MULTIPLE COMPARISON PROCEDURES

Rajender Parsad
I.A.S.R.I., Library Avenue, New Delhi – 110012
rajender@iasri.res.in

1. Introduction

Analysis of variance is used to test for the real treatment differences. When the null hypothesis that all the treatment means are equal is not rejected, it may seem that no further questions need to be asked. However, in some experimental situations, it may be an oversimplification of the problem. For example, consider an experiment in rice weed control with 15 treatments *viz.* 4 with hand weeding, 10 with herbicides and 1 with no weeding (control). The probable questions that may be raised and the specific mean comparisons that can provide their answers may be:

i) Is any treatment effective in controlling the weeds? This question may be answered simply by comparing the mean of the control treatment with the mean of each of the 14 weed-control treatments.

ii) Is there any difference between the group of hand-weeding treatments and the group of herbicide treatments? The comparison of the combined mean of the four hand weeding treatment effects with the combined mean of the 10-herbicide treatment effects may be able to answer the above question.

iii) Are there differences between the 4 hand weeding treatments? To answer this question one should test the significant differences among the 4 hand weeding treatments. Similar question can be raised about the 10-herbicide treatments and can be answered in the above fashion.

This illustrates the diversity in the types of treatment effects comparisons. Broadly speaking, these comparisons can be classified either as *Pair Comparison* or *Group Comparison.* In *pair comparisons*, we compare the treatment effects pairwise whereas in-*group comparisons*, the comparisons could be *between group comparisons, within group comparisons, trend comparisons, and factorial comparisons.* In the above example, question (i) is the example of the pair comparisons and question (ii) illustrates the between group comparison and question (iii) is within group comparison. Through trend comparisons, we can test the functional relationship (linear, quadratic, cubic, etc.) between treatment means and treatment levels using orthogonal polynomials. Factorial comparisons are related to the testing of means of levels of a factor averaged over levels of all other factors or average of treatment combinations of some factors averaged over all levels of other factors. For pairwise treatment comparisons there are many test procedures, however, for the group comparisons, the most commonly used test procedure is to partition the treatment sum of squares into meaningful comparisons. This can be done through contrast analysis either using single degrees of freedom contrasts or contrasts with multiple degrees of freedom.

Further, the comparisons can be divided into two categories viz. *planned comparisons* and *unplanned comparisons or data snooping*. These have the following meanings. Before the experiment commences, the experimenter will have written out a checklist, highlighting the comparisons or contrasts that are of special interest, and designed the experiment in such a

way as to ensure that these are estimable with as small variances as possible. These are the *planned comparisons*. After the data have been collected, the experimenter usually looks carefully at the data to see whether anything unexpected has occurred. One or more unplanned contrasts may turn out to be the most interesting, and the conclusions of the experiment may not be anticipated. Allowing the data to suggest additional interesting contrasts is called *data snooping*.

The most useful analysis of experimental data involves the calculation of a number of different confidence intervals, one for each of several contrasts or treatment means. The confidence level for a single confidence interval is based on the probability, that the random interval will be "correct" (meaning that the random interval will contain the true value of the contrast or function).

It is shown below that when several confidence intervals are calculated, the probability that they are all simultaneously correct can be alarmingly small. Similarly, when several hypotheses are tested, the probability that at least one hypothesis is incorrectly rejected can be uncomfortably high. Much research has been done over the years to find ways around these problems. The resulting techniques are known as *methods of multiple comparison*, the intervals are called *simultaneous confidence intervals*, and the tests are called *simultaneous hypothesis test*.

Suppose an experimenter wishes to calculate *m* confidence intervals, each having a $100(1-\alpha^*)\%$ confidence level. Then each interval will be individually correct with probability $1-\alpha^*$. Let S_j be the event that the j^{th} confidence interval will be correct and $\bar{S}_j$ the event that it will be incorrect $(j=1,...,m)$. Then, using the standard rules for probabilities of unions and intersections of events, it follows that

$$P(S_1 \cap S_2 \cap ... \cap S_m) = 1 - P(\bar{S}_1 \cup \bar{S}_2 \cup ... \cup \bar{S}_m).$$

This says that the probability that all of the intervals will be correct is equal to one minus the probability that at least one will be incorrect. If $m = 2$,

$$\begin{aligned} P(\bar{S}_1 \cup \bar{S}_2) &= P(\bar{S}_1) + P(\bar{S}_2) - P(\bar{S}_1 \cap \bar{S}_2) \\ &\le P(\bar{S}_1) + P(\bar{S}_2) \end{aligned}$$

A similar result, which can be proved by mathematical induction, holds for any number *m* of events, that is,

$$P(\bar{S}_1 \cup \bar{S}_2 \cup ... \cup \bar{S}_m) \le \sum_j P(\bar{S}_j),$$

with equality if the events $\bar{S}_1, \bar{S}_2, ..., \bar{S}_m$ are mutually exclusive. Consequently,

$$P(S_1 \cap S_2 \cap ... \cap S_m) \ge 1 - \sum_j P(\bar{S}_j) = 1 - m\alpha^* \qquad (1)$$

that is, the probability that the *m* intervals will simultaneously be correct is at least $1 - m\alpha^*$. The probability $m\alpha^*$ is called the *overall significance level* or *experiment wise error rate* or

family error rate. A typical value for α^* for a single confidence interval is 0.05, so the probability that six confidence intervals each calculated at a 95% individual confidence level will simultaneously be correct is at least 0.7. Although "at least" means "bigger than or equal to", it is not known in practice how much bigger than 0.7 the probability might actually be. This is because the degree of overlap between the events $\bar{S}_1, \bar{S}_2, ..., \bar{S}_m$ is generally unknown. The probability "at least 0.7" translates into an overall confidence level of "at least 70%" when the responses are observed. Similarly, if an experimenter calculates ten confidence intervals each having individual confidence level 95%, then the simultaneous confidence level for the ten intervals is at least 50%, which is not very informative. As *m* becomes larger the problem becomes worse, and when $m \geq 20$, the overall confidence level is at least 0%, clearly a useless assertion!

Similar comments apply to the hypothesis-testing situation. If *m* hypotheses are to be tested, each at significance level α^*, then the probability that at least one hypothesis is incorrectly rejected is at most $m\alpha^*$.

Various methods have been developed to ensure that the overall confidence level is not too small and the overall significance level is not too high. Some methods are completely general, that is, they can be used for any set of estimable functions, while others have been developed for very specialized purposes such as comparing each treatment with a control. Which method is best depends on which contrasts are of interest and the number of contrasts to be investigated. Some of these methods can also be used for identifying the homogeneous subsets of treatment effects. Such procedures are called as multiple range tests. Several methods are discussed in the sequel some of them control the overall confidence level and overall significance level.

Following the lecture notes on Fundamentals of Design of Experiments, let $\sum l_i t_i$ denote a treatment contrast, $\sum_i l_i = 0$ where t_i is the effect of treatment *i*. The (BLUE) and the confidence interval for the above contrast can be obtained as per procedure given in the aforementioned lecture notes. However, besides obtaining confidence intervals one may be interested in hypothesis testing. The outcome of a hypothesis test can be deduced from the corresponding confidence interval in the following way. The null hypothesis $H_0 : \sum_i l_i t_i = h$ will be rejected at significance level α in favour of the two-sided alternative hypothesis $H_1 : \sum_i l_i t_i \neq h$ if the corresponding confidence interval for $\sum_i l_i t_i$ fails to contain *h*.

In the following section, we discuss the confidence intervals and hypothesis tests based on several methods of multiple comparisons. A shorter confidence interval corresponds to a more powerful hypothesis test.

2. Multiple Comparison Procedures

The terminology "a set of simultaneous $100(1-\alpha^*)\%$ confidence intervals" will always refer to the fact that the overall confidence level for a set of contrasts or treatments means is (*at least*) $100(1-\alpha^*)\%$. Each of the methods discussed gives confidence intervals of the form

$$\sum_i l_i t_i \in \left(\sum_i l_i \hat{t}_i \pm w\sqrt{\hat{Var}\left(\sum l_i \hat{t}_i\right)} \right) \qquad (2)$$

where w,which we call the *critical coefficient*, depends on the method, the number of treatments v, on the number of confidence intervals calculated, and on the number of error degrees of freedom. The term

$$msd = w\sqrt{\hat{Var}\left(\sum l_i \hat{t}_i\right)} \qquad (3)$$

which is added and subtracted from the least square estimate in (2) is called the *minimum significant difference*, because if the estimate is larger than *msd*, the confidence interval excludes zero, and the contrast is significantly different from zero.

2.1 The Least Significant Difference (LSD) Method

Suppose that, following an analysis of variance *F* test where the null hypothesis is rejected, we wish to test $H_0 : \sum_i l_i t_i = 0$ against the alternative hypothesis $H_1 : \sum_i l_i t_i \neq 0$. For making pairwise comparisons, consider the contrasts of the type $t_i - t_j$ in which experimenters are often interested, are obtainable from $\sum_i l_i t_i$ by putting $l_i = 1, l_j = -1$ and zero for the other l's. The $100(1-\alpha)\%$ confidence interval for this contrast is

$$\sum l_i t_i \in \left(\sum l_i \hat{t}_i \pm t_{edf,\alpha/2}\sqrt{\hat{Var}\left(\sum l_i \hat{t}_i\right)} \right) \qquad (4)$$

where *edf* denotes the error degrees of freedom. As we know that the outcome of a hypothesis test can be deduced from the corresponding confidence interval in the following way. The null hypothesis will be rejected at significance level α in favour of the two-sided alternative hypothesis if the corresponding confidence interval for $\sum_i l_i t_i$ fails to contain *0*. The interval fails to contain *0* if the absolute value of $\sum l_i \hat{t}_i$ is bigger than $t_{edf,\alpha/2}\sqrt{\hat{Var}\left(\sum l_i \hat{t}_i\right)}$. The critical difference or the least significant difference for testing the significance of the difference of two treatment effects, say $t_i - t_j$ is $lsd = t_{edf,\alpha/2}\sqrt{\hat{Var}\left(\sum l_i \hat{t}_i\right)}$, where $t_{edf,\alpha/2}$ is the value of Student's *t* at the level of significance α and error degree of freedom. If the difference of any two-treatment means is greater than the *lsd* value, the corresponding treatment effects are significantly different.

The above formula is quite general and particular cases can be obtained for different experimental designs. For example, the least significant difference between two treatment effects for a randomized complete block (RCB) design, with *v* treatments and *r* replications is

$t_{(v-1)(r-1),\alpha/2}\sqrt{2MSE/r}$, where $t_{(v-1)(r-1),\alpha/2}$ is the value of Student's *t* at the level of significance α and degree of freedom *(v - 1)(r - 1).* For a completely randomized design with *v* treatments such that i^{th} treatment is replicated r_i times and $\sum_{i=1}^{v} r_i = n$, the total number of experimental units, the least significant difference between two treatment effects is $t_{(n-v),\alpha/2}\sqrt{MSE\left(\frac{1}{r_i}+\frac{1}{r_j}\right)}$.

It may be worthwhile mentioning here that the least significant difference method is suitable only for planned pair comparisons. This test is based on individual error rate. However, for those who wish to use it for all possible pairwise comparisons, should apply only after the *F* test in the analysis of variance is significant at desired level of significance. This procedure is often referred as *Fisher's protected lsd*. Since *F* calls for us to accept or reject a hypothesis simultaneously involving means. If we arrange the treatment means in ascending or descending order of their magnitude and keep the, means in one group for which the difference between the smallest and largest mean is less than the *lsd*, we can identify the homogeneous subsets of treatments. For example, consider an experiment that was conducted in completely randomized design to compare the five treatments and each treatment was replicated 5 times. *F*-test rejects the null hypothesis regarding the equality of treatment means. The mean square error (MSE) is 8.06. The means of five treatments tried in experiment are 9.8, 15.4, 17.6, 21.6 and 10.8 respectively. The *lsd* for the above comparisons is 3.75, then the homogeneous subsets of treatments are Group1: Treatment 1 and 5, group 2: treatment 2 and 3 and group 3: treatment 4. Treatments within the same homogeneous subset are identified with the same alphabet in the output from SAS.

2.2 Duncan's Multiple Range Test

A widely used procedure for comparing all pairs of means is the multiple range test developed by Duncan (1955). The application of Duncan's multiple range test (*DMRT*) is similar to that of *lsd* test. *DMRT* involves the computation of numerical boundaries that allow for the classification of the difference between any two treatment means as significant or non-significant. *DMRT* requires computation of a series of values each corresponding to a specific set of pair comparisons unlike a single value for all pairwise comparisons in case of *lsd.* It primarily depends on the standard error of the mean difference as in case of *lsd.* This can easily be worked out using the estimate of variance of an estimated elementary treatment contrast through the design.

For application of the *DMRT* rank all the treatment means in decreasing or increasing order based on the preference of the character under study. For example for the yield data, the rank 1 is given to the treatment with highest yield and for the pest incidence the treatment with the least infestation should get the rank as 1. Consider the same example as in case of *lsd.* The ranks of the treatments are given below:

Treatments	T1	T5	T2	T3	T4
Treatment Means	9.8	10.8	15.4	17.8	21.6
Rank	1	2	3	4	5

Compute the standard error of the difference of means (SE_d) that is same as that of square root of the estimate of the variance of the estimated elementary contrast through the design. In the present example this is given by $\sqrt{2(8.06)/5} = 1.796$. Now obtain the value of the least significant range $R_p = \frac{r_\alpha(p, edf) * SE_d}{\sqrt{2}}$, where α is the desired significance level, *edf* is the error degrees of freedom and $p = 2, \ldots, v$ is one more than the distance in rank between the pairs of the treatment means to be compared. If the two treatment means have consecutive rankings, then $p = 2$ and for the highest and lowest means it is v. The values of $r_\alpha(p, edf)$ can be obtained from Duncan's table of significant ranges.

For the above example the values of $r_\alpha(p, edf)$ at 20 degrees of freedom and 5% level of significance are $r_{0.05}(2,20) = 2.95$, $r_{0.05}(3,20) = 3.10$, $r_{0.05}(4,20) = 3.18$ and $r_{0.05}(5,20) = 3.25$. Now the least significant ranges R_p are

R_2	R_3	R_4	R_5
3.75	3.94	4.04	4.13

Then, the observed differences between means are tested, beginning with largest versus smallest, which would be compared with the least significant range R_v. Next, the difference of the largest and the second smallest is computed and compared with the least significant range R_{v-1}. These comparisons are continued until all means have been compared with the largest mean. Finally, the difference of the second largest mean and the smallest is computed and compared against the least significant range R_{v-1}. This process is continued until the differences of all possible $v(v-1)/2$ pairs of means have been considered. If an observed difference is greater than the corresponding least significant range, then we conclude that the pair of means in question is significantly different. To prevent contradictions, no differences between a pair of means are considered significant if the two means involved fall between two other means that do not differ significantly. For our case the comparisons will yield

4 vs 1: 21.6 - 9.8 = 11.8 > 4.13(R_5);

4 vs 5: 21.6 - 10.8 = 10.8 > 4.04(R_4);

4 vs 2: 21.6 - 15.4 = 6.2 > 3.94(R_3);

4 vs 3: 21.6 - 17.6 = 4.0 > 3.75(R_2);

3 vs 1: 17.6 - 9.8 = 7.8 > 4.04(R_4);

3 vs 5: 17.6 - 10.8 = 6.8 > 3.94(R_3);

3 vs 2: 17.6 - 15.4 = 2.2 < 3.75(R_2);

2 vs 1: 15.4 - 9.8 = 5.6 > 3.94(R_3);

2 vs 5: 15.4 - 10.8 = 4.6 > 3.75(R_2);

4 vs 1: 10.8 - 9.8 = 1.0 < 3.75(R_2);

We see that there are significant differences between all pairs of treatments except T3 and T2 and T5 and T1. A graph underlining those means that are not significantly different is shown below.

T1	T5	T2	T3	T4
9.8	10.8	15.4	17.8	21.6

It can easily be seen that the confidence intervals of the desired pairwise comparisons following (2) is

$$\sum_i l_i t_i \in \left(\sum_i l_i \hat{t}_i \pm \frac{r_\alpha(p, edf)}{\sqrt{2}} \sqrt{\hat{Var}\left(\sum_i l_i \hat{t}_i \right)} \right) \tag{5}$$

and least significant range in general is

$$lsr = \frac{r_\alpha(p, edf)}{\sqrt{2}} \sqrt{\hat{Var}\left(\sum_i l_i \hat{t}_i \right)}.$$

The methods of multiple comparison given in Sections 2.1 and 2.2 uses individual error rates (probability that a given confidence interval will not contain the true difference in level means). This may be misleading as is clear from inequality (1), i.e., if *m* simultaneous confidence intervals are calculated for preplanned contrasts, and if each confidence interval has confidence level $100(1-\alpha^*)\%$ then the overall confidence level is greater than or equal to $100(1-m\alpha^*)\%$. Therefore, the methods of multiple comparisons that utilizes experiment wise error rate or family error rate (Maximum probability of obtaining one or more confidence intervals that do not contain the true difference between level means) may be quite useful. In the sequel, we describe some methods of multiple comparisons that are based on family error rates.

2.3 Bonferroni method for preplanned comparisons

In this method the overall confidence level of $100(1-\alpha^*)\%$ for *m* simultaneous confidence intervals can be ensured by setting $\alpha^* = \alpha / m$. Replacing α by α / m in the formula (4) for an individual confidence interval, we obtain a formula for a set of simultaneous $100(1-\alpha^*)\%$ confidence intervals for *m* preplanned contrasts $\sum l_i t_i$ is

$$\sum_i l_i t_i \in \left(\sum_i l_i \hat{t}_i \pm t_{edf, \alpha/2m} \sqrt{\hat{Var}\left(\sum_i l_i \hat{t}_i \right)} \right) \tag{6}$$

Therefore, if the contrast estimate is greater than $t_{edf,\alpha/2m} \sqrt{\hat{Var}\left(\sum_i l_i \hat{t}_i \right)}$ the corresponding contrast is significantly different from zero.

It can easily be seen that this method is same as that of least significant difference with α in least significant difference to be replaced by α / m. Since $\alpha/(2m)$ is likely to be a typical value, the percentiles $t_{edf,\alpha/(2m)}$ may need to be obtained by use of a computer package.

When m is very large, $\alpha/(2m)$ is very small, possibly resulting in extremely wide simultaneous confidence intervals. In this case the Scheffe or Turkey methods described in the sequel would be preferred.

Note that this method can be used only for *preplanned* contrasts or any m preplanned estimable contrasts or functions of the parameters. It gives shorter confidence intervals than the other methods listed here if m is small. It can be used for any design. However, it cannot be used for data snooping. An experimenter who looks at the data and then proceeds to calculate simultaneous confidence intervals for the few contrasts that look interesting has effectively calculated a very large number of intervals. This is because the interesting contrasts are usually those that seem to be significantly different from zero, and a rough mental calculation of the estimates of a large number of contrasts has to be done to identify these interesting contrasts. Scheffe's method should be used for contrasts that were selected after the data were examined.

2.4 Scheffe Method of Multiple Comparisons

In the Bonferroni method of multiple comparisons, the major problem is that the m contrasts to be examined must be preplanned and the confidence intervals can become very wide if m is large. Scheffe's method, on the other hand, provides a set of simultaneous $100(1-\alpha^*)\%$ confidence intervals whose widths are determined only by the number of treatments and the number of observations in the experiment. It is not dependent on the number of contrasts are of interest. It utilizes the fact that every possible contrast $\sum l_i t_i$ can be written as a linear combination of the set of $(v-1)$ treatment - versus - control contrasts, $t_2 - t_1, t_3 - t_1, ..., t_v - t_1$. Once the experimental data have been collected, it is possible to find a $100(1-\alpha^*)\%$ *confidence region* for these $v-1$ treatment - versus - control contrasts. The confidence region not only determines confidence bounds for each treatment - versus - control contrasts, it determines bounds for *every* possible contrast $\sum l_i t_i$ and, in fact, for *any number* of contrasts, while the overall confidence level remains fixed. For mathematical details, one my refer to Scheffe (1959) and Dean and Voss (1999). Simultaneous confidence intervals for all the contrasts $\sum l_i t_i$ can be obtained from the general formula (2) by replacing the critical coefficient w by w_s where $w_s = \sqrt{aF_{a,edf,\alpha}}$ with a as the dimension of the space of linear estimable functions being considered, or equivalently, a is the number of degrees of freedom associated with the linear estimable functions being considered.

The Scheffe's method applies to any m estimable contrasts or functions of the parameters. It gives shorter intervals than Bonferroni method when m is large and allows data snooping. It can be used for any design.

2.5 Tukey Method for All Pairwise Comparisons

Tukey (1953) proposed a method for making all possible pairwise treatment comparisons. The test compares the difference between each pair of treatment effects with appropriate adjustment for multiple testing. This test is also known as Tukey's honestly significant difference test or Tukey's HSD. The confidence intervals obtained using this method are shorter than those obtained from Bonferroni and Scheffe methods. Following the formula (2),

one can obtain the simultaneous confidence intervals for all the contrasts of the type $\sum l_i t_i$ by replacing the critical coefficient w by $w_t = q_{v,edf,\,\alpha} / \sqrt{2}$ where v is the number of treatments and edf is the error degree of freedom and values can be seen as the percentile corresponding to a probability level α in the right hand tail of the studentized range distribution tables.

For the completely randomized design or the one-way analysis of variance model, $\hat{Var}(\hat{t}_i - \hat{t}_j) = \sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$, where r_i denotes the replication number of treatment i $(i = 1,2,\ldots,v)$. Then Tukey's simultaneous confidence intervals for all pairwise comparisons $t_i - t_j$, $i \neq j$ with overall confidence level at least $100(1-\alpha^*)\%$ is obtained by taking $w_t = q_{v,\,n\text{-}v,\,\alpha} / \sqrt{2}$ and $\hat{Var}(\hat{t}_i - \hat{t}_j) = \sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$. The values of $q_{v,n-v,\alpha}$ can be seen the studentized range distribution tables.

When the sample sizes are equal $(r_i = r; i = 1,...,v)$, the overall confidence level is *exactly* $100(1-\alpha^*)\%$. When the sample sizes are unequal, the confidence level is *at least* $100(1-\alpha^*)\%$.

It may be mentioned here that Tukey's method is the best for all pairwise treatment comparisons. It can be used for completely randomized designs, randomized complete block designs and balanced incomplete block designs. It is believed to be applicable (conservative, true α level lower than stated) for other incomplete block designs as well, but this has not yet been proven. It can be extended to include all contrasts but Scheffe's method is generally better for these types of contrasts.

2.6 Dunnett Method for Treatment-Versus-Control Comparisons

Dunnett (1955) developed a method of multiple comparisons for obtaining a set of simultaneous confidence intervals for preplanned treatment-versus-control contrasts $t_i - t_1 (i = 2,..., v)$ where level 1 corresponds to the control treatment. The intervals are shorter than those given by the Scheffe, Tukey and Bonferroni methods, but the method should not be used for any other type of contrasts. For details on this method, a reference may be made to Dunnett (1955, 1964) and Hochberg and Tamhane (1987). In general this procedure is, therefore, best for all treatment-versus-control comparisons. It can be used for completely randomized designs, randomized complete block designs. It can also be used for balanced incomplete block designs but not in other incomplete block designs without modifications to the corresponding multivariate t-distribution tables given in Hochberg and Tamhane (1987).

However, not much literature is available for multiple comparison procedures for making simultaneous confidence statement about several test treatments with several control

treatments comparisons. A partial solution to the above problem has been given by Hoover (1991).

2.7 Hsu Method for Multiple Comparisons with the Best Treatment

"Multiple comparisons with the best treatment" is similar to multiple comparisons with a control, except that since it is unknown prior to the experiment which treatment is the best, a control treatment has not been designated. Hsu (1984) developed a method in which each treatment sample mean is compared with the best of the others, allowing some treatments to be eliminated as worse than best, and allowing one treatment to be identified as best if all others are eliminated. Hsu calls this method RSMCB, which stands for *Ranking, Selection and Multiple Comparisons with the Best treatment.*

Suppose, first, that the best treatment is the treatment that gives the largest response on average. Let $t_i - max(t_j)$ denote the effect of the i^{th} treatment minus the effect of the best of the *other* $v-1$ treatments. When the i^{th} treatment is the best, $max(t_j)\ (j \neq i)$ will be the effect of the second-best treatment. So, $t_i - max(t_j)$ will be positive if treatment i is the best, zero if the i^{th} treatment is tied for being the best, or negative if the treatment i is worse than best.

If the best treatment factor level is the level that gives the smallest response rather than the largest, then Hsu's procedure has to be modified by taking $t_i - min(t_j)$ in place of $t_i - max(t_j)$.

To summarize, Hsu's method for multiple comparisons selects the best treatment and identifies those treatments that are significantly worse than the best. It can be used for completely randomized designs, randomized block designs and balanced incomplete block designs. For using it in other incomplete block designs, modifications of the tables is required.

3. Multiple Comparison Procedures using SAS/SPSS/MINITAB

The MEANS statement in PROC GLM or PROC ANOVA can be used to generate the observed means of each level of a treatment factor. The TUKEY, BON, SCHEFFE, LSD, DUNCAN, etc. options under MEANS statement causes the SAS to use Tukey, Bonferroni, Scheffe's, least significant difference, Duncan's Multiple Range Test methods to compare the effects of each pair of levels. The option CLDIFF asks the results of above methods be presented in the form of confidence intervals. The option DUNNETT ('1') requests Dunnett's 2-sided method of comparing all treatments with a control, specifying level '1' as the control treatment. Similarly the options DUNNETTL ('1') and DUNNETTU ('1') can be used for right hand and left hand method of comparing all treatments with a control. The pairwise comparisons of treatments in unbalanced data can be performed using LSMEANS statement of SAS.

To Specify Post Hoc Tests for GLM Procedures in SPSS: From the menus choose:

Analyze→ General Linear Model → From the menu, choose Univariate, Multivariate, or Repeated Measures → In the dialog box, click Post Hoc → Select the factors to analyze and move them to the Post Hoc Tests For list → Select the desired tests. Please note that Post hoc

tests are not available when covariates have been specified in the model. GLM Multivariate and GLM Repeated Measures are available only if you have the Advanced Models option installed. For obtaining adjusted treatment means and performing pairwise comparisons of effects, choose options and estimated marginal means.

Multiple comparison procedures available in MINITAB are under two different options viz. for one-way ANOVA and General Linear Model. The multiple comparison procedure available under one-way ANOVA are: Tukey's, Fisher's, Dunnett's, and Hsu's MCB. Tukey and Fisher ‘s methods may be used for obtaining confidence intervals for all pairwise treatment differences. Dunnett’s method is to be used for obtaining a confidence interval for the difference between each treatment mean and a control mean. Hsu's MCB provides a confidence interval for the difference between each level mean and the best of the other level means. Tukey, Dunnett and Hsu's MCB tests use a family error rate, whereas Fisher's LSD procedure uses an individual error rate. In one-way analysis of variance, the individual error rate displayed is exact in all cases, including equal and unequal sample sizes. In one-way analysis of variance, the family error rate displayed is exact for equal sample sizes. If levels have unequal sample sizes, the true family error rate for Tukey, Fisher, and MCB are slightly smaller than stated, resulting in conservative confidence intervals. The Dunnett family error rates are exact for unequal sample sizes.

One can choose out of the following four multiple comparison procedures under option General Linear Model viz. Tukey, Dunnett, Bonferroni or Sidak. The Tukey (also called Tukey-Kramer in the unbalanced case) and Dunnett methods are extensions of the methods used by one-way ANOVA. Except Dunnett’s method, all the three methods are for making simultaneous tests for all possible pairwise treatment comparisons. The Tukey approximation has not proven to be conservative for comparing more than three means. Bonferroni and Sidak methods are conservative methods based upon probability inequalities. The Sidak method is slightly less conservative than the Bonferroni method. Here, Conservative" means that the true error rate is less than the stated one. Following choices are available for multiple comparisons:

– Pairwise comparisons or comparisons with a control
- Which means to compare
- The method of comparison
- Display comparisons in confidence interval or hypothesis test form
- The confidence level, if you choose to display confidence intervals
- The alternative, if you choose comparisons with a control

To specify multiple comparison procedures in MINITAB:
From the menus choose: Stat→ANOVA→One way or General Linear Model→In the Dialog Box click on Comparisons→Select the desired Multiple Comparison Procedure. In General Linear Model, Select the comparisons all pairwise comparisons or comparisons with a control, then the terms on which comparisons are required → Check on the Multiple Comparison Procedure required and Check **Test** for multiple comparison output.

4. Conclusions

Each of the methods of multiple comparisons at subsections 2.3 to 2.7 allows the experimenter to control the overall confidence level, and the same methods can be used to

control the experiment wise error rate when multiple hypotheses are to be tested. There exist other multiple comparison procedures that are more powerful (*i.e.* that more easily detect a nonzero contrast) but do not control the overall confidence level nor the experiment wise error rate. While some of these are used quite commonly, however, we don't advocate their use.

The selection of the appropriate multiple comparison method depends on the desired inference. As discussed in Section 3 that for making all possible pairwise treatment comparisons, the Tukey's method is not conservative and gives smaller confidence intervals as compared to Bonferroni, Sidak and Scheffe's methods. Therefore, one may choose Tukey's method for making all possible pairwise comparisons.

For more details on methods of multiple comparisons, one may refer to Steel and Torrie (1981), Gomez and Gomez (1984) and Montgomery (1991), Hsu (1996), Dean and Voss (1999).

Acknowledgements: The discussion held with Professor Daniel Voss through E-mail during the preparation of this lecture is gratefully acknowledged.

References

Dean,A. and Voss,D.(1999). *Design and Analysis of Experiments*. Springer Texts in Statistics, Springer, New York. 67-97.

Duncan,D.B. (1955). Multiple range and multiple F-Tests. *Biometrics*, **11**, 1-42.

Dunnett,C.W.(1955). A multiple comparisons procedure for comparing several treatments with a control. *Jour. Am. Statist. Assoc*., **50**, 1096-1121.

Dunnett,C.W.(1964). New tables for multiple comparisons with a control. *Biometrics*, **20**, 482-491.

Gomez,K.A. and Gomez, A.A. (1984). *Statistical Procedures for Agricultural Research, 2nd Edition*. John Wiley and Sons, New York.

Hayter,A.J. (1984). A proof of the conjecture that the Tukey-Cramer multiple comparison procedure is conservative. *Ann. Statist.*, **12**, 61-75.

Hochberg,Y. and Tamhane,A.C.(1987). *Multiple Comparison Procedures.* John Wiley and Sons, New York.

Hoover,D.R.(1991). Simultaneous comparisons of multiple treatments to two (or more) controls. *Biom. J.*, **33**, 913-921.

Hsu,J.C. (1984). Ranking and selection and multiple comparisons with the best. *Design of Experiments: Ranking and Selection (Essays in Honour of Robert E.Bechhofer)*. Editors: T.J.Santner and A.C.Tamhane. 23-33, Marcel Dekker, New York.

Hsu,J.C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall. London.

Montgomery,D.C.(1991). *Design and Analysis of Experiments, 3rd edition*. John Wiley & Sons. New York. 67-79.

Peiser,A.M. (1943). Asymptotic formulas for significance levels of certain distributions. *Ann. Math. Statist.*, **14**, 56-62. (Correction 1949, *Ann.Math. Statist*., **20**, 128-129).

Scheffe,H.(1959). *The Analysis of Variance*. John Wiley & Sons. New York.

Steel, R.G.D. and Torrie, J.H.(1981). *Principles and procedures of statistics: A biometrical approach*. McGraw-Hill Book Company, Singapore, 172-194.

Turkey,J.W.(1953). The problem of multiple comparisons. *Dittoed manuscript of 396 pages*, Department of Statistics, Princeton University.