Published February 19, 2014

Symposium: Statistical Concepts

Multiple Comparison Procedures—Cutting the Gordian Knot

David J. Saville*

ABSTRACT

Multiple comparison procedures (MCPs), or mean separation tests, have been the subject of great controversy since the 1950s. Essentially, these procedures are an attempt at simultaneously formulating and testing pairwise comparison hypotheses using data from a single experiment. An unacceptable operating characteristic of most MCPs is their "inconsistency," an idea that is illustrated in this article. This characteristic led to the development of a "practical solution" to the MCP problem, which is to "cut the Gordian knot" by abandoning any attempt at simultaneous formulation and testing. Instead, I recommend using the simplest multiple comparison procedure, the unrestricted least significant difference procedure, to (i) formulate new hypotheses at a known "false discovery rate" (in the null case) such as 5%, and (ii) independently test interesting new hypotheses in a second experiment. I also discuss the implications for sample size calculations of the choice of MCP.

Multiple comparison procedures are commonly used to test for "significant" differences between treatment means in experiments, even in cases where the set of treatments has clear structure and has been derived with obvious questions in mind. In these cases, the use of an MCP is inappropriate, as has been pointed out a multitude of times (e.g., Swallow, 1984; Little, 1978). To quote Swallow (1984), MCPs "were developed for cases where the treatment set lacked structure, that is, where the treatments were just a collection of varieties or perhaps chemicals with no particular inter-relationships. Most treatment designs are not of this type. Usually, the treatment set has a structure, and the statistical analysis should recognize that structure." This can be achieved by specifying appropriate contrasts between the treatments, with each contrast addressing a particular question of interest to the researcher; in many cases, these contrasts can be chosen to be "orthogonal" (mutually independent) but this is not essential. Journals often encourage researchers to thus tailor their statistical analysis to the objectives of the research, but the specification of appropriate contrasts is not a skill easily acquired by researchers and help from a biometrician is not always available, so this encouragement is often to no avail. As an example of such advice, in the instructions to authors of Agronomy Journal (September, 2012), the statistical methods section is largely devoted to warning of the limitations of MCPs, with the closing statement, "[w]hen treatments have a logical structure, orthogonal contrasts among treatments should be used."

Saville Statistical Consulting Limited, PO Box 69192, Lincoln 7640, New Zealand. Received 11 Oct. 2012. Accepted 18 Dec. 2012. *Corresponding author (savillestat@gmail.com).

Published in Agron. J. 107:730–735 (2015)
doi:10.2134/agronj2012.0394
Available freely online through the author-supported open access option.

Thus, the long-running debate on the relative merits of the many different MCPs is relevant only to the minority of studies in which such a procedure is appropriate.

I first introduce some necessary statistical terminology, then discuss the general topic of MCPs in relation to various types of error rate and in relation to the levels of conservatism of some of the better known MCPs. The idea of inconsistency is then introduced and discussed, with particular attention being paid to Fisher's restricted least significant difference (LSD) procedure, the MCP most commonly used in Agronomy Journal. Next, the implication of the choice of MCP for required sample size (estimated using a power analysis) is discussed. Finally, a practical solution to the problem of best choice of MCP is described. This consists of using the simplest of procedures, the unrestricted LSD procedure, with the proviso that it be regarded as an hypothesis formulation tool, with any interesting pairwise hypotheses thus formulated requiring testing in a second, independent experiment.

RESULTS AND DISCUSSION

Types of Statistical Error

In drawing statistical conclusions from an experiment, the hope is that all of your decisions will be correct. For example, if there is truly no difference (between two treatment means, for example), then the correct decision is to decide that there is no difference (Table 1). Similarly, if there is truly a difference, then the correct decision is to decide that there *is* a difference (Table 1).

In real life, not all decisions will be correct, and statisticians refer to two types of error that you can make. Loosely speaking, one error is to find things that are not there, and the second is to not find things that *are* there. These errors are sometimes referred to as false positives and false negatives, respectively.

Formally, the first type of error (Type I) is to erroneously declare a null effect to be real or non-zero (Table 1). If your statistical

Abbreviations: HSD, honest significant difference; LSD, least significant difference; MCP, multiple comparison procedure.

Table 1. The four cases that can occur in relation to a statistically based decision concerning whether a particular difference is zero or not. In two cases the correct decision is made and in two cases an error occurs.

		THE TRUTH	
		No difference	Difference
DECISION	No difference	Correct decision	Type II error
	Difference	Type I error†	Correct decision‡

† Probability of Type I error = 0.05 (for 5% level significance test).

‡ Probability of correct decision (above) = power of test.

test has a 5% level of significance, this means that it has been constructed so that the probability of making a Type I error is 0.05 (or 5%). With a 1% level test, the Type I error rate is 0.01 (or 1%).

Conversely, a Type II error is to fail to find a real (non-zero) effect (Table 1). When designing your experiment, you may carry out a "power analysis" to determine the sample size required to reduce the Type II error rate to a specified low level. The "power" of the test is the probability of deciding that there is a non-zero difference when the difference truly is non-zero; that is, the power is the inverse of the Type II error rate (so the two add up to 1.00). (More details on power analysis are given below.)

In an ideal experiment, you are able to minimize the chance of both types of error. In real life, resources are limiting, so if you specify a low chance of one type of error, you automatically increase the chance of the other type of error; for example, if you decide to reduce the chance of a Type I error by insisting on a 1% significant difference rather than a 5% significant difference (before you report the difference as non-zero), then you increase the chance of a Type II error (not finding a truly non-zero difference). The only way to reduce both chances is to do more work or adopt a more sophisticated experimental design, perhaps involving a fancy block layout to reduce the residual error term.

When studying error rates associated with different MCPs, statisticians have found it relatively easy to study experiments in which there are truly *no* effects but have found it less convenient to study experiments in which there are real effects (which can occur in an infinite variety of ways). As a result, statisticians in general have spent more time studying Type I errors than Type II errors. Researchers, on the other hand, are often more worried about not detecting a real effect, so are more concerned about Type II errors than Type I errors. There has therefore historically been a difference in mind set—statisticians have tended to be more preoccupied with Type I errors, while researchers have tended to be more preoccupied with Type II errors.

Traditional Hypothesis Testing Scenario

The traditional hypothesis testing scenario goes something like the following:

A group of researchers has an idea (or hypothesis), such as "legumes outyield nonlegumes."

The researchers plan an experiment to test this idea (hypothesis); for example, they may decide to include six treatments, consisting of three legumes and three nonlegumes.

Their biometrician advises that the "contrast" (comparison) corresponding to the idea (hypothesis) of interest is the average of the treatment mean yields for legumes minus that for nonlegumes. This is a *pre-specified contrast*. For information on how to match ideas (hypotheses) to contrasts or comparisons, see Cochran and Cox (1957) or Saville and Wood (1991).

After the data have been collected, an analysis of variance (ANOVA) is performed that includes this contrast. An F test of the hypothesis "true contrast value = 0" is performed as part of the ANOVA; this is either significant or not.

Conclusion: The idea (hypothesis) that "legumes outyield nonlegumes" is either confirmed, or not, by the experiment.

What if There Are No Prior Ideas?

If there are no prior ideas (hypotheses), there are two cases that occur. The first, rather trivial case is that the experiment involves only two treatments (here there *is* presumably a prior plan, which is to compare the two treatments). In this case, there is good news!—statisticians agree on how to statistically compare the two treatment means (use an ANOVA, which, by default, includes an F test of the hypothesis "true difference between the two treatment means = 0").

The second case is that the experiment involves more than two treatments, but there are no prespecified contrasts (or ideas/hypotheses) (here the researcher presumably *did* have reasons for doing the experiment, but the underlying ideas have not been articulated). In this case, there is bad news—statisticians do not agree on how to statistically compare the treatment means (two at a time) and have suggested scores of different MCPs, which have been the subject of great controversy since the 1950s.

Multiple Comparison Procedures

Essentially, MCPs are an attempt at simultaneously formulating and testing pairwise comparison hypotheses using data from a single experiment. Statisticians view this as similar to data-dredging. They think, "what if all treatments are truly equal?" and worry about the number of false 5% level significances, or Type I errors, that can occur in this case (e.g., with 20 treatments, there are ${}_{20}C_2 = 20 \times 19/2 = 190$ comparisons, so with a Type I error rate of 0.05, there would be, on average, $190 \times 0.05 = 9.5$ Type I errors).

Such thinking about the probability of committing Type I errors in the null case generates a desire to build conservatism into an MCP procedure. The problem is: how much conservatism should be built into an MCP? Unfortunately, no one agrees on the answer. To date, scores of MCPs have been proposed, and new MCPs are currently being proposed, all with differing amounts of conservatism.

Ordering of Multiple Comparison Procedures by Level of Conservatism

Of those MCPs that are commonly used by research workers, the three most conservative MCPs, in terms of the ability of the researcher to declare significant differences between means, are Bonferroni, Tukey's honest significant difference (HSD), and Student–Newman–Keuls. These are all based on thinking about the *experiment-wise* Type I error rate, which is the probability that at least one Type I error occurs in an experiment that truly includes no treatment effects.

On the other hand, the least conservative MCPs are Duncan's multiple-range test and the unrestricted LSD procedure. These are based on thinking about the *comparison-wise* Type I error rate, which is the probability of Type I error per null comparison.

Fisher's restricted LSD procedure is somewhere in between the above extremes in terms of level of conservatism. If the overall F

test is statistically significant, then it reduces to the unrestricted LSD procedure. Overall, it is variable in its level of conservatism.

What is the Natural Unit?

A related question is: "What is the natural unit for the statistical analysis?"

If the answer is the *comparison* (between any two treatment means), then the MCP to use is the unrestricted LSD procedure, which will falsely declare 5% of null differences to be significant. In this case, the comparison-wise Type I error rate is 5%.

If the answer is the *experiment,* then the MCP to use is the more conservative Tukey's HSD procedure, which will falsely declare a null difference to be significant in only 5% of null experiments. In this case, the experiment-wise Type I error rate is 5%. This MCP has a much reduced comparison-wise Type I error rate but pays the price of a much increased Type II error rate.

But why stop there? If the answer is the *project*, consisting of several experiments, then the MCP to use is an even more conservative, yet-to-be-invented procedure that will falsely declare a null difference to be significant in only 5% of null projects. This procedure would have an even lower comparison-wise Type I error rate but would pay an even higher price in terms of an increased Type II error rate.

The logical next step in this sequence of possible natural units is the *research program,* consisting of several projects. Here the MCP to use would be even more conservative!

Or, if your statistician is particularly keen on not making errors of the Type I variety, he or she may want the natural unit to be the *lifetime of your statistician*. This would have particularly dire consequences in terms of your hope of achieving a statistically significant effect!

In conclusion, I would argue that for the researcher, the *individual comparison* is the natural unit. Once you depart from it, there is no natural stopping point (experiment, project, research program, lifetime, ...?). Consequently, in ad hoc situations, the unrestricted LSD should be used with the full understanding that the false discovery rate is whatever the researcher chooses his or her Type I error rate to be.

Table 2. Mean oat grain yields from four fictitious experiments, one per bear, with data analysis carried out by Goldilocks using the Fisher's restricted LSD procedure. The pooled standard error of the mean (SEM), overall ***F*** value, and its significance are also presented, and the last row gives the "significance" of the difference between MiteyOat and TrustyOat for each experiment (reproduced from Saville and Rowarth, 2008).

	Mean oat grain yield			
Oat cultivar	Baby Bear	Mama Bear	Papa Bear	Grandpa Bear
	kg/ha			
WonderOat	5030	5030	5030	5030
Various no. 1	5160	5260	5460	5760
MegaOat	4910	4910	4910	4910
MiteyOat	4450	4450	4450	4450
TrustyOat	5550	5550	5550	5550
FlakeyOat	4870	4870	4870	4870
Various no. 2	5120	5120	4520	4320
TrendyOat	4910	4910	4910	4910
Statistics				
SEM	200	200	200	200
Overall *F* value	2.43	2.57	3.82	5.98
Significance of overall *F*	ns	*	**	***
LSD(5%)	ns	588	588	588
LSD(1%)	ns	ns	801	801
LSD(0.1%)	ns	ns	ns	1080
Difference, MiteyOat – TrustyOat	ns	*	**	***

* Significant at 0.05 probability level; ns, not significant.

** Significant at the 0.01 probability level.

*** Significant at the 0.001 probability level.

Inconsistency of Multiple Comparison Procedures

In general, the more conservative an MCP, the more inconsistent it is. The term *inconsistency* is now defined, and its undesirability is explained.

By definition, an MCP is called *inconsistent* if for any two treatment means, the probability of judging them to be "significantly" different depends on either the number of treatments included in the analysis or the values of the other treatment means (Saville, 1990).

Goldilocks and the Four Bears

To illustrate this idea, I will use an example from Saville and Rowarth (2008), including their Table 1.

In this fictitious example, I borrowed the terminology of Carmer and Walker (1982) and Saville (1985) and considered the case of a statistician, Goldilocks, who has analyzed data for four clients, Baby Bear, Mama Bear, Papa Bear, and Grandpa Bear (the latter has recently come to live with the family). The Bears are all keen porridge eaters and each had performed an experiment with eight oat (*Avena sativa* L.) cultivars in an attempt to increase their oat production. All four experiments were laid out in a randomized complete block design with four replications. By coincidence, each experiment included six common cultivars, and these cultivars happened to have identical oat yield data in all four experiments (Table 2). The other two cultivars differed between the experiments and their names are not specified; their oat yield data differed in their means but not in their standard deviations between experiments. Goldilocks' statistical analysis using Fisher's restricted LSD procedure for each of the four experiments is summarized in Table 2 in terms of mean oat yield for each cultivar, the pooled standard error of the mean (SEM) (which turned out to be 200 in all four experiments), the overall *F* value and its significance, and the LSD for 5%, 1%, and 0.1% level tests. To illustrate the notion of inconsistency, the significance of the difference between two of the cultivars, MiteyOat and TrustyOat, as determined by Goldilocks using the Fisher's restricted LSD procedure, is also included.

Astonishingly, the significance of the difference between MiteyOat and TrustyOat varied widely among the four experiments (from not significant to 0.1% significant), in spite of the fact that their mean yields were identical, the pooled SEMs were identical, and the residual degrees of freedom (21) were identical among experiments. The reason for this variation can be traced back to the decision on which two "various" cultivars were included. When the two "various" cultivars had yields that were similar to the experimental average, a low overall *F* value was calculated (Table 2). When the two "various" cultivars had extreme yields (one low, one high), a high overall *F* value was calculated (Table 2). That is, the mean yields for the two

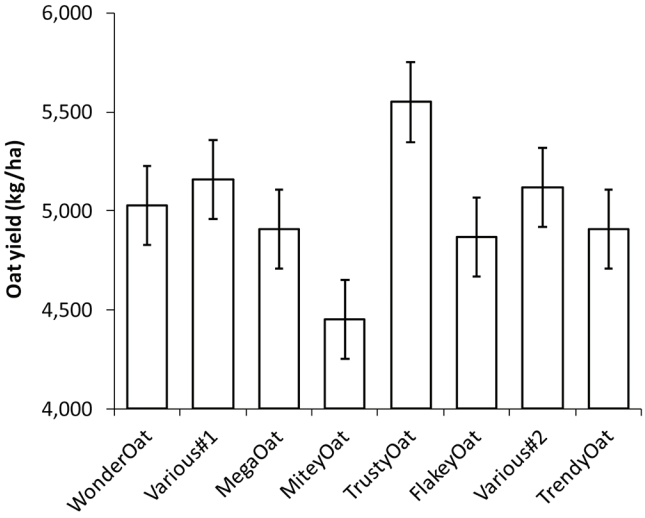

Fig. 1. Cultivar mean oat yields in Baby Bear's experiment. Vertical bars are SE bars (using pooled standard error of the mean of 200). Note that the mean difference between MiteyOat and TrustyOat is 1100 kg/ha, which is 5.5 times the SE, yet Fisher's restricted LSD procedure declares this difference to be not statistically significant.

"various" cultivars determined the statistical significance of the overall *F* value and hence the significance of the difference between MiteyOat and TrustyOat. Such statistical subtleties aside, however, one can imagine what Grandpa Bear would say about the way in which Goldilocks had handed out significant differences: "Why would Goldilocks be so kind to an old bear but so unfair to a young fledgling-experimenter bear?"

This sort of inconsistency in the results is something that no practicing biometrician would ever want to have to defend (like: "would you like to share your office block with four angry bears?"). In all four experiments, the *t* value for the comparison of MiteyOat with TrustyOat is given by $t = 1100/(200 \times \sqrt{2}) = 3.889$, which is 0.1% significant (given 21 df). It defies common sense to override this simple test in Baby Bear's experiment by arguing that the other cultivar means are not sufficiently spread out (hence a low *F*), so this comparison should be declared "not significant." This nonsignificant result also contradicts what a journal reader would decide after inspecting a bar graph of the results such as that shown in Fig. 1; the usual rule of thumb is that means that differ by more than about 3 SEs are significantly different ($P < 0.05$), yet in this case the two means differ by 5.5 SEs, yet are declared to be not significantly different by Fisher's LSD procedure. The logical response of researchers such as the Bears would be to ensure good results by including an old, low-yielding oat cultivar in their trials (to increase the overall *F* value). Thus the statistical procedure would lead to a nonsensical waste of resources.

If the data from the four experiments were to be reanalyzed using the unrestricted LSD procedure, then the significance of the difference between MiteyOat and TrustyOat would be 0.1% in all four experiments. That is, the unrestricted LSD procedure is consistent; in fact, it is the only consistent procedure.

The Inconsistency of Other Multiple Comparison Procedures

The four data sets given in Table 2 were constructed to illustrate the inconsistency of Fisher's restricted LSD procedure because it is the MCP most commonly used in Agronomy Journal. Other well-known MCPs behave in a fairly consistent manner between these four data sets, and different examples are required to illustrate their inconsistency. Such examples are given in Saville (1990) for Tukey's HSD procedure.

Table 3. The significance of the difference between MiteyOat and TrustyOat in Baby Bear's experiment for multiple comparison procedures of varying levels of conservatism.

Multiple comparison procedure†	Significance
Fisher's restricted LSD	ns
Bonferroni	*
Tukey's HSD	*
Student–Newman–Keuls	*
Duncan's multiple-range test	**
Unrestricted LSD	***

* Significant at 0.05 probability level; ns, not significant.

** Significant at the 0.01 probability level.

*** Significant at the 0.001 probability level.

† LSD, least significant difference; HSD, honest significant difference.

Interestingly, if six well-known MCPs are used to analyze Baby Bear's data in Table 2, the significance of the difference between MiteyOat and TrustyOat varies markedly, from not significant to 0.1% significant (Table 3). Fisher's restricted LSD procedure is at one extreme, yielding a nonsignificant result. With the Bonferroni, Tukey's HSD, or Student–Newman–Keuls procedures, the difference between MiteyOat and TrustyOat is 5% significant (Table 3). Using Duncan's multiple-range test, the difference is 1% significant, while with the unrestricted LSD procedure, the difference is 0.1% significant. This ordering of the significance levels of the results from the various MCPs reflects their levels of conservatism and inconsistency, except that Fisher's restricted LSD is variable and unfairly disadvantaged by this example and would more typically be placed between Student–Newman–Keuls and Duncan's multiple-range test in the ordering.

In summary, the most conservative MCPs, such as Tukey's HSD, Bonferroni procedures, and the Student–Newman–Keuls test are also the most inconsistent procedures. Duncan's multiple-range test is the most consistent of the alternatives to the unrestricted LSD procedure. More recently, in genomics applications, the false discovery rate has been introduced by Benjamini and Hochberg (1995) and developed further by Storey and Tibshirani (2003) and others; such procedures, of necessity, also suffer from the problem of inconsistency.

Effect of Choice of Multiple Comparison Procedure on Sample Size Required

Statistics textbooks give various formulas for the determination of sample size (*n*) using a power analysis, but these formulas usually produce results similar to one another, which is surprising because the line of reasoning varies considerably. The common theme, however, is that they are normally based on considerations of a *t*-test, which in the context of MCPs means that they are appropriate for just one choice of MCP, the unrestricted LSD procedure (which is a multiple *t*-test).

For example, the formula for sample size (*n*) that I use in my statistics courses is based on assuming that the estimated sample size (*n*) is relatively large, so percentiles of the normal distribution (or equivalently the *t* distribution with residual

degrees of freedom = ∞) can be used in the formula. The proviso is that if the resulting estimate for n turns out to be small, then appropriate percentiles of t need to be substituted and a few iterations performed. The formula is

$$n = 2\left[\frac{(1.96+1.645)s}{\Delta}\right]^2 \quad [1]$$

where s is the expected pooled standard deviation, Δ is the minimum difference of interest (between two means), the required significance of difference = 5% (two-tailed test), and the required power of the test = 95% (probability of detecting differences of >Δ).

Example: If I substitute $s = 400$ and $\Delta = 800$, then the estimated sample size is

$$n = 2\left(\frac{3.605\times 400}{800}\right)^2 = 7 \text{ replicates} \quad [2]$$

As an aside, if there are to be eight experimental treatments and seven replicates laid out in a randomized block design, this means the residual degrees of freedom will be $7 \times 6 = 42$, which is large enough for no adjustment to be required to the percentiles of t used in Eq. [1] or [2].

More importantly, note that this estimated sample size only applies if you plan on analyzing your data using the unrestricted LSD procedure!

Sample Size Required with Tukey's Honest Significant Difference Procedure

For example, if the plan is to analyze data using Tukey's HSD procedure with eight experimental treatments, the formula for the sample size (n) required is

$$n = 2\left[\frac{(3.03+1.645)s}{\Delta}\right]^2 \quad [3]$$

where the constant 1.645 is unchanged because this relates to the required power of 95%; however, the 1.96 is replaced by the value of 3.03 (= $4.29/\sqrt{2}$, where 4.29 is obtained from standard tables of the range for eight treatments, again assuming ∞ residual degrees of freedom).

Example for eight treatments: If I substitute $s = 400$ and $\Delta = 800$, then the estimated sample size is

$$n = 2\left(\frac{4.675\times 400}{800}\right)^2 = 11 \text{ replicates} \quad [4]$$

Example for 20 treatments: If the plan is to include 20 experimental treatments, however, then the formula for the sample size (n) required is

$$n = 2\left[\frac{(3.54+1.645)s}{\Delta}\right]^2 \quad [5]$$

where the original 1.96 is replaced by the value of 3.54 (= $5.01/\sqrt{2}$, where the 5.01 is obtained from standard tables of the range for 20 treatments, again assuming ∞ residual degrees of freedom). If I again substitute $s = 400$ and $\Delta = 800$, then the estimated sample size is

$$n = 2\left(\frac{5.185\times 400}{800}\right)^2 = 14 \text{ replicates} \quad [6]$$

That is, the required sample size is seven replicates for the unrestricted LSD, 11 replicates for Tukey's HSD with eight treatments, and 14 replicates for Tukey's HSD with 20 treatments. In the latter case, the researcher needs to have twice as many replicates with Tukey's HSD as with the unrestricted LSD.

In general, the morals are that the calculation of sample size must be appropriate for the planned MCP and that this calculation should be performed routinely as part of the experimental design process.

The Practical Solution

The practical solution is as follows:

1. Abandon the idea of simultaneously formulating and testing all possible pairwise comparison hypotheses using data from a single experiment.
2. Instead, use the simplest MCP (the unrestricted LSD procedure) to formulate new hypotheses at a known "false discovery rate" (e.g., 5% of null comparisons), then independently test these new hypotheses in a second experiment using an appropriate set of preplanned pairwise comparisons.
3. This is normal scientific practice, so this solution fits well with the way in which reputable scientists operate.

That is, on the basis of the discussion above and in Saville (1985, 1990, 2003), the recommendation to researchers is that if the use of a MCP is appropriate, the unrestricted LSD procedure is the best choice given the proviso that it should be treated solely as an hypothesis generating method, not as a method for simultaneous formulation and testing of hypotheses.

This is a very similar conclusion to those reached independently by others (e.g., Rothman, 1990; Perneger, 1998), who have pointed out the undesirability of making adjustments for multiplicity.

Cultivar Evaluation Trials

As an aside, agronomists who conduct annual series of cultivar evaluation trials may find the formal structure of hypothesis formulation (1 yr) and testing (next year) hard to relate to. At the start of each season, they may have hypotheses, of varying strengths of conviction, that certain cultivars are the best in terms of yield or quality. These hypotheses will be strengthened or weakened by the data from the current season. Hence, they may prefer to view the process as one of "continuous modification of hypotheses" by trial work over successive seasons rather than as a more formal formulation and testing of hypotheses. In this case, the same recommendation stands, that this modification of hypotheses be done on the basis of the unrestricted LSD rather than any other MCP.

Advantages of Using the Unrestricted Least Significant Difference Procedure

Advantages of using the unrestricted LSD procedure are as follows:

1. It is the only consistent MCP.
2. It is the simplest procedure.
3. Its calculation is the easiest to check, so arithmetic errors are minimized.
4. It is the most flexible MCP, catering to unequal sample sizes and, if necessary, to unequal variances.
5. It has a constant Type I error rate (e.g., 5%), with all other MCPs having variable, nominal (in name only) Type I error rates.
6. It has maximum power, so it has the greatest chance of generating an interesting new pairwise comparison hypothesis.
7. The required sample size is easily calculated and the formula is given in standard statistics textbooks.

General Contrasts and Report Writing

This "practical solution" applies also to general contrasts (such as legumes vs. nonlegumes), not just pairwise contrasts.

Within a single experiment, there may be a mix of prespecified contrasts (perhaps including some pairwise comparisons) and contrasts or comparisons that have become interesting as a result of the experiment. For all contrasts (or ideas or hypotheses), the key thing when writing a report on an experiment is to be completely honest and to clearly describe which ideas you had before the experiment and which ideas were formulated as a result of the experiment. For preplanned pairwise comparisons and general contrasts, your experiment confirms or denies each hypothesis. For post-planned pairwise comparisons and general contrasts, however, your experiment has led you to formulate each hypothesis that needs to be confirmed or denied in a second experiment.

CONCLUSIONS

In summary, I suggest using the same statistical procedure in all cases:

- preplanned pairwise comparisons
- preplanned general contrasts
- post-planned pairwise comparisons (the MCP case)
- post-planned general contrasts

The statistical procedure is the traditional *F* test of a contrast, which is mathematically equivalent to carrying out a *t*-test of a contrast, which is also equivalent to performing an LSD test (or, more precisely, a *least significant contrast* test). This suggestion, of statistically analyzing both preplanned and post-planned contrasts in an identical manner, is statistically defensible only if the report on the experiment distinguishes clearly between testing and forming ideas and hypotheses.

ACKNOWLEDGMENTS

Thanks to Barry Glaz and Kathleen Yeater for encouraging me to write this article, the American Society of Agronomy for supporting my trip to Cincinnati for the associated symposium, and the journal reviewers for improvements to the manuscript.

REFERENCES

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B 85:289–300.

Carmer, S.G., and W.M. Walker. 1982. Baby Bear's dilemma: A statistical tale. Agron. J. 74:122–124. doi:10.2134/agronj1982.00021962007400010031x

Cochran, W.G., and G.M. Cox. 1957. Experimental designs. 2nd ed. John Wiley & Sons, New York.

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

Perneger, T.V. 1998. What's wrong with Bonferroni adjustments. BMJ 316:1236–1238. doi:10.1136/bmj.316.7139.1236

Rothman, K.J. 1990. No adjustments are needed for multiple comparisons. Epidemiology 1:43–46. doi:10.1097/00001648-199001000-00010

Saville, D.J. 1985. Multiple comparison procedures. Proc. Agron. Soc. N.Z. 15:111–114.

Saville, D.J. 1990. Multiple comparison procedures: The practical solution. Am. Stat. 44:174–180.

Saville, D.J. 2003. Basic statistics and the inconsistency of multiple comparison procedures. Can. J. Exp. Psychol. 57:167–175.

Saville, D.J., and J.S. Rowarth. 2008. Statistical measures, hypotheses, and tests in applied research. J. Nat. Resour. Life Sci. Educ. 37:74–82.

Saville, D.J., and G.R. Wood. 1991. Statistical methods: The geometric approach. Springer-Verlag, New York.

Storey, J.D., and R. Tibshirani. 2003. Statistical significance for genome-wide studies. Proc. Natl. Acad. Sci. 100:9440–9445. doi:10.1073/pnas.1530509100

Swallow, W.H. 1984. Those overworked and oft-misused mean separation procedures: Duncan's, LSD, etc. Plant Dis. 68:919–921.