart, possibly with less white area and more gray and black area, i.e. more significant differences. Do we publish both charts. Will the ordinary reader understand the differences between the charts or will the reader be confused?

FIGURE 2
Comparisons of Overall Mathematics Proficiency
Based on Appropriate Tests of Statistical Significance



Note: Reading across, from left to right, this chart shows whether the average proficiency of each state or territory is lower than, the same as, or higher than that of other participants.

*Significance determined by an application of the Bonferroni procedure based on 780 comparisons by comparing the difference between the two means with four times the square root of the sum of the squared standard errors.

For any given state:

■ Overall average proficiency statistically significantly higher than comparison state.

□ No statistically significant difference from comparison state

■ Overall average proficiency statistically significantly lower than comparison state.

Issue 7: Fine Line Between Data Snooping and Preplanned Comparisons.

Whether or not a set of comparisons is preplanned or the result of data snooping seems to be straightforward. However, as illustrated by our previous example on the distribution of teachers by age and race, we often do a certain amount of preplanning and then proceed to data snooping to decide what it is we will actually talk about in our report.

Issue 8: Difficulty in Implementing These Procedures.

Multiple comparison procedures are not built into sample survey software, as they are into many ordinary statistical analysis packages such as SAS. This makes the implementation of these procedures more tedious. We need guidelines for our Center staff and our contractors so that we are all carrying out the procedures in the same way.

There is much resistance by the analysts. Their argument is that these procedures are too conservative and are less powerful. This is true, and again we do not know how conservative. For the analyst, this means that he will not find as many significant differences using multiple comparison procedures as he would without them. In addition, these procedures are more difficult and more time consuming. The analyst has to go through this process of deciding how many comparisons are involved; he must think more carefully about what it is he wants to say and this takes a lot of time when a number of different families of comparisons are being made in a survey report.

Issue 9: Displaying the Results of Multiple Comparison Procedures.

It is often difficult to display the results of multiple comparison procedures, both in terms of confidence intervals and in terms of hypothesis tests. Figure 1 is one example of how one might display the results of multiple tests in a fairly understandable way.

Issue 10: Substitution for More Complex Analyses.

One of the reasons the issue of multiple comparisons arises so frequently in our survey reports is that it is often used when we should really be doing more complex analyses - analysis of variance and loglinear models. Instead, we are making many pairwise comparisons using t-tests. We need to raise the level of analysis that is done on our survey data. It has only been very recently that software has become available to perform these types of analyses on survey data and it will take time for analysts to become familiar with this software and begin to use it. In addition, the whole area of analysis of data from complex surveys and how one takes account of the sample design is a new and developing area.

Issue 11: Robustness of the Bonferroni Procedure to Violations of the Assumption of Normality.*

There is some evidence (6) that if the assumption of normality is violated, the tail probabilities of the test statistic (z or t) will be poorly estimated. Thus, if the distribution of the variable being analyzed is nonnormal, the Bonferroni adjustment may not provide an accurate approximation of the overall alpha level. More work needs to be done to examine the nature of the distribution of our variables and the effect on the Bonferroni procedure.

4. Conclusion

So, what guidelines do I have for our analysts who use multiple comparisons? Not many. If an author needs several conclusions to hang together to draw his conclusion, this set of comparisons can be considered to be a family. Beyond this, we tend at NCES to use a variablewise approach rather than a global approach for a whole table or a whole report. This is really the only guideline we have been able to come up with so far.

Where do we go from here? We can continue to try to come up with guidelines to help the analyst use multiple comparison procedures; we can discontinue the use of multiple comparison procedures - some of our analysts would be happy with this approach; we could change the alpha level at which we test hypotheses, possibly using 0.10 rather than 0.05. This would help the power issue a little. Many Federal agencies have already switched to an alpha level of 10%. The Center is considering this change and will be bringing in some outside experts to provide opinions on these issues. We are also looking into some of the recent proposed modifications to the Bonferroni procedure designed to achieve higher power (7,8).

Two other options seem to have less validity. When doing many tests, some authors switch to an alpha level of 0.01. This approach is not very appealing; at times we will be overcorrecting and at other times undercorrecting. Others have recommended the use of orthogonal comparisons. Many analysts will answer the statistician who makes this suggestion by informing him that his research questions do not come in orthogonal sets. This is often not a viable alternative, although it should be given some consideration. Even if the set of comparisons is orthogonal, it is not clear that an adjustment to the alpha level need not be made.

*Issue 11 was brought up at the presentation of this paper at the ASA meetings, by Fritz Scheuren and Juliet Shaffer.

The issue of multiple comparisons is a difficult one, even more difficult in a survey setting than in more ordinary applications. Guidelines are difficult to write. What this paper has tried to do is to lay out some of the issues which arise when one tries to use multiple comparison procedures in the setting of a federal agency.

REFERENCES

(1) Miller, Rupert. Simultaneous Statistical Inference, 1966. McGraw-Hill.

(2) Miller, Rupert. Developments in Multiple Comparisons 1966-1976. JASA 72 469-478.

(3) Hochberg,Y. and Tamhane, A. Multiple Comparison Procedures. Wiley, 1987.

(4) Current Index of Linear Models: 1975-1988. American Statistical Association.

(5) Sirken, M., Shimizu, I., French, D., and Brock, D. Manual on Standards and Procedures for Reviewing Statistical Reports. NCHS, 1990.

(6) Shaffer, Juliet, University of California at Berkeley. Personal Communication.

(7) Shaffer, Juliet, "Modified Sequentially Rejective Multiple Test Procedures". JASA, 81, 826-831.

(8) Simes, R.J. "An Improved Bonferroni Procedure for Multiple Tests of Significance", Biometrika, 73, 751-54.

TABLE 1

DISTRIBUTION OF TEACHERS BY
BY SEX AND BY RACE/ETHN, AND AVG AGE

	Sex		Race/Ethn					Avg
	M	F	W	B	H	A	I	Age
ALL SCHOOLS								
PUBLIC								
URBAN								
School Level								
Elem					X			
Sec					Y			
Combined								
Minority Enrollment								
Lt 20%								
GE 20%								
School Size								
Lt 150								
150 to 499								
500-749								
750+								
SUBURBAN								
.								
.								
RURAL								
.								
.								
PRIVATE								
.								
.								

TABLE 2

Table 2.5.--Percentage of eighth graders who report enrolling in science course with laboratory, science course without laboratory, or no science course, by selected background characteristics

Background Characteristics	Course Enrollment		
	Science Course with Laboratory	Science Course without Laboratory	No Science
TOTAL	21.5	74.2	4.4
SEX			
Male	22.1	73.4	4.5
Female	20.9	74.9	4.3
RACE/ETHNICITY			
Asian and Pacific Islander	25.1	65.7	9.3
Hispanic	19.2	72.5	8.3
Black	19.5	74.4	6.0
White	21.9	74.8	3.2
American Indian and Native Alaskan	21.2	73.4	5.3
SCIENCE TEST QUARTILE			
Lowest Quartile	19.1	74.0	6.9
25-49%	18.9	76.8	4.4
50-74	21.1	75.7	3.2
Highest Quartile	25.6	72.4	2.0
SCIENCE HOMEWORK			
None	19.3	68.7	12.0
Less than 1 Hour	20.5	76.7	2.7
1 Hour	22.5	75.1	2.5
2 Hours	24.2	73.6	2.1
3 Hours	26.4	71.4	2.2
4-6 Hours	29.5	68.1	2.4
7-9 Hours	28.3	69.7	2.0
10 Hours or more	35.3	64.4	0.3
SCHOOL TYPE			
Public	21.5	73.9	4.6
Catholic	18.6	79.1	2.3
Independent	48.0	48.1	3.9
Other Private	21.5	76.1	2.4

SOURCE: U.S. Department of Education, National Center for Education Statistics, "National Education Longitudinal Study of 1988: Base Year Student Survey."

This paper is intended to promote the exchange of ideas among researchers and policy makers. The views are those of the author, and no official support by the U.S. Department of Education is intended or should be inferred.