Tests based on the t-test and Multiple range tests

One of the strengths of contrasts is that they should, in general, be made *a priori*. This means that the investigator is not conducting a wide search for anything that might be significant, but rather a directed set of tests against certain, predetermined contrasts. Although we know that each contrast has an α probability of error, most investigators feel comfortable doing several contrasts as long as (1) the overall F test in ANOVA indicates some significant effects and (2) the number of tests does not exceed the d.f. of the model.

However, investigators do not always have *a priori* contrasts. Some investigators are interested in detecting differences among the treatment levels and determining which treatments are different from which other treatments. Frequently, the interest is in doing all possible pairwise tests. Since these test do not involve *a priori* decisions I, will refer to them as Post-hoc, or Post-ANOVA, tests and the error rate will be an issue. At the very least, these techniques should not be done unless the ANOVA indicates that statistically meaningful, significant differences in treatments levels exist. Once you have found out some treatment(s) are "different", we will want to determine which one(s) are different?

Tests derived from the t-test

If we had done a t-test on the individual pairs of treatments, the test would have been done as $t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$. If the difference between $\bar{Y}_1 - \bar{Y}_2$ was large enough, the t value would have been greater than the $t_{critical}$ and we would conclude that there was a significant difference between the means. Since we know the value of $t_{critical}$ we could figure out how large a difference is needed for significance for any particular values of MSE, n_1 and n_2. We do this by replacing t with $t_{critical}$ and solving for $\bar{Y}_1 - \bar{Y}_2$.

$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, so

$$t_{critical}\sqrt{MSE\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \bar{Y}_1 - \bar{Y}_2 \quad \text{or} \quad \bar{Y}_1 - \bar{Y}_2 = t_{critical} S_{\bar{Y}_1 - \bar{Y}_2}$$

This value is the exact width of an interval $\bar{Y}_1 - \bar{Y}_2$ which would give a t-test equal to $t_{critical}$. Any larger values would be "significant" and any smaller values would not. This is called the "Least Significant Difference". $LSD = t_{critical} S_{\bar{Y}_1 - \bar{Y}_2}$

This least significant difference calculation can be used to either do pairwise tests on observed differences or to place a confidence interval on observed differences. However, since it is based on the t-test, each test is done with an α probability of error, that is a 5% probability of error on every test if $\alpha = 0.05$. Fisher suggests a "protection" against inflated error which was, if the ANOVA does not indicate significant differences don't use the LSD. This is called Fisher's protected LSD. Subsequent researchers have come up with other types of protection that are discussed below.

Note that the calculations above assume equal variance and use the ANOVA value of MSE (mean squared error) as the estimate of variance. This approach has been necessary historically, but PROC MIXED will fit separate variances and use them in these post-ANOVA calculations. The t-test on

unpooled variances uses $\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ as the standard error. The degrees of freedom are not known exactly for this t-test denominator. While SAS uses the Satterthwaithe adjustment for the t-test, use the "DDFM=Kenward-Rogers" in PROC MIXED.

Post-hoc tests

The LSD has an α probability of error on each and every test. The whole idea of ANOVA is to give a probability of error that is α for the whole experiment, so, much work in statistics has been dedicated to this problem. Some of the most common and popular alternatives are discussed below. Most of these are also discussed in your textbook.

The LSD is the LEAST conservative of those discussed, meaning it is the one most likely to detect a difference and it is also the one most likely to make a Type I error when it finds a difference. However, since it is unlikely to miss a difference that is real, it is also the most powerful. The probability distribution used to produce the LSD is the t distribution.

Bonferroni's adjustment: Bonferroni pointed out that in doing k tests, each at a probability of Type I error equal to α, the overall experimentwise probability of Type I error will be NO MORE than $k*\alpha$, where k is the number of tests. Therefore, if we do 7 tests, each at α=0.05, the overall rate of error will be NO MORE than =.35, or 35%. So, if we want to do 7 tests and keep an error rate of 5% overall, we can do each individual test at a rate of $\alpha/k = 0.05/7 = 0.007143$. For the 7 tests we have an overall rate of $7*0.007143 = 0.05$. The probability distribution used to produce the LSD is the t distribution.

Duncan's multiple range test: This test is intended to give groupings of means that are not significantly different among themselves. The error rate is for each group, and has sometimes been called a familywise error rate. This is done in a manner similar to Bonferroni, except the calculation used to calculate the error rate is $[1-(1-\alpha)^{r-1}]$ instead of the sum of α. For comparing two means that are r steps apart, where for adjacent means r=2. Two means separated by 3 other means would have r = 5, and the error rate would be $[1-(1-\alpha)^{r-1}] = [1-(1-0.05)^4] = 0.1855$. The value of a needed to keep an error rate of α is the reverse of this calculation, $[1-(1-0.05)^{1/4}] = 0.0127$.

For the Bonferroni with 7 tests the error rate for α=0.05 up to 0.35, and the adjusted value was 0.007143. The corresponding values for Duncan's adjustment are 0.301663 and 0.007301.

Tukey's adjustment The Tukey adjustment allows for **all possible pairwise tests**, which is often what an investigator wants to do. Tukey developed his own tables (see Appendix table A.5 in your book, "Selected percentiles of the studentized range distributions). For "t" treatments and a given error degrees of freedom the table will provide 10%, 5% and 1% error rates. These are experimentwise rates of Type I error.

Scheffé's adjustment This test is the most conservative. It allows the investigator to do not only all pairwise tests, but **all possible tests**, and still maintain an experimentwise error rate of α. "All possible" tests includes not only all pairwise tests, but comparisons of all possible combinations of treatments with other combinations of treatments (recall contrasts). The calculation is based on a square root of the F distribution, and can be used for range type tests or confidence intervals. The test is more general than the others mentioned. For the special case of pairwise comparisons, the statistic is $\sqrt{(t-1)}*F_{t-1,\, n(t-1)}$ for a balanced design with t treatments and n observations per treatment.

Place the post-hoc tests above in order from the one most likely to detect a difference (and the one most likely to be wrong) to the one least likely to detect a difference (and the one least likely to be wrong). **LSD is first, followed by Duncan's test, Tukey's and finally Scheffé's**. Dunnett's is a special test that is similar to Tukey's, but for a specific purpose, so it does not fit well in the ranking. The Bonferroni approach produces an upper bound on the error rate, so it can become excessively conservative for a given number of tests. It is a useful approach if you want to do a few tests, fewer than allowed by one of the others (e.g. you may want to do just a few and not all possible pairwise). In this case, the Bonferroni may be better.

Dunnett's adjustment – Dunnett's adjustment is a special adjustment intended only for testing **all treatments against one other treatment**, usually a control. For "t-1" treatments to be tested against the control the test gives an experimentwise rate of Type I error.

SAS is capable of applying these tests in both GLM and MIXED. We will concern ourselves only with PROC MIXED since GLM cannot handle unequal variance models and in some advanced designs does not correctly calculate the standard errors.

PROC MIXED estimates the treatment level means as "least squares means" in a statement called LSMEANS that follows the model statement. Least squares are not simple raw data means, but rather means of means or predicted values.

Pairwise comparisons can be specified with the PDIFF option and adjustments are specified as ADJUST=adjustment. If and adjustment is not specified an LSD is done. The available adjustments in PROC MIXED include: BON (Bonferroni), DUNNETT, SCHEFFE, and TUKEY.

See the SAS example.

Unfortunately, SAS does not currently express these results as a range test, but only as pairwise tests. A SAS macro is available for this purpose.

In the SAS examples a TUKEY adjustment is applied to the mouse diet and Spock judge examples. A new example on handicap discrimination is used to show several adjustments. DUNNETT's test was done separately, without Saxton's macro.