

Multiple Comparisons

Often in the context of planning an experiment or analyzing data after an experiment has been completed, we find that comparison of specific pairs or larger groups of treatment means are of greater interest than the simple question posed by an analysis of variance - do at least two treatment means differ? It may be that embedded in a group of treatments there is only one "control" treatment to which every other treatment should be compared, and comparisons among the non-control treatments may be uninteresting. One may also, after performing an analysis of variance and rejecting the null hypothesis of equality of treatment means want to know exactly which treatments or groups of treatments differ. To answer these kinds of questions requires careful consideration of the hypotheses of interest both before and after an experiment is conducted, the Type I error rate selected for each hypothesis, the power of each hypothesis test, and the Type I error rate acceptable for the group of hypotheses as a whole.

Comparisons or Contrasts

If we let \bar{T}_i represent a treatment mean and c_i a weight associated with the i th treatment mean then a comparison or contrast can be represented as:

$$c_1\bar{T}_1 + c_2\bar{T}_2 + c_3\bar{T}_3 + \dots + c_k\bar{T}_k,$$

where $\sum_{i=1}^k c_i = 0$. It can be seen that this contrast is a linear combination of treatment means (other contrasts such as quadratic and cubic are also possible). All of the following are possible comparisons:

$$\bar{T}_1 - \bar{T}_2$$

$$\bar{T}_2 - \bar{T}_3$$

$$\bar{T}_1 - \bar{T}_3$$

$$\frac{\bar{T}_1 + \bar{T}_2}{2} - \bar{T}_3$$

$$\frac{\bar{T}_2 + \bar{T}_3}{2} - \bar{T}_1$$

$$\frac{\bar{T}_1 + \bar{T}_2 + \bar{T}_3}{3} - \bar{T}_4$$

$$\frac{\bar{T}_1 + \bar{T}_2}{2} - \frac{\bar{T}_3 + \bar{T}_4}{2}$$

because they are weighted linear combinations of treatment means and the

weights sum to zero $\left(\sum_{i=1}^k c_i \right)$.

For example, previously we have performed comparisons between two treatment means using the t -statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with $(n_1 + n_2) - 2$ degrees of freedom. This statistic is a "contrast." The numerator of this expression follows the general form of the contrast outlined above with the weights c_1 and c_2 equal to 1 and -1, respectively:

$$(1) \times (\bar{x}_1) + (-1) \times (\bar{x}_2)$$

However, we also see that this contrast is divided by the pooled within cell or within group variation. So, a contrast is actually the ratio of a linear combination of weighted means to an estimate of the pooled within cell or error variation in the experiment:

$$t = \frac{(1) \times (\bar{x}_1) + (-1) \times (\bar{x}_2)}{\sqrt{\text{within cell variation}}}$$

with $\left(\sum_{i=1}^k n_i\right) - k$ degrees of freedom. For a non-directional null hypothesis t could be replaced by F :

$$F = \frac{[(1) \times (\bar{x}_1) + (-1) \times (\bar{x}_2)]}{MS_{error}}$$

with 1, and $\left(\sum_{i=1}^k n_i\right) - k$ degrees of freedom. In general, a contrast is the ratio of a linear combination of weighted means to the mean square within cells times the sum of the squares of the weights assigned to each mean divided by the sample size within cells:

$$\frac{c_1 \bar{x}_1 + c_2 \bar{x}_2 + \dots + c_k \bar{x}_k}{MS_{error} \left(\sum_{i=1}^k c_i^2 / n_i \right)}$$

where the c_i 's are the weights assigned to each treatment mean, $\sum_{i=1}^k c_i = 0$, n_i is the number of observations in each cell and MS_{error} is the within cell variation pooled from the entire experiment (the within cell mean square estimated from a variance partition). For a comparison of two treatment means $c_1 = 1$ and $c_2 = -1$, so:

$$t = \frac{(1) \times (\bar{x}_1) + (-1) \times (\bar{x}_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} (MS_{error})}$$

$n_1+n_2 -2$ degrees of freedom, or

$$F = \frac{[(1) \times (\bar{x}_1) + (-1) \times (\bar{x}_2)]^2}{\frac{(1)^2}{n_1} + \frac{(1)^2}{n_2} (MS_{error})}$$

with 1, and $\left(\sum_{i=1}^k n_i\right) - k$ degrees of freedom. More generally; where ψ^2 indicates the contrast

$$F = \frac{\psi^2}{MS_{error} \sum_{i=1}^k c_i^2 / n_i}$$

with 1, and $\left(\sum_{i=1}^k n_i\right) - k$ degrees of freedom.

The F - statistic outlined above provides a parametric test of the null hypothesis that the contrasted means are equal. Similar statistics can be elaborated for rank like non-parametric tests. Hollander and Wolfe (1973) outline several non-parametric contrast estimators.

Experiment and Comparison - Wise Error Rates

In an experiment where two or more comparisons are made from the data there are two distinct kinds of Type I error. The comparison - wise error rate is the probability of a Type I error set by the experimenter for evaluating each comparison. The experiment - wise error rate is the probability of making at least one Type I error when performing the whole set of comparisons. If we let a_c the comparison - wise error rate, a_e the experiment - wise error rate, and j the number of contrasts performed, then if the contrasts are planned in advance of the experiment (*a priori*) and done in place of the analysis of variance the relationship between a_c and a_e is given by these equations (the Dunn-Sidak correction):

$$\alpha_e = 1 - (1 - \alpha_c)^j$$

$$\alpha_c = 1 - (1 - \alpha_e)^{1/j}$$

An approximate estimate of the relationship between α_c and α_e is given by the Bonferroni correction

$$\alpha_e = j\alpha$$

$$\alpha_c = \alpha_e/j$$

As j increases the Bonferroni approximation departs markedly from the exact calculation given by the Dunn-Sidak correction. In the table below $\alpha_c = 0.05$ and the values tabulated represent estimates of α_e for various numbers of contrasts.

j	Dunn-Sidak	Bonferroni
1	0.05	0.05
2	0.0975	0.10
3	0.142625	0.15
4	0.1854	0.20
5	0.2262	0.25
10	0.40126	0.50
20	0.6415	1.0

Note that the value of α_e estimated under the Dunn-Sidak correction assumes that all contrasts performed are mutually independent. If some of the contrasts performed are dependent then the value of α_e given by the Dunn-Sidak correction will be an overestimate of α_e . Therefore, unless it is known that the set of contrasts are independent (orthogonal) then we can only provide an interval

estimate of α_e . For completely dependent contrasts $\alpha_e = \alpha_c$ for all j contrasts, so

$$\alpha_c \leq \alpha_e \leq [1 - (1 - \alpha_c)^j]$$

The above results apply for planned or *a priori* comparisons. When comparisons are performed after the data have been examined (*a posteriori*) or subjected to an analysis of variance then controlling the experiment - wise error rate requires an even larger penalty. If we let m equal the number of possible contrasts of size g then

$$\alpha_m = 1 - (1 - \alpha_c)^m$$

and α_m is said to be the family - wise error rate. For example, if an experiment consisting of $k = 5$ treatments was performed and one or more pairs of treatment means were examined after the experiment then the exponent m , the number of *possible* pairwise comparisons is $(k(k - 1))/2 = 10$. For $\alpha_c = 0.05$, α_e would be 0.40126. Had only 2 or 3 pairwise contrasts been performed *a priori* then α_e would have been much smaller. *A posteriori* contrasts involving comparing the average of 2 means to a third mean, the average of two means to the average of two other means, or other families of contrasts could also be performed. However, the experiment - wise error rate grows very rapidly since a penalty must be taken for each *possible* comparison in each family examined rather than just for the actual number of *a posteriori* comparisons made. This is because once you have looked at the results of the experiment one can snoop out the comparisons that are likely to be significantly different. One is therefore more prone to snoop out Type I errors.

Which error rate should we pay most attention to in planning and analyzing experiments? This again is a matter of judgment and must be balanced against the acceptable contrast and experiment - wise Type II error rate. Since to achieve a low experiment - wise error rate requires an even lower contrast - wise Type I error rate, the contrast - wise Type II error rate will be high. If it is more costly to the researcher to permit even one Type I error in a set of contrasts then the experiment - wise error rate should be minimized. On the other hand, if failing to detect a true treatment effect is more costly then less emphasis should be placed on minimizing the experiment - wise Type I error rate. Although no rule of thumb exists regarding an acceptable value for α_e , I recommend that the experiment - wise Type I error rate be set at 10 to 15%.

Further Reading

Jones, D. 1984. Use, misuse, and role of multiple-comparison procedures in ecological and agricultural entomology *Environmental Entomology* 13: 635-649.

Chew, V. 1976. Comparing treatment means: a compendium. *Hortscience* 11: 348-357.

Hays, W.L. 1981. *Statistics*. 3rd edition, Chapter 12. Holt, Rinehart, and Winston.

Hollander, M. and D.A. Wolfe. 1973. *Nonparametric Statistical Methods*. Wiley, New York.