Multiple Comparison Tests for Balanced, One-Factor Designs

Term Paper

FRST 533

Dec. 15, 2005

Craig Farnden

1.0 Introduction

Frequently after completing an analysis of variance test in a single factor experimental design, further analysis is desired to explore potential differences between many pairs of factor level means. Most textbooks describing common statistical procedures (i.e. Sokal and Rohlf 1995, Steel et al 1997, Hicks and Turner 1999, Sheskin 2004, Kutner et al 2005) recognize the preferred method of planning for such contrasts in advance based on expected relationships that are logical extensions of the experimental design. Such *a priori* contrasts must be completely independent (i.e. no one mean can be used in two separate comparisons) and the number of contrasts is limited by the available degrees of freedom.

It is also acknowledged, however, that desired comparisons cannot always be planned (they may be suggested by the experimental results themselves) or may be greater in number than the available degrees of freedom. A considerable number of statistical methods have arisen to deal with such situations. Such methods fall under the general category of multiple comparison tests or *a posteriori* comparisons.

In making a series of multiple comparisons, the general objective is to detect population differences where they occur, while keeping the cumulative probability of a Type I error below the desired probability for the entire experiment. While a series of difference-between-means t-tests could be employed to detect such differences, the sum of Type I error probabilities could easily escalate above the experiment wise α-level, particularly where a large number of contrasts are required. Controlling the experiment wise Type I error rate, then, is a major feature of multiple comparison tests.

2.0 Analysis Methods

Six multiple comparison methods are presented in this paper. The methods selected are a commonly applied subset of those available for balanced, single factor studies. The descriptions of various methods below have been compiled from a number of textbooks and manuals (Hicks and Turner 1999, Kutner et al 2005, SAS Institute 2003, Sheskin 2004, Sokal and Rohlf 1995, Steel et al 1997), which are used in this paper both as sources for the methods and as interpretive guides to their application – no original sources (i.e. journal papers) have been consulted.

Bonferroni Method

The Bonferroni (also Bonferroni-Dunn) method of paired comparisons allows any number of unplanned comparisons between two means, and permits comparisons for both balanced and unbalanced designs. It is based on ensuring that the probability of no Type I errors across all tests is at least 1-α. In order to ensure this requirement, the allowable error rate is divided up amongst all pairs to be tested to get a new rate (α') that is used for each paired test:

$$\alpha' = \frac{\alpha}{k}$$ where k equals the total number of comparisons to be completed

For each test of factor level means a and b:

H_0: $\mu_a = \mu_b$
H_1: $\mu_a \neq \mu_b$

a t-test can be established whereby:

$$t_{calc} = \frac{\left(\bar{y}_a - \bar{y}_b\right) - 0}{\sqrt{MSE\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}} \qquad \text{and} \qquad t_{crit} = t_{1-\alpha'/2, df_{error}}$$

For each case tested, if $t_{calc} > t_{crit}$, then the null hypothesis must be rejected, indicating a significant difference between the means. Alternatively, a p-value can be determined for t_{calc} and compared to α' (a convenient method where p-values have been calculated using a statistics software program such as SAS). If the p-value < α', the null hypothesis must be rejected.

Use of α' in this method is considered to result in a conservative test: the probability of no Type I errors will almost always be underestimated. With this method, the worst case scenario is that the cumulative probability of one or more Type I errors across all tests will equal α.

Scheffé Method

The Scheffé method allows for the same types of comparisons as the Bonferroni method, with the addition of allowing comparisons between groups of means (i.e. H0: $\mu_a = (\mu_b+\mu_c)/2$). For pairwise comparisons where:

H_0: $\mu_a = \mu_b$
H_1: $\mu_a \neq \mu_b$

the test statistic for the Scheffé test is:

$$S = \frac{\hat{L}}{s(\hat{L})} \qquad \text{where } \hat{L} = \sum_{j=1}^{J} c_j \bar{y}_{\bullet j} \quad \text{and} \quad s(\hat{L}) = \sqrt{MSE \times \left(\sum_{j=1}^{J} c_j^2 \times \frac{1}{n_j}\right)} \quad \text{and} \quad \sum_{j=1}^{J} c_j = 0$$

For the purposes of a multiple comparison between all pairs of factor level means in a trial, the value of c_a in all cases is 1/2, and for c_b it is –1/2.

The test statistic is compared to a critical value:

$$S_{crit} = \sqrt{(J-1) \times F_{1-\alpha, J-1, n-J}}$$ (note that the F value is the same as that for the ANOVA)

For the purposes of simultaneous multiple comparisons in a balanced trial (where all values of n_j are equal), it is most useful to feed the value of S_{crit} back into the calculation of S to get a critical difference between means where:

$$\hat{L} = S_{crit} \times s\left(\hat{L}\right) \qquad \text{and} \qquad diff_{crit} = 2 \times \hat{L}$$

For any pairs of means with a difference greater than $diff_{crit}$ the null hypothesis should be rejected: there is a significant difference between the means.

Tukey's HSD Test

The Tukey HSD (*Honestly Significant Difference*) test uses the studentized range satistic (q) to find significant differences between any pair of means out of a family of J means in a balanced design. For each possible pair of means a and b:

H_0: $\mu_a = \mu_b$
H_1: $\mu_a \neq \mu_b$

a test can be established whereby:

$$q_{calc} = \frac{\bar{y}_a - \bar{y}_b}{\sqrt{\frac{MSE}{n}}} \qquad \text{and} \qquad q_{crit} = q_{\alpha,J,df_{error}}$$

where J = number of treatments, and n = number of observations per treatment.

If the value of q_{calc} exceeds the value of q_{crit} from the tables of the studentized range, then the null hypothesis must be rejected, indicating a significant difference between the means.

As with the special case of a Scheffé test for comparison of all means in a balanced design, we can determine a single value for the critical significant difference by substituting the critical value back into the formula for q_{calc}:

$$\left(\bar{y}_a - \bar{y}_b\right)_{crit} = q_{calc} \times \sqrt{\frac{MSE}{n}}$$

For any pairs of means with a difference greater than $\left(\bar{y}_a - \bar{y}_a\right)_{crit}$ the null hypothesis should be rejected: there is a significant difference between the means.

Student-Newman-Keuls Test (SNK)

The Student-Newman-Keuls (also Newman-Keuls) test is similar to Tukey's HSD test in that it uses the same equation for q_{calc} and the same tables for the studentized range of critical values. Instead of using a single critical value of the test statistic for all pairs of means, however, the critical value will vary depending on how many other treatment means are ranked between the two being tested.

To start this procedure, all of the factor level means are ranked in a list, and all possible ranges of means are determined. Construct a table of possible ranges. For a list of 5 means i, ii, iii, iv and v, the possible ranges will include:

i to v	ii to v	iii to v	iv to v
i to iv	ii to iv	iii to iv	
i to iii	ii to iii		
i to ii			

where range (i to v) includes 5 means with a value of $k = 5$, and range (iii to iv) includes 2 means with a value of $k = 2$.

For each row in the table of ranges, start with the means with the widest separation (a and b) such that for:

$H_0: \mu_a = \mu_b$
$H_1: \mu_a \neq \mu_b$

a test can be established whereby:

$$q_{calc} = \frac{\bar{y}_a - \bar{y}_b}{\sqrt{\frac{MSE}{n}}} \qquad \text{and} \qquad q_{crit} = q_{\alpha,k,df_{error}}$$

(note that the value *J* from the Tukey procedure has been replaced by *k* in the determination of q_{crit})

If the value of q_{calc} exceeds the value of q_{crit} from the tables of the studentized range, then the null hypothesis must be rejected, indicating a significant difference between the means.

In cases where a significant difference is found, move to the next pair of means (to the right) in the row and repeat the test. When each row is completed, move to the next row. Note that for any pair of means in the table for which no significant difference is detected, all pairs of means in subordinate ranges (any cells in the table that are either lower in the same column or in equal or lower rows in all columns further to the right) will also have no significant difference and will not need to be tested.

Duncan's New Multiple Range Test

Duncan's test is very similar to the SNK procedure. It differs from the SNK test in that it uses a variable significance level (α') depending on the number of means in the range being tested:

$$\alpha' = 1 - (1 - \alpha)^{k-1}$$

This difference is accommodated by using a different set of tables in which the family wise significance level (α) is used to enter the table, but critical values of the test statistic contained within the tables have been adjusted based on the number of means in the range and their effect on α'. With the exception of the alternate table of critical values, the Duncan test is conducted in an identical manner as the SNK procedure.

Dunnett's Test

Dunnett's test is a special case of multiple comparisons where all factor level means are compared to the control, but no other pairs of means are tested.

For each test of a treatment mean a:

$H_0: \mu_a = \mu_0$
$H_1: \mu_a \neq \mu_0$

A Dunnett modified t statistic can then be established whereby:

$$t_{D\,calc} = \frac{\left(\bar{y}_a - \bar{y}_0\right)}{\sqrt{\frac{2MSE}{n}}} \qquad \text{and} \qquad t_{Dcrit} = t_{D\,1-\alpha,J-1,df_{error}}$$

For each case tested, if $t_{D_{calc}} > t_{D_{crit}}$, then the null hypothesis must be rejected, indicating a significant difference between the means.

As with the special case of a Scheffé test for comparison of all individual means in a balanced design and the Tukey HSD test, we can determine a single value for the critical significant difference by substituting the critical value back into the formula for $t_{D_{calc}}$:

$$\left(\bar{y}_a - \bar{y}_0\right)_{crit} = t_{D\,calc} \times \sqrt{\frac{2MSE}{n}}$$

3.0 Example

In a trial to test growth enhancing treatments in repressed lodgepole pine stands, 10 different treatment regimes were tested that included selected combinations of thinning and fertilization in a randomized complete block design. For a variety of reasons including the confounding effects of further thinning by snowshoe hares in fertilized plots, no treatment interactions can be reliably evaluated, and each treatment combination will be evaluated as a discrete treatmen. Given the detection of at least one significant difference between treatment means, differences between individual pairs of treatment means can then be evaluated using multiple comparison techniques.

The trial consists of three blocks, each containing 10 treatment units (including an untreated control). The response variable is 'apparent' site index, based on measurement of the last five years of growth (a modified growth intercept method was used to convert five years of leader growth at various reference heights to apparent site index) on the best five trees in a plot. The means by treatment unit are listed in Table 1.

Table 1. Site index (m) data by treatment unit.

	Block		
Treatment	1	2	3
0	9.62	9.42	9.82
1	10.36	10.32	11.88
2	13.7	11.7	11.88
3	13.08	10.38	11.7
4	18.06	17.26	17.24
5	10.02	10.3	10.1
6	15.1	16.06	14.18
7	15.94	13.18	13.38
8	17.72	15.7	16.52
9	15.14	15.8	16.34

Multiple Comparison Results

A one-factor analysis of variance was conducted using SAS (output in Appendix 1) where:

H_0: $\mu_0 = \mu_2 = \mu_3 = \ldots = \mu_9$
H_1: at least one pair of means is not equal

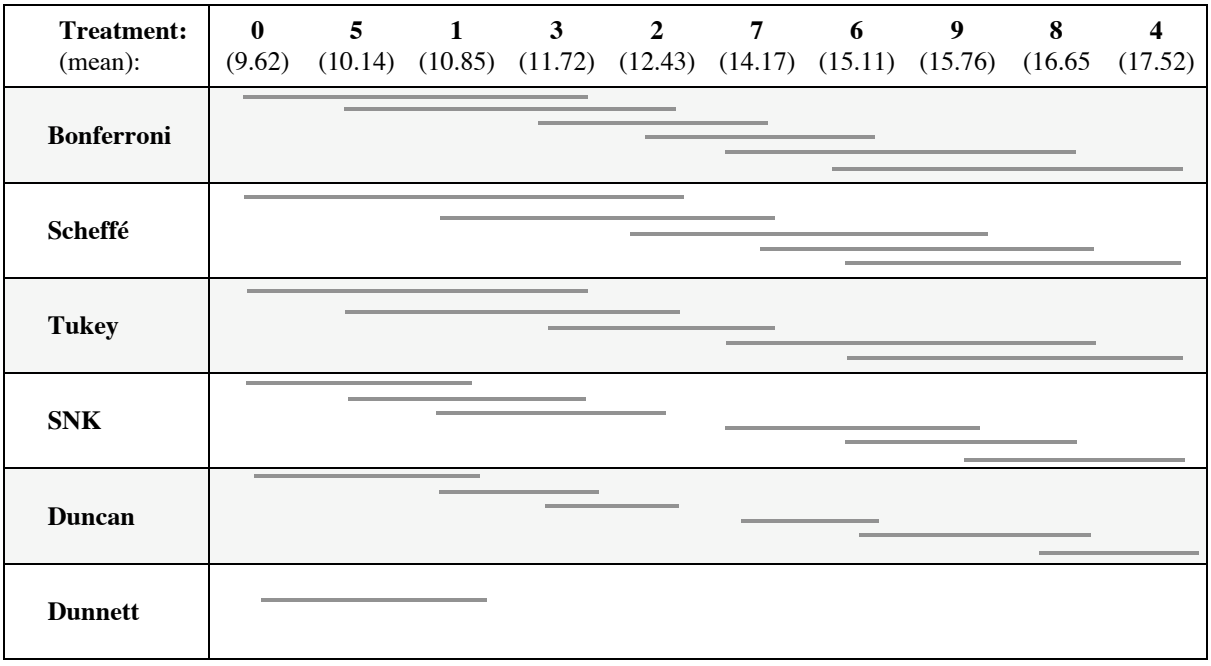
Evaluation of the residual plot confirmed the assumption of homogenous variance, and both the normality plot and normality tests confirmed the assumption of a normal distribution of errors. The F-value for the analysis was 26.21, with a corresponding p-value of <0.0001 ($<\alpha$ = 0.05). At least one pair of treatment means, therefore, is significantly different.

For each of the multiple comparison tests discussed above, an additional SAS run was completed to test for differences between pairs of treatment means (SAS output is included in Appendix 2). Results of these analyses are presented in three different manners in Table 2 and Figures 1 & 2.

Table 2. Variation in multiple range test results using a ranked list of means and letter codes to illustrate groupings. Any two means with the same single-letter code cannot be said to be significantly different. For example, using the Bonferroni test, treatment 2 is a member of groups b, c and d, and cannot be distinguished as significantly different from other members of those groups. It would, however, be significantly different from the mean for treatments 0, 9, 8 and 4. In the case of the Dunnett test, the letters simply indicate which means are not significantly different from the control (treatment 0).

		Method					
Treatment	Mean	Bonferroni	Scheffé	Tukey	SNK	Duncan	Dunnett
0	9.62	a	a	a	a	a	a
5	10.14	ab	a	ab	ab	a	a
1	10.85	ab	ab	ab	abc	ab	a
3	11.72	abc	ab	abc	bc	bc	
2	12.43	bcd	abc	bc	c	c	
7	14.17	cde	bcd	cd	d	d	
6	15.11	def	cde	de	de	de	
9	15.76	ef	cde	de	def	e	
8	16.65	ef	de	de	ef	ef	
4	17.52	f	e	e	f	f	

Figure 1. For each of the six tests, groups of means that are not significantly different from each other are indicated by a solid, horizontal bar. For an individual mean *x*, the range indicated by the collection of bars that falls below *x* indicates all means from which it cannot be significantly distinguished. For the Dunnett test, the bar indicates which means cannot be distinguished from the control.



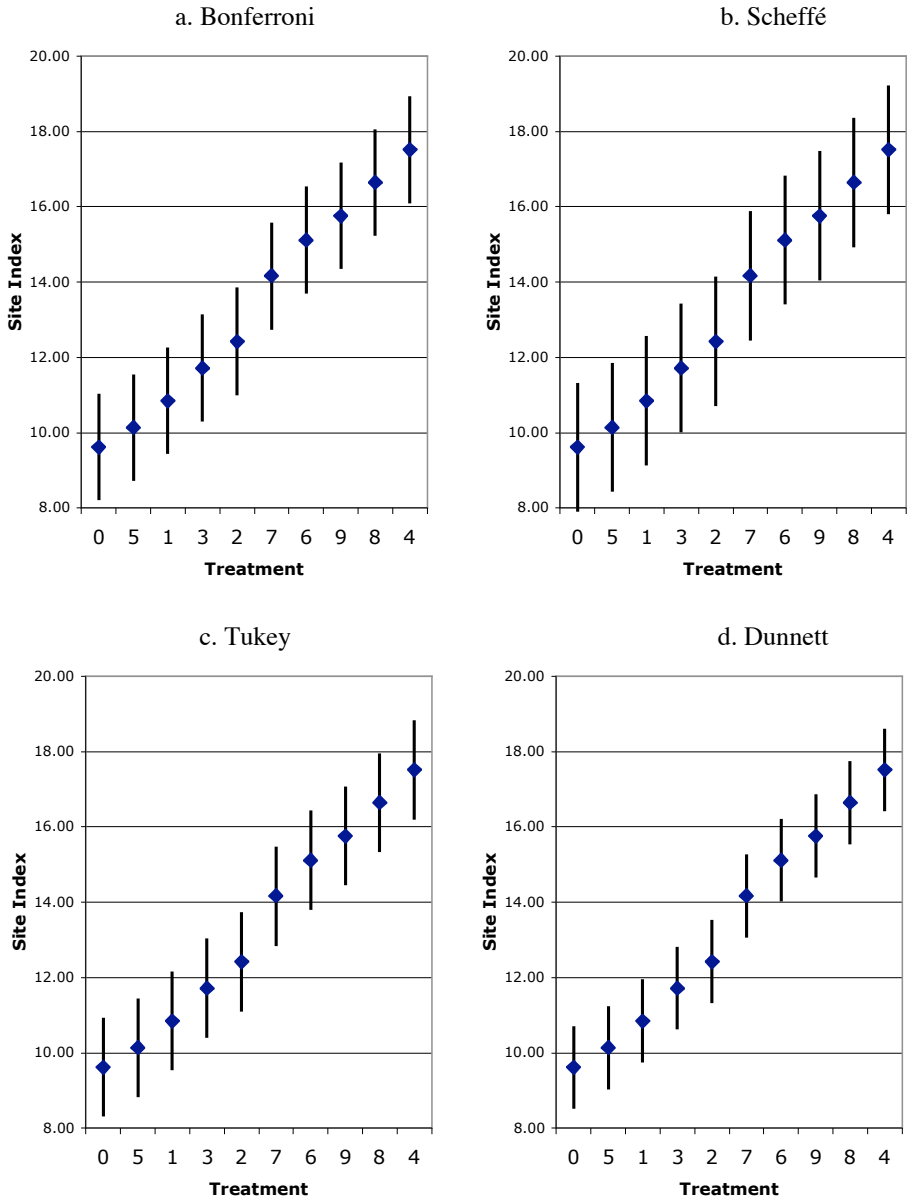


Figure 2. Differences between means as depicted by error bars. Each tail of the error bars represents 1/2 of the critical difference between means (a fixed critical difference is not available for the SNK and Duncan tests, so they cannot be depicted in this manner). Where tails for two means overlap, the means are not significantly different. For the Dunnett test, the only overlap of concern is with the control (treatment 0) – no other differences should be interpreted from this chart.

The methods of data presentation in Table 2 and Figure 1 are capable of showing results for all six methods, with Figure 1 providing perhaps a quicker overall perception of the groupings. A presentation such as the one in Figure 2 has the advantage of displaying the degree of similarity or dissimilarity between adjacent or nearby means, but makes it difficult to distinguish significant versus non-significant differences under marginal conditions. Still other presentations are available that combine features of those presented here.

In contrasting the various methods, it is interesting to note the number of other means for which each treatment mean is significantly different (Table 3). Other than for the Dunnett test, the Scheffé method distinguished the least number of differences, followed in order by Bonferroni, Tukey, SNK and Duncan. The Dunnett test, for differences from the control, was on par with the other tests which made the finest level of distinctions between means.

Table 3. Number of other means (by treatment) for which no significant difference was found in each of the tests.

	Treatment									
	0	**5**	**1**	**3**	**2**	**7**	**6**	**9**	**8**	**4**
Bonferroni	4	5	5	6	6	6	6	5	5	4
Scheffé	5	5	6	6	6	7	6	6	5	4
Tukey	4	5	5	6	5	6	5	5	5	4
SNK	3	4	5	4	3	3	4	5	4	3
Duncan	3	3	4	3	2	2	4	3	4	2
Dunnett	3	-	-	-	-	-	-	-	-	-

Discussion

The varying results obtained in the 6 multiple comparison tests in this paper result from the relative importance placed on Type I and II errors. Methods that place the most stringent controls on maintaining the experiment wise α level will fail in more cases to distinguish between means that are truly different. Such methods are considered to be highly conservative. Methods that are less conservative attempt to make a higher number of separations between truly different means while at the same time minimizing the risk of inflating Type I errors.

The relative conservatism of these methods, along with their approach to and/or success at controlling Type I errors is provided in Table 4. The ranking agrees with the results in Table 3 for the example in this paper.

The choice of multiple comparison methods depends on the situation being tested. The scope of this paper has been fairly limited to pair-wise comparison of means in a one factor study with a balanced design. The ranking of conservatism and the availability of methods may change for other situations.

The choice of methods within the scope of this paper is largely a function of the degree and type of risk that one is willing to take. Most authors emphasize the absolute control of the experiment wise α level and favor the Tukey test (α level is controlled but detects more differences than Scheffé and Bonferroni). Going beyond the selection of a single test, however, many authors recommend use of several tests. For pairs of means that are always

separated as significantly different or combined as not significantly different regardless of the test used, there is no problem. For the marginal pairs, it is worth considering whether or not a difference is important to the phenomena being studied or the conclusions to be made. If not, there is no cause for concern. If the distinction is important, then further study may be warranted.

Table 4. Ranking of multiple comparison methods by conservatism.

Method	Conservatism and comments	Sources
Scheffé	A highly conservative test usually associated with complex comparisons; not normally recommended for the type of analysis in this report. p(no Type I errors) ≥ (1-α).	Hicks and Turner 1999 Kutner et al 2005 Sheskin 2004 Sokal and Rohlf 1995 Steel et al 1997
Bonferroni	A highly conservative (similar to Scheffé) test often used for multiple pair wise comparisons. p(no Type I errors) ≥ (1-α)	Dallal 2001 Kutner et al 2005 Sheskin 2004 Sokal and Rohlf 1995
Tukey	A moderately conservative test that exactly protects the experiment wise α level: p(no Type I errors) = (1-α). Possibly the most commonly recommended procedure for making all possible pair wise tests.	Dallal 2001 Hicks and Turner 1999 Kutner et al 2005 Sheskin 2004 Sokal and Rohlf 1995 Steel et al 1997
Dunnett	A moderately conservative test that exactly protects the experiment wise α level: p(no Type I errors) = (1-α). The Dunnett test will, on average, find more means that are significantly different from the control than does the Tukey test as it maintains more power through making a smaller number of comparisons.	Sheskin 2004 Sokal and Rohlf 1995 Steel et al 1997
SNK	A moderately conservative test that protects the experiment wise α level: p(no Type I errors) = (1-α). Once commonly used, it has become less so due to concerns regarding detection of differences no replicated in other methods with similar α level protection.	Dallal 2001 Hicks and Turner 1999 Sheskin 2004 Sokal and Rohlf 1995 Steel et al 1997
Duncan	A lightly conservative test that does not protect the experiment wise α level; results will often be close to those obtained with a set of independent t-tests at the α level.	Dallal 2001 Steel et al 1997

References

Dallal, G.E. 2001. Multiple comparison procedures. http://www.tufts.edu/~gdallal/mc.html

Hicks, C.R., and Turner, K.V. 1999. Fundamental concepts in the design of experiments. 5th edition. Oxford University Press, New York NY.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. 2005. Applied linear statistical models. 5th edition. McGraw-Hill Irwin, New York NY..

Sheskin, D.J. 2004. Handbook of parametric and nonparametric statistical procedures. 3rd edition. Chapman & Hall / CRC Press, Boca Raton FL.

SAS Institute 2003. SAS online doc9.1. SAS Institute, Cary NC. http://support.sas.com/91doc

Sokal, R.R. and Rohlf, F.J. 1995. Biometry. 3rd edition. WH Freeman and Co., New York NY.

Steel, R.G.D., Torrie, J.H. and Dickey, D.A. 1997. Principles and procedures of statistics, a biometrical approach. 3rd edition. McGraw-Hill Co. Inc., New York NY.

Appendix 1. Analysis of Variance for Example Data

Appendix 2. Multiple Comparison Test Output from SAS