OnLine Journal of Biological Sciences 1 (8): 723-727, 2001

Multiple Comparisons

Hülya Atil and Yakut Unver
Faculty of Agriculture, Ege University, Turkey

Abstract: Multiple comparison procedures allow detecting differences among treatment means. These procedures include pairwise, multiple comparisons, planned orthogonal and non-orthogonal contrasts and orthogonal polynomials. In many situations the use of multiple comparison procedure has been either appreciated or not the most meaningful method of analysing the data. The objectives of many animal experiments are to detect meaningful relationships among treatments and associated responses and also to identify relationships among biological responses. In this study, Least Significant Differences test, Student –Newman-Keuls test, Tukey test, Duncan test, Scheffe test, Hartley test, Dunnet t-test, Bonferroni test, Waller-Duncan test, William t-test, confidence interval, orthogonal polynomials and orthogonal contrasts are discussed.

Key words: Multiple comparisons, pairwise comparisons, contrasts

Introduction

One objective of an analysis of variance is to compare means of the response variable. When comparing more than two means, the analysis of variance would indicate if the means are significantly different from each other but will not indicate which means differ from other. Generally, it is not sufficient to show that the differences between treatment means using F-test in the analysis of variance. So, researchers want to compare the treatment means depend on treatment properties. There are two types of error during the comparison of means. These are "comparisonwise error" and "experimentwise error" (Steel and Torrie, 1980). There is no rule to which one of them is more appropriate.

Multiple comparison methods are available the provide more detailed information about the differences among the means. Some of the more popular methods include Least Significant Difference (LSD) test, Student-Newman-Keuls (SNK) test, Tukey test, Duncan test, Scheffe test, Hartley test, Dunnett test, Bonferroni-t test, Waller-Duncan test, William-t test, confidence interval, orthogonal polynomials and orthogonal contrasts.

With 6 treatments there are 15 possible pairs of treatments that could be compared using these methods. Clearly it does not make sense to take all 15 comparisons. The 6 treatment means would be replaced by 15 differences, which can hardly aid interpretation. The misleading impression would be given of having estimated 15 quantities from just 6 means. Significance levels calculated as described above the meaningless. So which comparisons should be made? With 6 treatments there are 5 degrees of freedom, corresponding to 5 'independent' differences among the 6 quantities. Thus it is usually misleading to make more than 5 comparisons.

The comparisons made should correspond to the objectives the experimenter had in studying these particular treatments. The experimenter should have chosen these treatments because he wanted to make certain comparisons between them. It should be possible to state the objectives of the experiment in terms of comparisons between treatments.

Interpreting the results of experiments is that the appearance of the results may suggest additional questions beyond the specified number of orthogonal comparisons. The temptation to make comparisons suggested by the data must always be resisted, as the nominal confidence level will then always be invalid. In multiple comparisons, Type I error levels have different values because more than two means are tested. Sometimes it is possible to reject the hypothesis, which should be accepted. Type I error is called as an "experimental error". Multiple comparisons can be subdivided into 3 groups according to precaution of Type I and Type II errors (Yildiz and Bircan, 1994):

I. Group: In this group, tests are not consisting any precaution, for obstruction the increase in experimental error. Therefore these tests have greater error rate by increasing the number of comparisons. The Least Significant Differences test (LSD) has these properties.

II. Group: In this group, tests are consisting of the strongest precaution for preventing the experimental error. This group includes Tukey, Scheffe and Dunnett tests. In LSD test, some comparisons can be significant which should be non-significant while for these tests some differences of means can be non-significant which should be significant.

III. Group: This group is between first two groups. In this group, tests are consisting of the precaution against the experimental error rates according to differences between means. Duncan and Student-Newman-Keuls tests have these attributes.

Least Significant Differences Test (LSD): The LSD is a valid test procedure for planned comparisons and more appropriate when the numbers of means are less. LSD can be investigated by two ways:

1. Ordinary LSD procedure (protected LSD),
2. Unprotected LSD procedure (Bek and Efe, 1995).

The ordinary or unprotected LSD test is the most appropriate for pairwise treatment comparisons. It should be used only for independent comparisons and comparisons between means that are adjacent in rank (Lowry, 1992). The first procedure is called LSD test. This can be used only the condition of rejection of H_0 hypothesis. However, this is not necessary for the second procedure. There are t value tables for unprotected LSD procedure (Chew, 1976).

$$LSD = t_{\alpha,v}\sqrt{\frac{2HK\Omega}{r}} \quad \text{(for equal r)}$$

LSD is given by
where, LSD : critical value, v : error degrees of freedom and r: replication number for per treatment.

$$LSD = t_{\alpha,v}\sqrt{HK\Omega\left[\frac{1}{r_1}+\frac{1}{r_2}\right]} \quad \text{(for unequal r)}$$

where, r_1 and r_2 are the replication number for i^{th} and j^{th} treatments, respectively.
It is not appropriate to use for more then three means. Because by increasing number of treatments Type I error is increased (Bek and Efe, 1995).
When all of the means are compared with each other the power of test is decreased because the comparisons will be done are not independent (Ikiz *et al.*, 1996).

Student-Newman-Keuls' Test (SNK): Each of the three persons named contributed to this that, but it is also referred to as the Newman-Keuls' test, or simply Keuls' method. (Steel and Torrie, 1980). It provides for multiple critical values from the distribution of the Studentized range. In this test, the difference of any two means is compared by using a changeable value according to the number of rank among means instead of the consistent critical value. The most important property of this test, it is not necessary to reject H_0 hypothesis. The test protects the nominal Type I error as well as does Tukey' s HSD and is slightly more sensitive but more cumbersome to use (Gill, 1973).
SNK is given by

$$W_i = Q'_{\alpha(p_i,v)}\sqrt{\frac{HK\Omega}{r}} \quad \text{(for equal r)}$$

where, p_i: the number of rank among means, v: degrees of freedom and $Q'_{\alpha(p_i,v)}$ special table value for SNK.

$$W_r = Q'_{\alpha(r_s,v)}\sqrt{HK\Omega\left[\frac{r_1+r_2}{2r_1r_2}\right]}$$

(for unequal r)
(Sokal and Rohlf, 1969).

where , r_1: replication number in 1^{sd} group and r_2: replication number in 2nd group.
This method is not used for estimating confidence interval. Also, the error rate is an experimentwise error (Bek and Efe, 1995). In this test, while H_0 hypothesis is not rejected it can be found significant differences between means.

Tukey test (HSD): Tukey (1953) developed a honestly significant difference (HSD) test, which uses a t-like statistics based upon the standardized range (Gill, 1973). It is appropriate for all pairwise comparisons. The error rate is experimentwise error. This test can also be used for calculating confidence intervals for all difference of means.

When the replication numbers are equal:

$$T = Q'_{\alpha,t,v}\sqrt{\frac{HK\Omega}{r}}$$

where, t: number of treatment, Q': table value using in SNK.

When there is not big differences between group replications, Bancroft (1968), recommended harmonic means of replication numbers (r_h) instead of r:

$$r_h = \left[\frac{\frac{1}{r_1}+\frac{1}{r_2}\ldots\ldots+\frac{1}{r_t}}{t}\right]^{-1}$$

and test statistics:

$$T = Q'_{\alpha,t,v}\sqrt{\frac{HK\Omega}{r_h}}$$

$$= Q'_{\alpha,t,v}\sqrt{HK\Omega\left[\frac{1}{\frac{1}{t}\left(\frac{1}{r_1}+\frac{1}{r_2}\ldots\ldots+\frac{1}{r_t}\right)}\right]}$$

When the replication numbers are unequal:

$$T = Q'_{\alpha,t,v}\sqrt{HK\Omega\left(\frac{1}{2}\left(\frac{1}{r_i}+\frac{1}{r_j}\right)\right)}$$

When the variances have big heterogeneous, in this case another procedure like Tukey is used (Gill, 1971):

First of all, sample variance is calculated for difference of means by t(t-1)/2 pair of comparisons:

$$HK\Omega_{ij} = \left(\frac{HK\Omega_i}{r_i}+\frac{HK\Omega_j}{r_j}\right), \quad i \neq j$$

and degrees of freedom for this test is given by

$$v = \frac{HK\Omega_{ij}}{\frac{\left(\frac{HK\Omega_i}{r_i}\right)^2}{r_i-2}+\frac{\left(\frac{HK\Omega_j}{r_j}\right)^2}{r_j-1}}$$

Test statistics, minimum significant differences (MSD) is

$MSD = t_{(\alpha T/2,V)} * HKO_0$

where, $t_{(\alpha T/2,V)}$: Student t value, α_T is according to compared the number of mean is given by,

$\alpha_T 1-(0.95)^{2/t(t-1)}$

When the replication numbers are not equal and variances are heterogeneous for group of means (Spjotvell and Stoline, 1973; Hochberg, 1976):
The test statistics is

$$T = Q'_{\alpha_T,v}\sqrt{\frac{HK\Omega}{r_s}}$$

where r_s = the smallest one r_1 and r_2.

The degree of unbalanced between replication number is given by

$U = (\text{Max } r_i / \text{Min } r_i), \qquad (r_1, r_2, \ldots, r_t)$

This test is known as "Modified Tukey Test Statistics".

When the below conditions are valid, Modified Tukey test is preferred to Scheffe test:

a) each U value and $t \geq 15$,
b) $U \leq 2$ and $6 \leq t < 15$,
c) $U \leq 1.2$ and $t < 6$.

Chew (1976) recommended the mean of Tukey's Q' value and SNK's Q' value to exterminate the hardness of Tukey test.

So,

$$T_{.} = \left[\frac{1}{2}\left(Q'_{\alpha,p,v} + Q'_{\alpha,(p),v}\right) * \sqrt{\frac{HK\Omega}{r}}\right]$$

This test is known as "Tukey' s multiple range test".

Duncan test: This test developed by Duncan in 1955. It is more powerful than HSD and SNK tests, but it does not protect adequately the Type I error rate. For that reason, it is mostly misused. However , Type I error level is bigger than the one that SNK has (Bek and Efe, 1995). Another property of this test that it is not necessary the rejection of H_0 hypothesis for this test. In this test, α significance level is equal to 1-(1-α). Duncan test considers the ordered set of treatment mean yields and tests each group of two, three, four, or more, successive means to assess whether the spread between the highest and lowest of each group is significantly large in terms of the experimental error mean square, allowing for the number of treatments in the group being tested.
The test departs from the SNK procedure where a constant significance level is used at all testing stages.

Test statistics is

$$D_{.} = Q_{\alpha,(p,v)}\sqrt{\frac{HK\Omega}{r}} \qquad \text{(for equal r)}$$

where, $Q_{\alpha,(p,v)}$: critical value, p : the number of rank among the means put in order, v : error degrees of freedom and r : replication number for each group.
This method is not used for estimating of confidence interval. It cannot use for unequal r, in this case Tukey or SNK tests can be used.

Scheffe test: Scheffe test is appropriate for general comparisons. This test can be used only in the condition of rejection of H_0 hypothesis. The error rate is experimentwise error. So that, Scheffe (1959) recommended that the α level is 0.10 (Bek and Efe, 1995). In this method, all possible contrasts can be tested for significance or confidence intervals. This test has a large critical value. So that, it is conservative in this sense and the power may be low.
Many statisticians tell that this technique overprotects against Type I error when making orthogonal comparisons and do not recommend it in this case (Steel and Torrie, 1980).

Test statistics is

$$S = \sqrt{(p-1) * F_{\alpha,(p-1),v)} * \frac{2HK\Omega}{r}}$$

This test is also used for estimation of confidence interval. When the replication numbers of each of group are not equal, Scheffe test is more appropriate than Tukey test (HSD).

Hartley test: Hartley is developed the new method calling "Hartley Test" for comparing the group means. In this method a Q value is used for differences between means according to their magnitude. For this method test statistics (D) is calculated and bigger differences from D are accepted significant.

Test statistics is

$$D = Q * S_{\bar{x}}$$

where, Q: table value.
This method shows the more differences are significant. But, Tukey is more prefer method than Hartley, because of confidently and useable (Düzgünes, 1963).

Dunnett test: In some cases, researcher's goal is to detect the differences between control and other treatments. It is not necessary the rejection of H_0 hypothesis for this test. Dunnett test has one critical value like Tukey and Scheffe tests. The error rate is experimentwise error. By this test, (t-1) comparison can be done in t treatment.

Test statistics is

$$\left.\begin{array}{l} DT = Q'_{\alpha,p,v}\sqrt{\dfrac{2HK\Omega}{r}} \\ \text{and} \\ DT = t_{.} * \sqrt{\dfrac{2HK\Omega}{r}} \end{array}\right\} \text{(for equal r)}$$

where, Q' and t_0 table values for Dunnett test. (Dunnett' s t_0 values and Student-t values are same for $t = 2$ treatment. But for more than two treatments, Dunnett' s values are bigger than Student-t values).

When control and other treatment groups have different replication numbers is given by

$$DT = t_{.} * \sqrt{HK\Omega * \left(\frac{1}{r_0} + \frac{1}{r_i}\right)} \quad \text{(for unequal r)}$$

where r_0: replication number of control group, r_i: replication number of other i^{th} treatment group.

Bonferroni t test: If a scientist is interested in advance in making a relatively small number of comparisons is interested in making a relatively small number of comparisons among means, and contrasts prove to be nonorthogonal, a procedure involving Bonferroni-t test (Gill, 1973).

Test statistics is

$$Q = \sum 1_i * \mu_i$$

where 1_i: orthogonal coefficients and μ_i : mean of treatment. In this test, if the number of contrast involving to comparison is near to total number of treatment, sensitivity decreases. Also, error is dividing to number of comparison (orthogonal or nonorthogonal) whereas for other tests, error is valid for whole experiment.

Waller-Duncan test: Enough information cannot be obtained about this test. Waller and Duncan are developed this test in 1969. The test is mostly similar to LSD with some differences. This test has the maximum accuracy degree among multiple

comparison tests (Waller and Duncan, 1969). In this test, in spite of rejection of H_o all treatments are not evaluation at the same sensitivity. In the other word, in spite of F test value is near to F table value H_O can be rejected. In this case, Waller-Duncan test cannot compare the means with same error level.

William t test: Enough information cannot be obtained about this test. However, this test is preferred when the control group and treatments have different doses with quantitative variables e.g. hormone, vitamin and fertilizer. Also, this test has similar application area with Dunnett, at the same time it is more powerful than Dunnett test.

Confidence interval: The multiple comparison tests except LSD are used if there is not prior knowledge about experiment, confidence interval is an alternative method for this situation (Ikiz, 1997).
The most confidentional way for multiple comparisons that to obtain the confidence intervals for treatment means and can be shown and interpreted them with graphics.

Confidence interval is

$$\bar{X} \pm t_{\alpha/2,s} \sqrt{\frac{HK()}{r}}$$

Orthogonal polynomials: If the treatments are quantitative variables (e.g. 0, 30, 60, 90 doses), orthogonal polynomials are used. This can be doses of medicine or fertilizer and acid concentrations.
It can be used only these doses have same interval (0, 30, 60 etc.) and have same number of replication. In experiments have these properties the effects of treatments can be tested by linear, quadratic or cubic, etc. For this, the fitting of the orthogonal polynomials is done by using appropriate coefficients (Pascal' s triangle).

Orthogonal contrasts: This method's goal is to compare the some of treatments instead of all of them. It is not necessary the rejection of H_0 hypothesis for this test. This test is recommended if the variables are qualitative and have prior knowledge of experiment.
There are some rules of orthogonal contrasts:

1. Sum of the coefficients must be equal to zero ($\sum_{i=1}^{t} l_i = 0$ for $l_1, l_2, \ldots\ldots l_t$),
2. Paired comparisons for t treatments must be orthogonal

($\sum_{i=1}^{t} l_{1i} l_{2i} = 0$ for the coefficients of the first comparison are $l_{11}, l_{12}, \ldots\ldots l_{1t}$ and the second one's are $l_{21}, l_{22}, \ldots\ldots l_{2t}$).

By this way, (t-1) comparison can be done. Whenever comparisons are orthogonal, ignoring the significance of treatments contrasts can be done (Ikiz *et al.*, 1996).

Results and Discussion

There are a lot of methods to compare the treatment means. Researcher, should make every attempt to select the statistical procedure, which is appreciate for their data analysis and interpretation rather than merely following the approach commonly used by some other scientists in their field. The problem of choosing sensible comparisons between should always be related to the type and structure of the treatments.

Pairwise, nonindependent, multiple comparison procedures are appropriate for comparing sets of unstructured, qualitative treatments, especially when the objective is to perform preliminary data analysis or to choose a best treatment (Lowry, 1992).
According to experiment and result of analysis of variance, the most suitable method is chosen. But, some researchers offer different suggestions. For example, according to Chew (1976), Scheffe test is suitable for all positions and than Tukey, SNK and Duncan tests are better than others, however Gill (1980), recommended the Tukey test.
For some of the multiple comparison tests (Unprotected LSD, SNK, Duncan, Tukey and Dunnett tests) are not necessary the rejection of H_0 hypothesis.
Bek and Efe (1995), explained the LSD test is the mostly prefer test, because of easy application, compare some of means from all means and can be obtain smaller differences are significant. But, by increasing compared means, error rate is increased so that some other alternative methods are developed. The unprotected LSD procedure often has maximum power among the procedures. LSD test is the most appropriate for pairwise treatment comparisons.
Duncan is common method, but its disadvantage like LSD is error rate is increased by increasing the number of compared means.
In SNK test, number of significant mean difference is more than Tukey test, whereas less then Duncan test (Bek and Efe, 1995). Student-Newman-Keuls and Protected LSD procedures were closer to the nominal experimentwise error rate and would be preferred for pairwise analysis (Schwertman and Carter, 1995). Scheffe and Tukey tests are multiple comparison tests, which can be used confidence interval of means.
When the number of comparisons is similar in size into the number of treatment groups the Bonferroni procedure is more powerful than the Scheffe procedure (Schwertman and Carter, 1995). The Bonferroni procedure is the recommended for nonorthogonal contrasts, because it splits the Type I error rate equally among all comparisons (Gill, 1973).
If there are a lot of means and researcher want to obtain smaller differences are significant, in this case, must be used Duncan, if researcher want to obtain only bigger differences must be used SNK, Tukey or Scheffe tests (Yildiz and Bircan, 1994).
Dunnett and William-t tests are used to compare between control group and other treatments.
If factors are qualitative and prior knowledge is available orthogonal contrasts is suitable. Procedures for planned contrasts often partition over all treatment effects into sets of preplanned, independent, individual comparisons. The researcher should strive the choose experimental treatments such that all biologically meaningful comparisons are orthogonal (Lowry, 1992).
If factors are quantitative (doses and doses ratio are equal), in this case orthogonal polynomials are used. The regression approach, using orthogonal polynomials, assesses relationship among treatments and response when treatments are graded levels of quantitative factors.
If numbers of replications are not equal in each group, Scheffe test is better than Tukey test. But, for paired comparisons it is better to use Tukey test instead of Scheffe test (Bek and Efe, 1995).
Multiple comparison tests are calculated differently according to equal or not equal number of replications. From these tests only SNK test can be used only number of replications are equal. Also tests except SNK and William-t tests can be used

condition of heterogeneous variance.

Lehmann and Shaffer (1977) are strongly recommend against Duncan's multiple range test, because of the derivation of error rates in Duncan's test depends on a monotonicity condition (consistently increasing or consistently decreasing and not oscillating) that does not necessarily occur with real data.

Gill (1980) gave some recommendations to the researchers: 1). Use designed contrasts (preferably orthogonal) to maximize sensitivity. 2). Do not compare all pairs of means if treatments are quantitative or otherwise structurally related. 3). If one must resort to comparing all pairs of means, use Tukey (HSD) test or SNK test to protect Type I error rate. Do not use the LSD test or Duncan test, both of which badly distort the calibration of strength of evidence.

As a result according to these properties, mostly suitable method is chosen by researcher the appropriate use of multiple comparison procedures assists animal science researches in making correct, useable conclusions and interpretations of them. We can suggested to researcher choosing the most appropriate multiple comparisons procedure should be take into consideration to experiment's structure, whether the experiment preplanned or not and the structure of variables (qualitative or quantitative). If there is no a specific rule, we can suggest the using of confidence interval.

References

Bancroft, T. A., 1968. Topics in Intermediate Statistical Methods. Volume I. Iowa State University Press, USA.

Bek, Y. and E. Efe, 1995. Arastirma ve Deneme Metodlari I. Çukurova Üniversitesi Ziraat Fakültesi Ders Kitabi, Yayin No: 71, Adana.

Chew, V., 1976. Comparing Treatment Means: A Compendium. Hortscience, 11: 348.

Düzgünes, O., 1963. Arastirmalarda Istatistik Prensipleri ve Metodlari. Ege Üniversitesi Matbaasi, Izmir.

Gill, J. L., 1971. Analysis of Data with Heterogeneous Variance: A Review. J. Dairy Sci., 54: 369.

Gill, J. L., 1973. Current Status of Multiple Comparisons of Means in Designed Experiments. J. Dairy Sci., Vol: 56, No: 8.

Gill, J. L., 1980. Design and Analysis of Experiments: In the Animal and Medical Science. I-II-III. Iowa State Univ. Press, Iowa, USA.

Hochberg, Y., 1976. A Modification of the T-Method of Multiple Comparisons for One-Way Layout with Unequal Variances. J. Amer. Stat. Assoc., 71: 200.

Ikiz, F., H. Püskülcü and S. Eren, 1996. Istatistige Giris (IV. Baski). Ege Üniversitesi Basimevi, Bornova, Izmir.

Ikiz, F., 1997. Deneme Planlama ve Degerlendirme. Ders Notlari, Ege Universitesi, Muhendislik Fakultesi, Izmir.

Lehmann, E. L. and J. P. Schaffer, 1977. On a Fundamental Theorem in Multiple Comparisons. J. Am. Stat. Assoc., 72: 576.

Lowry, S. R., 1992. Use and Misuse of Multiple Comparisons in Animal Experiments. J. Anim. Sci., 70:1971.

Scheffe, H., 1959. The Analysis of Variance. Willey, New York.

Schwertman, N. C. and N. J. Carter, 1995. A More Practical Scheffe - Type Multiple Comparison procedure for Commonly Encountered Numbers of Comparisons. J. Statist. Comput. Simul., 53: 181.

Sokal, R. R., and M. R., Rohlf, 1969. Biometry, Freeman, San Francisco.

Spjotvell, E. and M. R. Stoline, 1973. Extension of the Method of Multiple Comparisons to Include Cases with Unequal Sample Sizes. J. Amer. Stat. Assoc., 68: 975.

Steel, R. G. D. and J. H. Torrie, 1980. Principles and Procedures of Statistics. A Biometrical Approach (2nd Ed). McGraw-Hill Book Co, New York.

Tukey, J. W., 1953. The Problem of Multiple Comparisons. Unpublished notes, Princeton Univ., Princeton, New Jersey.

Waller, R. A. and D. B. Duncan, 1969. A Bayes Rule for the Symmetric Multiple Comparison Problem. Journal of the American Statistical Association, 64, 1484. (Corrigendum 1972), 67: 253.

Yildiz, N. and H. Bircan, 1994. Arastirma ve Deneme Metodlari. Atatürk Üniversitesi Ziraat Fakültesi Ders Kitabi, Yayin No: 697, Erzurum.