Chapter 1: Multiple Comparison Procedures

1.1 Introduction

The omnibus F-test in a one-way ANOVA is a test of the null hypothesis that the population means of all *k* samples are equal. Note that rejection of this null hypothesis (when *k* > 2) does not provide us with very much detailed information. In other words, rejection of the null hypothesis does not tell us which means differ significantly from other means--apart from the fact that the smallest and largest means are different, of course! And so, if the null hypothesis is rejected, a search for which differences are significant is in order. These search techniques are known as *multiple comparison (MC) procedures*.

1.2 Fisher's Least Significant Difference (LSD) Method

You may recall from your introduction to ANOVA that there are definite problems associated with multiple *t*-tests. Specifically, whenever you carry out multiple *t*-tests (rather than one-way ANOVA) on some set of means with the alpha *per comparison* (α_{PC}) set at .05, the probability of at least one Type I error can be much greater than .05. To illustrate the point, if you carried out *c* tests (or contrasts), and if each test was independent of all the others, then the *maximum* probability of at least one Type I error in the set, or the *familywise* alpha level would be given by:

$$\alpha_{FW} \leq 1 - (1 - \alpha_{PC})^{C} \qquad \textbf{(1.1)}$$

The reason this formula gives you a *maximum* (rather than an exact value) for α_{FW} should be clear: If you fail to reject the null hypothesis in all *k* cases, then the probability of at least one Type I error is zero. If you reject all *c* null hypotheses, then α_{FW} will be equal to the right hand portion of equation 1.1.

In actual fact, there are dependencies among all possible pairwise comparisons, and this makes things more difficult. It is not possible to determine exactly the familywise (FW) alpha for several nonindependent *t*-tests. However, it is known that under all circumstances:

$$\alpha_{FW} \leq c\,\alpha_{PC} \qquad \textbf{(1.2)}$$

Equation 1.2 captures what is known as the **Bonferroni inequality**. Howell (1997, p. 362) explains it as follows: "...the probability of occurrence of one *or more* events can never exceed the sum of their individual probabilities. This means that when we make three comparisons, each with a probability = [.05] of Type I error, the probability of *at least* one Type I error can never exceed .15."

All of the foregoing has been concerned with doing multiple *t*-tests *in place of* the one-way ANOVA. The use of *t*-tests *after* rejection of the null hypothesis for an omnibus *F*-test is somewhat different. This technique is known as **Fisher's least significant difference (LSD)** test, or sometimes as **Fisher's protected t**. Under certain circumstances, Fisher's LSD test will

ensure that the FW alpha level is no greater than the per comparison (PC) alpha. In order to understand what these circumstances are, we must first clarify some terminology used by Howell (1997). What Howell calls the *complete null hypothesis* is just the null hypothesis for a one-way ANOVA, which states that all population means are equal. If you have 5 treatments, for example, the complete null hypothesis specifies that $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. When you are doing multiple comparisons, it may be that the complete null hypothesis is not true, but a more *limited null hypothesis* may be. For example, it may be that $\mu_1 = \mu_2 < \mu_3 = \mu_4 < \mu_5$.

According to Howell (1997, p. 369), "When the complete null hypothesis is true, the requirement of a significant overall *F* ensures that the familywise error rate will equal α. Unfortunately, if the complete null hypothesis is *not* true but some other more limited null hypothesis involving several means is true, the overall *F* no longer affords protection for *FW*."

Therefore, the LSD technique is not recommended *except* when you have three treatment levels. In this case the FW error rate will remain at α. To see why, consider the following scenarios: First, consider the case where the complete null hypothesis is true: $\mu_1 = \mu_2 = \mu_3$. In this case, the probability that you will commit a Type I error with your overall *F*-test is α; and any subsequent Type I errors that might occur (i.e., when you carry out the three possible pairwise *t*-tests) will not affect the FW error rate. (Note that this is true for any number of means *when the complete null hypothesis is true*.)

If the complete null hypothesis is not true, but a more limited null hypothesis is true, then it must be the case that two of the means are equal and different from the third. For example, it may be that $\mu_1 = \mu_2 < \mu_3$. In this case, it is not possible to make a Type I error when carrying out the omnibus *F*-test. (You can only make a Type I error when the null hypothesis is true, and the null hypothesis for the omnibus *F* is the *complete* null hypothesis--and we have just said that it is *not* true.) In this case, there will be only one pairwise comparison for which the null hypothesis is true, and therefore only one opportunity to make a Type I error. And so the probability of Type I error will be α_{PC}.

1.3 Calculation of *t* when *k* > 2

Before we go on, it should be noted that when you have more than 2 independent samples (i.e., when $k > 2$), t-values for multiple comparisons are computed in a way that differs somewhat from what you learned previously. Before everyone panics, let me emphasise that the difference is very slight, and that the reasons for the difference even makes sense!

Let us begin with the *t*-test for 2 independent samples. The formula boils down to:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}} \qquad \textbf{(1.3)}$$

And the standard error of the difference between 2 independent means is often calculated with this formula:

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{SS_1 + SS_2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \textbf{(1.4)}$$

But note that that $(SS_1 + SS_2)/(n_1 + n_2 - 2)$ is really a pooled variance estimate, or s^2_p. And so equation 1.4 can be rewritten as:

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \textbf{(1.5)}$$

Finally, note that when $n_1 = n_2$, equation 1.5 can be rearranged to give:

$$S_{\bar{X}_1-\bar{X}_2} = \sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{S_P^2\left(\frac{2}{n}\right)} = \sqrt{\frac{2S_P^2}{n}} \quad \textbf{(1.6)}$$

It should be clear that the main component of the denominator of the *t*-ratio is a pooled estimate of the population variance. Normally when you perform an independent samples t-test, this variance estimate is based on 2 samples. However, the more samples you can base it on, the more accurately it will estimate the population variance. And so, when you have *k* independent samples/groups (where $k > 2$), and when the homogeneity of variance assumption is tenable, it makes abundant sense to use a pooled variance estimate that is based on all *k* samples: $(SS_1 + SS_2 + ... SS_k) / (n_1 + n_2 + ... n_k - k)$. This formula should look familiar to you, by the way, because it is the MS_{within} or MS_{error} for the one-way ANOVA.

Finally then, when calculating t-ratios in the context of more than two independent samples, when all sample sizes are equal (and when the homogeneity of variance assumption is tenable), it is customary to use the following formula:

$$t = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2MS_{error}}{n}}} \quad \textbf{(1.7)}$$

1.4 The studentized range statistic, *q*

In our discussion to this point, we have glossed over another problem that arises when one has more than 2 samples and does multiple *t*-tests. It is well known that as sample size (*n*)

increases, so does the magnitude of the sample range. (In case you've forgotten, the range is the highest score in the sample minus the lowest score.) Imagine drawing random samples of various sizes from a normally distributed population with $\sigma = 10$. The effect of sample size on *expected value* of the range is clear in the following table:

Table 1.1 Expected values of the range for samples of various sizes drawn from a normal population with $\sigma = 10$.

Sample Size	Expected Value of Range
2	11
5	23
10	31
20	37
50	45
100	50
200	55
500	61
1000	65

Note that the effect of increased sample size on the expected value of the range is most pronounced when the sample sizes are relatively small. That is, an increase from $n = 2$ to $n = 5$ results in more than doubling of the expected value of the range (11 to 23). But an increase from $n = 500$ to $n = 1000$ results in a very modest increase in the expected value of the range (from 61 to 65).

If you're wondering what this has to do with multiple *t*-tests, remember that when you do a *t*-test, you are really comparing two scores that are drawn from a normal distribution: The two scores you compare are two sample means, and the normal distribution is the sampling distribution of the mean. Note that if you really have 5 samples (and 5 means), then the expected difference between the largest and smallest means (i.e., the range for the set of 5 means) will be much larger than if there are only 2 samples (and 2 means). In other words, if you were to draw 5 random samples from a normal population and compare the smallest and largest sample means, "the observed t-ratio would exceed the critical t-ratio far more often than the probability denoted by the nominal value of alpha" (Glass & Hopkins, 1984, p. 369).

Not surprisingly, there is a statistic that does take into account the number of samples: It is called the studentized range statistic, or *q*. The value of *q* is calculated by subtracting the smaller of two sample means from the larger, and dividing by the standard error of the mean:

$$q = \frac{\overline{X}_L - \overline{X}_S}{\sqrt{\frac{MS_{error}}{n}}} \qquad \textbf{(1.8)}$$

This formula looks very similar to formula 1.7, but note that MS_{error} is not multiplied by 2 when you calculate *q*. Perhaps the difference between *t* and *q* is even clearer in equation 1.9:

$$q = t\sqrt{2} \quad \textbf{(1.9)}$$

Just as there are critical values of t, there are critical values of q. Many statistics texts (including Howell, 1997) have tables of critical values of q. Most such tables show *Error df* down the left side of the table. This refers to the degrees of freedom for MS_{error} in the overall ANOVA. And across the top of the table is something like: *r = Number of Steps Between Ordered Means*. This refers to the number of means encompassed by the two means being tested. For example, if you have a set of 5 sample means rank ordered from smallest (M_1) to largest (M_5), the number of means encompassed by M_1 and M_5 would be 5--or in other words, $r = 5$.

1.5 The Tukey HSD Method of Multiple Comparisons

The Tukey HSD test is designed to make all pairwise comparisons between means while maintaining the FW error probability at α. (That is, if you set $\alpha = .05$ for the Tukey HSD test, the probability of at least one Type I error will be no greater than .05.) The test statistic is the studentized range statistic q as defined in equation 1.8. The critical value of q for *all* pairwise comparisons is the critical value with the *maximum value of r* for that set of means. For example, if there are 5 means in the set, then all mean differences are treated as if the two means were 5 steps apart.

Let us look at an example with $k = 5$ independent groups (with 8 in each group). The means of the 5 groups, and the results of a one-way ANOVA are shown below (and in Table 12.1 in Howell, 1997):

Table 1.2 Group means and ANOVA sumary table

Group	Mean	Source	*df*	*MS*	*F*	*p*
M-S	4	Between	4	874.40	27.33	< .01
M-M	10	Error	35	32.00		
S-S	11					
S-M	24					
Mc-M	29					

There are two ways we could go about carrying out the Tukey HSD test. On the one hand, we could calculate q for each of the possible pairwise comparisons, and compare the obtained value to the critical value. (We would reject the null hypothesis of no difference between the means if the obtained value of q was equal to or greater than the critical value.)

The other approach is to rearrange equation 1.8 in such a way that we can calculate a *critical difference* between means. To be consistent with Howell (1997), I will call this critical difference W (for width). W = the smallest width (or difference) between means that will be significant. The formula for calculating W is:

$$W_{crit} = q_{crit}\sqrt{\frac{MS_{error}}{n}} \quad \textbf{(1.10)}$$

As described earlier, the critical value of q depends on the degrees of freedom associated with the error term (from the overall ANOVA), and on the number of means in the set. It also depends on the alpha level you have chosen. For this example, df = 35 (from the ANOVA summary table), and $r = 5$. If we set $\alpha = .05$, then the critical value of $q = 4.07$. (Note that there is no entry for df = 35, so I averaged the values for df = 30 and df = 40.) The square root of $(MS_{error}/n) = 2$. And so $W = 4.07(2) = 8.14$. In other words, any difference between means that is equal to or greater than 8.14 will be declared significant at the .05 level.

The next step is to construct a table of mean differences as follows (each cell entry is the mean at the top of the column minus the mean at the left end of the row):

Table 1.3 Mean pairwise differences between conditions.

		M-S (4)	**M-M** (10)	**S-S** (11)	**S-M** (24)	**Mc-M** (29)
M-S	(4)		6	7	20	25
M-M	(10)			1	14	19
S-S	(11)				13	18
S-M	(24)					5

All significant differences (i.e., those larger than 8.14) are shown in the shaded area of the table. It is clear that there are no significant differences amongst the 3 smallest means; and that there is no significant difference between the 2 largest means. Furthermore, the largest 2 means do differ significantly from the smallest 3. This information is sometimes conveyed by writing down the treatments (i.e., just the group names/codes in this example) and underlining the homogeneous subsets. For the present results, for example:

<u>M-S M-M S-S</u> <u>S-M Mc-M</u>

1.6 The Newman-Keuls Method

The **Newman-Keuls** (NK) method of multiple comparisons is very similar to the Tukey HSD method. But whereas the Tukey test uses only one critical value of q (i.e., the critical value for the largest value of r), the NK test uses k-1 critical values of q. So in the previous example with $k = 5$ treatments, you would need 4 different critical values of q. Using df = 35 once again, the critical values of q would be:

$q_2 = 2.875$
$q_3 = 3.465$
$q_4 = 3.815$
$q_5 = 4.070$

The subscripts on the q's indicate the value of r, or the number of means encompassed by the 2 means being compared. Thus, when you compare the largest and smallest means, the critical value of $q = 4.07$, just as it was for the Tukey HSD test. But as the number of means encompassed by the 2 being compared decreases, so does the critical value of q. These critical values can be converted to critical differences (W) using equation 1.10:

$$W_2 = 2.875(2) = 5.75$$
$$W_3 = 3.465(2) = 6.93$$
$$W_4 = 3.815(2) = 7.63$$
$$W_5 = 4.070(2) = 8.14$$

The mean differences in Table 1.3 would then be compared to these critical differences. Note that with this set of critical differences, we would conclude that the difference between M-S and M-M (10-4) and the difference between M-S and S-S (11-4) are both significant (which we were unable to do with the Tukey HSD test). Thus, the results of the Newman-Keuls test on these data could be summarized as follows:

M-S	M-M	S-S	S-M	Mc-M

1.7 Comparison of Tukey and NK Methods

The Tukey and NK methods both use the studentized range statistic, q. The main difference between them is that the Tukey method limits the FW alpha level more strictly. (In fact, it limits the probability of at least one Type I error to α.) It achieves this by treating all pairwise comparisons as if the 2 means were k steps apart, and using just one critical value of q. Note that for the initial comparison of the smallest and largest means, the Tukey and NK tests are identical. But for subsequent comparisons, the NK method will reject the null hypothesis more easily, because the critical value of q becomes smaller as the number of means in the range decreases.

Some researchers and statisticians shy away from the NK test, because they feel that it is too liberal--i.e., that the probability of Type I error is too high. We will return to this issue later.

1.8 Linear Contrasts

Before going on to other MC methods, we must understand what a **linear contrast** is. A pairwise t-test (see equation 1.7) is a special kind of linear contrast that allows comparison of one mean with another mean. In general, linear contrasts allow us to compare one mean (or set of means) with another mean (or set of means).

It may be easier to understand what a linear contrast is if we first define a **linear combination**. According to Howell (1997, p. 355), "a linear combination [L] is a weighted sum of treatment means ":

$$L = a_1\overline{X}_1 + a_2\overline{X}_2 + ... + a_k\overline{X}_k = \sum_{i=1}^{k} a_i\overline{X}_i \qquad \textbf{(1.11)}$$

A linear combination becomes a linear contrast when we impose the restriction that the *coefficients* must sum to zero ($\Sigma a_i = 0$).

The contrast coefficients are just positive and negative numbers (and zeroes) that define the hypothesis to be tested by the contrast. Table 1.4 shows the coefficients for making all possible pairwise contrasts when there are 5 means.

Table 1.4 Coefficients for all pairwise contrasts involving 5 sample means.

Means being compared	**1**	**2**	**3**	**4**	**5**	**Σa_i**
1 v 5	1	0	0	0	-1	0
1 v 4	1	0	0	-1	0	0
1 v 3	1	0	-1	0	0	0
1 v 2	1	-1	0	0	0	0
2 v 5	0	1	0	0	-1	0
2 v 4	0	1	0	-1	0	0
2 v 3	0	1	-1	0	0	0
3 v 5	0	0	1	0	-1	0
3 v 4	0	0	1	-1	0	0
4 v 5	0	0	0	1	-1	0

1.9 Simple versus Complex Contrasts

Contrasts that involve only two means, with contrast coefficients equal to 1 and -1, are called **simple** or **pairwise contrasts**. All of the contrasts in Table 1.4, for example are simple contrasts.

Complex contrasts involve 3 or more means. For example, with a set of 5 means, it may be hypothesised that the mean of the first two means is different from the mean of the last 3 means. One set of coefficients that would work for this particular contrast is:

3 3 -2 -2 -2

It may not be immediately obvious how I arrived at these coefficients, so let's work through it in steps.

Step 1: Mean of groups 1 & 2 = $(M_1 + M_2)/2 = (1/2)M_1 + (1/2)M_2$

Step 2: Mean of groups 3-5 = $(M_3 + M_4 + M_5)/3 = (1/3)M_3 + (1/3)M_4 + (1/3)M_5$

Step 3: We could use as our coefficients the fractions in the right-hand portions of the equations in Steps 1 and 2. Note that one set of coefficients (which set is arbitrary) would have to be made negative so that the sum of the coefficients is zero. Thus, our coefficients could be:

1/2 1/2 -1/3 -1/3 -1/3

Step 4: Some texts recommend stopping at this point, but others suggest that it is easier to work with coefficients that are whole numbers. To convert the coefficients to whole numbers, multiply each one by the lowest common denominator. In this case, that means multiplying each one by 6. Doing so yields the set of coefficients shown above.

*Note that there is a shortcut that does not involve so many steps. The coefficient for the first set of means in a contrast equals the **number** of means in the second set; and the coefficient for the second set of means equals the **number** of means in the first set. Finally, the coefficient for one of the two sets is arbitrarily selected and made negative. (Note that if you had 10 means, and had coefficients of 6 and -4, these could be reduced to 3 and -2.)*

In the example given earlier, there are 3 means in the second set, and so the coefficient for the first set of means is 3; and there are 2 means in the first set, so the coefficient for the second set of means is 2. Making the 2's negative yields the set of coefficients listed earlier.

1.10 Testing the Significance of a Linear Contrast

Many MC methods use a modified *t*-ratio, or an *F*-ratio as the test statistic. Here, we will focus on the modified *t*-test. (See Howell (1997, 1992) for an explanation of the *F*-test--and note that $F = t^2$.)

Before looking directly at the *t*-ratio for a linear contrast, let me remind you that in general, ***t* = (statistic - parameter) / (standard error of statistic).** In the case of a single sample *t*-test, the statistic is a sample mean, the parameter is the (null hypothesis) population mean, and the term in the denominator is the standard error of the mean (i.e., the sample standard deviation divided by the square root of the sample size. For an independent samples *t*-test (equation 1.3), the statistic is the difference between 2 sample means; the parameter is the (null hypothesis) difference between the two population means (which is usually equal to zero); and the term in the denominator is the standard error of the difference between two independent means (see equations 1.4 and 1.5).

When you test the significance of a linear contrast, the statistic in the numerator of the t-ratio is the linear contrast L (equation 1.11). Note that L is computed using *sample means*. Therefore, L is really an *estimate* of a corresponding contrast that uses *population means*. This contrast that uses population means (rather than sample means) is the parameter for our *t*-ratio. In almost all cases, however, the null hypothesis specifies that this parameter equals zero, and so it can be left off the formula.

The final piece of the formula is the **standard error of the linear contrast**, which is computed as follows:

$$S_L = \sqrt{MS_{error} \sum \frac{a^2}{n}} \quad \textbf{(1.12)}$$

And when all sample sizes are equal, this reduces to:

$$S_L = \sqrt{\left(\frac{MS_{error}}{n}\right) \sum a^2} \quad \textbf{(1.13)}$$

Finally, note that for simple (pairwise) contrasts with $n_1 = n_2 = n$, the coefficients are 1 and -1, so formula 1.13 reduces to:

$$S_L = \sqrt{\frac{2MS_{error}}{n}} \quad \textbf{(1,14)}$$

This is the same as the denominator of the *t*-ratio shown in equation 1.7.

Finally then, the t-ratio for a linear contrast is calculated with equation 1.15:

$$t = L/s_L \quad \textbf{(1.15)}$$

1.11 Planned versus Post Hoc Comparisons

There is a very important distinction between *planned* (or *a priori*) and *post hoc* (or *a posteriori*) contrasts. Glass and Hopkins (1984, p. 380) say this about the distinction:

> In planned contrasts, the hypotheses (contrasts) to be tested must be specified *prior* to data collection. MC methods which employ planned comparisons can be advantageous if the questions that the researcher is interested in are a relatively small subset of questions. The distribution theory and probability statements for these MC methods are valid only if there is no chance for the user to be influenced by the data in the choice of which hypotheses are to be tested. The rationale for planned contrasts is similar to that for "one-tailed" *t*-tests--to be valid, the decision must be made *a priori*. Post hoc MC techniques do not require advance specification of the hypotheses (contrasts) to be tested. The Tukey and NK MC methods are considered to be post hoc methods since there is no delimitation as to which pairs of means will be contrasted.

To further illustrate how important this distinction really is, consider the scenario described by Howell (1997, p. 350). If you have 5 means, there will be 10 possible pairwise comparisons. Let us imagine that the complete null hypothesis is true (i.e., all 5 population

means are equal). Let us also imagine that the smallest and largest sample means differ by enough to allow rejection of the null hypothesis for a pairwise comparison, but that no other pairwise comparisons would allow rejection of the null hypothesis. Now imagine that you are going to carry out *one* pairwise contrast. As Howell points out:

> If you have to plan your single comparison in advance, you have a probability of .10 of hitting on the 1 comparison out of 10 that will involve a Type I error. If you look at the data first, however, you are certain to make a Type I error, assuming that you are not so dim that you test anything other than the largest difference.

1.12 Dunn's Method of Multiple Comparisons (Bonferroni *t*)

Dunn's MC method uses the *Bonferroni inequality* (described in section 1.2 of these notes). It consists of doing multiple planned (*a priori*) *t*-tests with $\alpha_{PC} = \alpha_{FW}/c$, where c = the number of contrasts. Recall that the probability of at least one Type I error in a family of c contrasts can be no greater than $c\alpha_{PC}$. For example, if 5 significance tests are carried out with $\alpha_{PC} = .01$, the FW alpha cannot exceed .05. Note that in a case like this, you could simply look up the critical value of t with the appropriate degrees of freedom and $\alpha = .01$. But if you were to carry out 3 significance tests rather than 5, then you would set $\alpha_{PC} = .05/3 = .0167$. It is not possible to find a critical t-value for $\alpha = .0167$ in an ordinary table of critical t-values. (Try it if you don't believe me!) Dunn's important contribution was to provide a table of critical values (see Appendix t' in Howell, 1997). For this example with $c = 3$ comparisons and $\alpha_{FW} = .05$, if we let $df = 30$, the critical value of t' is 2.54. That is, for each of the 3 comparisons, the null hypothesis could only be rejected if the computed value of t was equal to or greater than 2.54.

Note as well that for $c = 5$ comparisons and $\alpha_{FW} = .05$, the critical value of t' is equal to the critical value of t with $\alpha = .01$: If $df = 30$, for example, $t' = 2.75$, and $t = 2.75$. Why? Because when $\alpha_{FW} = .05$ and $c = 5$, $\alpha_{PC} = .01$.

Finally, note that Dunn's method is very flexible in that it allows any number of simple and/or complex contrasts. However, it also requires that the contrasts be ***planned*** *(see section 1.11).*

1.13 Dunnett's Test (for comparing all treatments with a control)

The **Dunnett test** is designed for the situation where you wish to compare each of k-1 treatment conditions with one control condition. That is to say that there are $c = k$-1 *planned contrasts*. The test statistic is a modified t-ratio shown in equation 1.7. Note however that one of the sample means in the numerator is always the mean for the control group.

Critical t values for Dunnett's test are tabled in many statistics textbooks. See for example, Appendix t_d in Howell (1997, 1994). If we had Error $df = 30$, $\alpha_{FW} = .05$, and $k = 5$ treatments (including the control condition), then the critical value of t would be 2.58.

Note that with 5 treatments (including the control condition), there would be 4 comparisons. Those same 4 comparisons could have been carried out using Dunn's test. But

note that the critical value of t for Dunn's test in this same situation would be 2.66. That is, for each of the comparisons with the control condition, it would be harder to reject the null hypothesis with Dunn's test than with Dunnett's test. In other words, when you plan to compare each treatment with a control condition, Dunnett's test is more powerful than Dunn's test.

Finally, note that contrary to my earlier assertion, Howell (1997, p. 381) describes Dunnett's test as a post hoc procedure. There is disagreement as to whether the k-1 contrasts have to be planned or not. I side with Glass and Hopkins (1984) on this issue: They say that Dunnett's test may be used "where the plan is to compare each of the J - 1 means with one (and only one) *predesignated* mean (usually the mean of the control group)." The words "plan" and "predesignated" make it abundantly clear that Dunnett's test is **not** a post hoc test, and should be used in conjunction with **planned comparisons** of each treatment with a "predesignated" control.

1.14 The Scheffé Method of Multiple Comparisons

The **Scheffé test** is a very flexible **post hoc** MC method. It allows one to carry out any and all pairwise and complex contrasts while maintaining control over the FW alpha. As a result, the critical t (or F) value for the Scheffé test is greater than for any of the other MC procedures we will discuss. In other words, it is the *most conservative/least powerful* MC procedure we will consider.

The test statistic is t as defined in equation 1.15; and the critical value of t is calculated as follows:

$$t_S = \sqrt{(k-1)F_{FW(k-1,df_{error})}} \qquad \textbf{(1.16)}$$

The F_{FW} in equation 1.16 is the critical F-value with α equal to the desired FW alpha level. If the obtained t-ratio is equal to or greater than t_S, then one can reject the null hypothesis for that contrast.

Note that some books (e.g., Howell, 1997) describe the test statistic for the Scheffé method as an F-ratio. To carry out the test that way, you can simply use the square of equation 1.15 for calculating the test statistic, and the square of equation 1.16 for calculating the critical F-value.

Because the Scheffé test is so conservative, it is not generally recommended except for **post hoc data snooping**--especially when you wish to carry out several contrasts. In most other situations, another MC procedure is likely to be much more suitable.

1.15 Planned Orthogonal Contrasts

We have already discussed the important distinction between planned and post hoc contrasts (see section 1.11). We now must consider the special case of *planned **orthogonal***

contrasts (POC's). In this context, orthogonal means independent. Howell (1997, p. 360) gives a nice example that should help clarify the difference between orthogonal and non-orthogonal contrasts:

> Sometimes contrasts are independent of one another [i.e., orthogonal], and sometimes they are not. For example, knowing that M_1 is greater than the average of M_2 and M_3 tells you nothing about whether M_4 is likely to be greater than M_5. These two contrasts are independent. However, knowing that M_1 is greater than the average of M_2 and M_3 suggests that there is a better than 50:50 chance that M_1 is greater than M_2. These two contrasts are not independent.

Let us imagine that we have two contrasts involving a set of 5 means. Let a_1 - a_5 be the coefficients for the first contrast, and b_1 - b_5 for the second contrast. If this pair of contrasts is orthogonal then all of the following will be true: $\Sigma a = 0$; $\Sigma b = 0$; and $\Sigma ab = 0$. Here is an example of two contrasts that **are** orthogonal:

	Coefficients					Sum
Contrast 1	3	3	-2	-2	-2	0
Contrast 2	1	-1	0	0	0	0
a(b)	3	-3	0	0	0	0

The first contrast compares the mean of the first two treatments to the mean of the last three treatments; and the second contrast compares the means of the first two treatments. Note that the sum of the coefficients for each contrast is zero; and the sum of the products of the coefficients is also zero (3 - 3 + 0 + 0 + 0 = 0).

Now here is a pair of contrasts that are **not** orthogonal:

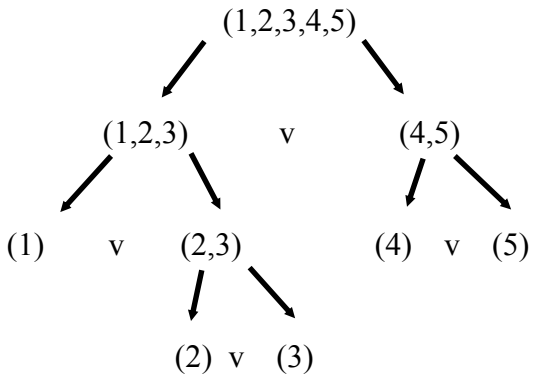
	Coefficients					Sum
Contrast 1	1	0	0	0	-1	0
Contrast 2	1	0	0	-1	0	0
a(b)	1	0	0	0	0	1

The first contrast looks for a difference between the first and fifth means; and the second contrast looks for a difference between the first and fourth means. It might seem that these two contrasts are independent. That is, you might think that knowledge that the first contrast is significant provides no information whatsoever about the second contrast. Nevertheless, the sum of the products of the coefficients does *not* equal zero (1+0+0+0+0 = 1); and therefore, despite what intuition might be telling you, these two contrasts are *not* orthogonal.

A *set* of contrasts (i.e., more than 2) is orthogonal if *each possible pair* of contrasts is orthogonal according to the rules described above. The maximum number of orthogonal

contrasts is k - 1, where k = the number of treatments. (Note that k - 1 also equals df_{error} in the one-way ANOVA.)

Howell (1997, pp. 361-362) explains a simple method for coming up with complete sets of k -1 orthogonal contrasts. It entails the use of a tree diagram. The trunk of the tree has all k treatments. The first contrast forms two branches of the tree. For example, if our first contrast compares the mean of treatments 1-3 to the mean of treatments 4 and 5, then one branch represents the first 3 treatments, and the other branch the last two treatments. In order for the complete set of contrasts to be orthogonal, all subsequent contrasts must avoid comparison of treatments on different limbs of the tree. According to Howell, this rule will find most, but not all possible sets of orthogonal contrasts. One possible set of orthogonal contrasts could be derived as follows:



The coefficients for the 4 contrasts diagrammed here would be as follows:

	Coefficients					Sum
(1,2,3) v (4,5)	2	2	2	-3	-3	0
(1) v (2,3)	2	-1	-1	0	0	0
(4) v (5)	0	0	0	1	-1	0
(2) v (3)	0	1	-1	0	0	0

If you wished to verify that this set of contrasts is indeed orthogonal, you could calculate the sum of the products of the coefficients (i.e., Σab) for each of the 6 possible pairs of contrasts. If Σab = 0 in all cases, then the set is orthogonal.

1.16 Testing the Significance of a Linear Contrast with an *F*-ratio

I have already discussed (in section 1.10) how to test the significance of a linear contrast using a modified *t*-test (see equation 1.15). That method is perfectly valid, and has at least these two advantages: 1) it makes explicit the fact that testing the significance of a contrast entails division of the contrast by its standard error; and 2) it is similar to the procedure used for several other MC methods (e.g., Dunn's test, Dunnett's test). However, one important fact about orthogonal contrasts is much more apparent if one uses *F*-ratios to test for significance.

Equation 1.15 shows the *t*-ratio for testing the significance of a linear contrast. Recall that $t^2 = F$. If we square both sides of equation 1.15, we get:

$$F(1, df_{error}) = \frac{L^2}{S_L^2} \qquad \textbf{(1.17)}$$

When all sample sizes are equal, the standard error of a contrast is computed as in equation 1.13. Substituting this into equation 1.17 gives:

$$F(1, df_{error}) = \frac{L^2}{\left(\frac{MS_{error}}{n}\right)\sum a^2} \qquad \textbf{(1.18)}$$

Finally, multiplying both numerator and denominator by *n*, and dividing both numerator and denominator by Σa^2 yields this equation, which can be found in Howell (1997, 1994):

$$F(1, df_{error}) = \frac{nL^2 \Big/ \sum a^2}{MS_{error}} \qquad \textbf{(1.19)}$$

In ANOVA, *F* is equal to the ratio of two "mean squares". The denominator in equation 1.19 is MS_{error} from the overall ANOVA. It should not surprise you, therefore that the numerator of equation 1.19 gives the "mean square of the contrast", or MS_{contrast}. Note as well that because any linear contrast has $df = 1$, the numerator is also equal to SS_{contrast}. (Recall that $MS = SS/df$, so if $df = 1$, $MS = SS$.) This relationship is shown in equation 1.20:

$$\text{MS}_{\text{contrast}} = \text{SS}_{\text{contrast}} = \frac{nL^2}{\sum a^2} \qquad \textbf{(1.20)}$$

And so equation 1.19 can be simplified to:

$$F(1, df_{\text{error}}) = MS_{\text{contrast}} / MS_{\text{error}} \qquad \textbf{(1.21)}$$

Now we are in a position to finally get to the point of this exercise: **When you have a complete set of k - 1 orthogonal contrasts, the sum of $SS_{contrast}$ will be equal to $SS_{treatments}$:**

$$\Sigma SS_{contrast} = SS_{treatments} \quad \textbf{(1.22)}$$

Or putting it another way, **orthogonal contrasts allow you to partition $SS_{treatments}$ into k - 1 independent, non-overlapping components**.

Planned orthogonal contrasts are carried out *regardless* of the significance or nonsignificance of the omnibus F, and they use a **per contrast error rate**. If you test significance with a t-ratio (equation 1.15), the critical value of t can be taken from an ordinary table of critical t-values with $df = df_{error}$. If you use the F-test shown in equation 1.21, the critical value of F can be taken from the central F-distribution with 1 and df_{error} degrees of freedom. As noted by Glass and Hopkins (1984, p. 384), "The requirement that the contrasts be both planned and orthogonal makes the POC procedure very different from multiple t-tests [or F-tests], even though the critical t-ratio [or F-ratio] is ostensibly the same." (See also section 1.11 of these notes.)

Finally, it should be noted that there is not universal agreement amongst statisticians and researchers on the requirement of orthogonality for planned contrasts. Keppel and Winer are prominent textbook authors who maintain that planned contrasts need not be orthogonal so long as they are **small in number** (i.e., about k - 1) and **meaningful** in the experimental context.

1.17 Trend Analysis

Trend analysis is really just a special case of orthogonal contrasts that can be used when the *independent variable is continuous*. Under these circumstances, in fact, trend analysis may be much more appropriate and informative than any of the MC procedures we have discussed so far.

Suppose for example that we had $k = 7$ seven groups of subjects in equally spaced age groups (10, 15, 20, 25, 30, 35, and 40 years of age), and that we took some measure of psychomotor performance. Imagine that the means and ANOVA results are as shown below:

Although it is not required that the step size between adjacent levels of the independent variable be constant (as in the example above, where the step size is 5), *trend analysis is definitely easier with a constant step size*. In this case, it is possible to look up the needed **contrast coefficients** in a table (e.g., *Appendix Polynomial* in Howell, 1997). (Should you ever need to work out your own coefficients for use with unequal intervals, see Howell, 1997, 1994, for guidance.)

Table 1.5 Mean psychomotor performance for 7 age groups (n=4 in each group)

Age	Mean
10	8.5
15	9.5
20	10.5
25	11.5
30	10.0
35	9.0
40	8.5

Source	*df*	*MS*	*F*	*p*
Between	6	4.905	4.12	< .01
Error	21	1.190		

Just as it is possible to carry out k - 1 planned orthogonal contrasts, one can test for k - 1 different types of trends (where k = the number of treatment levels). The first and simplest of these is a **linear trend** (sometimes called first-degree trend). A significant linear trend tells you that as the amount of the independent variable increases, the amount of the dependent variable increases (or decreases) in linear fashion. (Given this description, it may not surprise you to discover that the linear trend is closely related to Pearson r, as we will see later.) To test for a significant linear trend, we need to construct a contrast, L_{linear}. The coefficients can be obtained in a table of coefficients of orthogonal polynomials. For this example, with k = 7, the coefficients are: -3, -2, -1, 0, 1, 2, and 3. (Note that as with any linear contrast, the coefficients sum to zero.) Substituting these coefficients into equation 1.11 give us the following:

$$L_{\text{linear}} = -3(8.5) + (-2)(9.5) + (-1)(10.5) + 0(11.5) + 1(10.0) + 2(9.0) + 3(8.5) = -1.5$$

$$MS_{\text{linear}} = SS_{\text{linear}} = nL^2_{\text{linear}} / \Sigma a^2 = 4(-1.5)^2 / 28 = .321$$

$$F(1,21) = MS_{\text{linear}} / MS_{\text{error}} = .321 / 1.19 = .27$$

The critical F-value is taken from the central (i.e., ordinary) F-distribution, just as with POCs. Because the F-value is less than 1, we would conclude that there is **not** a significant linear trend in these data.

We could then move on to the second-degree **(quadratic)**, third-degree **(cubic)**, fourth-degree **(quartic)**, and so on, right up to the sixth-degree trend. (See Howell, 1997, 1994, for examples of quadratic functions.) If we were to do so--and if we did not make any calculation errors!--we would find that the sums of squares (SS) for these k -1 trends add up to $SS_{\text{treatments}}$. In practice, however, it is rare to go much beyond the third-degree, or cubic trend for two reasons: First, it is often the case that most of $SS_{\text{treatments}}$ is already exhausted by that point; and second, it is very difficult to understand trends beyond the third- or fourth-degree variety even if they do exist!

The coefficients for the linear, quadratic, and cubic trends with k = 7 treatments are as follows:

	Coefficients (a_1-a_7)							Σa^2
Linear	-3	-2	-1	0	1	2	3	28
Quadratic	5	0	-3	-4	-3	0	5	84
Cubic	-1	1	1	0	-1	-1	1	6

I will leave it to you to work out the tests of quadratic and cubic trends. You should end up with results similar to those shown in Table 1.6.

Table 1.6 Trend analysis summary

Source	*SS*	*df*	*MS*	*F*	*p*
Age	29.43	6	4.905	4.12	< .01
Linear	0.32	1	0.32	0.27	
Quadratic	24.11	1	24.11	20.26	< .01
Cubic	0.67	1	0.67	0.56	
Remainder	4.33	3	1.44	1.21	
Error	25.00	21	1.190		

NOTE: $SS_{remainder} = SS_{age} - SS_{linear} - SS_{quadratic} - SS_{cubic} = 29.43 - 0.32 - 24.11 - 0.67 = 4.33$.

The results summarized in Table 1.6 might be described as follows in a research report:

> The main effect of Age was significant, $F(6,21) = 4.12$, $MS_{error} = 1.19$, $p < .01$. As can be seen in Table 1.5, mean psychomotor performance improved with increasing age up to 25 years. Thereafter, performance deteriorated with increasing age. Trend analysis revealed that only the quadratic component was significant, $F(1,21) = 20.26$, $p < .01$, and that it accounted for 82% of the treatment effect. A test of all remaining components (pooled) was not significant, $F(5,21) = 0.89$, *n.s.*

1.18 Relationship Between Trends and Correlation Coefficients

The presence of a significant linear trend tells you that as the value of the independent variable increases, the value of the dependent variable also increases (or decreases if it is a negative linear relationship). It may not surprise you to discover, therefore, that there is a mathematical relationship between SS_{linear} and Pearson *r*:

$$|r| = \sqrt{\frac{SS_{linear}}{SS_{total}}} \qquad \textbf{(1.23)}$$

Similarly, eta-squared (the correlation ratio) can be computed as follows:

$$\eta^2 = SS_{treatments} / SS_{total} \quad \textbf{(1.24)}$$

Finally, note that it is possible to test for something called "non-linearity of regression", which may arise when we talk about multiple regression later on. You don't need to worry about this now, but for the sake of completeness, here's how you could go about doing it (if you were so inclined!):

1. $SS_{nonlinear} = SS_{treatments} - SS_{linear}$
2. $MS_{nonlinear} = SS_{nonlinear} / (k - 2)$
3. $F(k\text{-}2, df_{error}) = MS_{nonlinear} / MS_{error}$

1.19 Selecting a MC Procedure

Table 12.5 from Howell (1997, reproduced below) and Figure 17.3 from Glass and Hopkins (1984, see the Appendix) can be used to help you decide which procedure to use. Note that Howell's table labels Dunnett's test as post hoc, whereas Glass and Hopkins' flowchart indicates that the k-1 contrasts must be planned. As I stated earlier, I agree with Glass and Hopkins on this point.

TABLE 12.5 Comparison of alternative multiple comparison procedures

Test	Error Rate[a]	Comparison	Type	A Priori or Post Hoc
1. Individual *t*-tests	*PC*	Pairwise	*t*	*a priori*
2. Linear contrasts	*PC*	Any contrasts	*F*	*a priori*
3. Bonferroni *t* (Dunn's test)	*PE or FW*	Any contrasts	*t*[c]	*a priori*
4. Holm; Larzelere & Mulaik	*FW*	Any contrasts	*t*[c]	*a priori*
5. Fisher's LSD	*FW*[b]	Pairwise	*t*	*post hoc*
6. Newman-Keuls	*FW*[b]	Pairwise	Range	*post hoc*
7. Ryan (REGWQ)	*FW*	Pairwise	Range	*post hoc*
8. Tukey HSD	*FW*	Pairwise[d]	Range[c]	*post hoc*
9. Scheffé test	*FW*	Any contrasts	*F*[c]	*post hoc*
10. Dunnett's test	*FW*	With control	*t*[c]	*post hoc*

[a] *PC* = per comparison; *PE* = per experiment; *FW* = familywise
[b] Against complete H_0
[c] Modified
[d] Tukey HSD can be used for all contrasts, but is poor in this case

Finally, let me draw your attention to an important point raised by Howell (1997, p. 383):

> People often fail to realize that in selecting a [MC] test it is perfectly acceptable to compare each of several tests on your own data in terms of the size of the **critical values** [emphais added], and to select a test on that basis.... The important point is that these decisions have to be based on a consideration of the critical values, and not the final results. You can't just try out every test you have and choose the one that gives you the answers you like.

1.20 Miscellany & Conclusions

As you may have already noticed, the topic of MC procedures is one of the more controversial areas in statistics. Because people differ in how liberal or conservative they wish to be (with respect to the probability of making a Type I error), there can be serious disagreement.

Psychologists tend to be quite conservative--in fact, many of them seem to have a pathological fear of making a Type I error. For example, they are very suspicious of one-tailed hypothesis tests, even though they can be perfectly legitimate under the right circumstances. And when it comes to post hoc pairwise comparisons, many of them have a clear preference for Tukey's HSD test over the more liberal Newman-Keuls procedure.

Why are psychologists such a conservative bunch? I think it is because of the way many of them were taught about the relative costs of Type I and Type II errors. In my own undergraduate (and postgraduate) statistics courses, we were told that the Type I error is the more costly error. That is, if you make a Type I error, you will almost certainly waste a lot of time and money chasing after some treatment effect that does not really exist. At the same time, it was implied that Type II errors are not nearly so costly, because when you make a Type II error, you just fail to detect a real effect.

But when you stop and think about this for a moment, maybe we've had it all wrong. I think one could argue that the Type II error is the really costly one--at least in fields where subjects are readily available, and replications are easy and inexpensive. (In such fields of research, effects are not generally believed until they have been independently replicated anyway!) If I am unfortunate enough to make a Type I error, surely I (or someone else) will soon discover that the result cannot be replicated, and the error will be exposed. But if I perform an experiment that is somewhat lacking in power, and use very conservative statistical tests, I may **fail to discover a real effect**. (Note that a Type II error is really a failure to discover something!) Think of the potential costs associated with this failure to discover: It may be years before someone makes the discovery that should have been mine. Or worse yet, the discovery may never be made. From this more moderate point of view, making an occasional Type I error may not be as disastrous an event as many psychologists seem to believe.

Post hoc MC Procedures and the Significance of the Omnibus F

The traditional approach to **post hoc** MC procedures has been that they should be carried out only if the omnibus *F*-test is significant. However, views on this are changing. Note that for

the Fisher LSD test, a significant overall *F* is required. As Howell (1997, p. 351) puts it: "...the rationale underlying the error rates for Fisher's least significant different [*sic*] test, to be discussed in Section 12.4, required overall significance." However, the logic underlying many other post hoc MC procedures does not require significance of the overall *F*. Once again, I turn to Howell (1997, p. 351) for this explanation:

> First of all, the hypotheses tested by the overall test and a multiple-comparison test are quite different, with quite different levels of power. For example, the overall *F* actually distributes differences among groups across the number of degrees of freedom for groups. This has the effect of diluting the overall *F* in the situation where several group means are equal to each other but different from some other mean. Second, requiring overall significance will actually change the *FW* [alpha], making the multiple-comparison tests [more] conservative. The tests were designed, and their significance levels established, without regard to the overall *F*.

Let me remind you once again in closing that psychologists are a very conservative bunch when it comes to MC procedures and control over FW alpha levels. Therefore, should any of you be in a position next year of using a post hoc MC procedure (other than Fisher's LSD test) when the overall *F* is not significant, be prepared for objections from your project supervisor. He or she is almost certain to hold the traditional view that all post hoc tests require a significant overall *F*.

Review Questions

1. If a *t*-test is used to compare the smallest and largest means in a set of 5 means, which of the following is true? The probability of a Type I error is

a. greater than
b. equal to
c. less than

the α associated with the critical *t*-value.

2. Are there any circumstances under which multiple t-tests are recommended as a MC procedure? If so, describe those circumstances. If not, explain why not.

3. In making all pairwise comparisons among 6 means, MC methods that use a familywise Type I error rate will tend to make _____ (fewer/more) Type I errors and _____ (fewer/more) Type II errors than methods that employ a per comparison Type I error rate.

4. Why is it not appropriate to treat post hoc comparisons as if they were planned in advance?

5. Which of these is **not** a valid contrast?

a. $L = \overline{X}_2 - \overline{X}_3$

b. $L = \overline{X}_1 + \overline{X}_2 + \overline{X}_3 = 0$

c. $L = -\overline{X}_1 + 2\overline{X}_2 - \overline{X}_3$

d. $L = \overline{X}_1 - \overline{X}_2 - \overline{X}_3 + \overline{X}_4$

e. $L = 3\overline{X}_1 - \overline{X}_2 - \overline{X}_3 - \overline{X}_4$

6. Which of the valid contrasts above are *complex* contrasts?

7. What is the null hypothesis implicit in Question 5c?

8. Are these contrasts in Question 5 orthogonal?

a. a and c
b. a and d
c. a and e
d. c and e
e. d and e

Indicate which of the following methods is the proper MC method for the situations described in Questions 9-14.

a. Dunnett's test
b. Newman-Keuls test
c. POC's
d. Fisher's LSD test
e. Scheffé test
f. Tukey HSD test
g. Dunn's test (Bonferroni t)

9. You wish to carry out all pairwise comparisons and to maintain control over the FW error rate.

10. You have $k = 4$ independent groups, and prior to data collection intend to test these three null hypotheses: $\mu_1 = \mu_2$; $\mu_3 = \mu_4$; and $(\mu_1 + \mu_2) = (\mu_3 + \mu_4)$.

11. Which methods use a per comparison alpha?

12. You have 5 groups, and are only interested in pairwise comparisons of Group 1 (the control group) to each of the other 4 conditions.

13. You have 9 groups and a significant omnibus *F*, and now wish to do some "data snooping" involving both simple and complex contrasts.

14. You have 3 independent groups and a significant omnibus *F*, and now wish to make all pairwise comparisons.

15. In which of these ways does the Scheffé method differ from planned orthogonal contrasts (POC)?

a. in the coefficients used for a particular contrast
b. in computing the contrast (L)
c. in calculating *t* or *F*
d. in the critical value for the test statistic

16. Newman-Keuls and Tukey tests give identical results when:

a. there are $k = 3$ groups
b. $r = 2$ (i.e., the 2 means are adjacent)
c. comparing the extreme-most means
d. *n* is large

17. Under what conditions would you consider using trend analysis?

18. If you had an experiment with 6 (equally spaced) levels of a continuous independent variable, how many orthogonal trends could be evaluated?

19. What are the coefficients for evaluating the cubic trend with $k = 6$ groups?

20. Assume that you have a significant linear trend in your data. How would you calculate the percentage of the treatment effect that is accounted for by that linear trend?

21. Assume that $SS_{treatment} = 45.8$, $SS_{linear} = 41.7$, $SS_{total} = 111.4$, and that the relationship between your continuous independent variable and the dependent variable is **negative**. What is the value of Pearson *r*?
