Although the $F$-test may allow us to declare that the means are significantly different, it tells us nothing about \textit{which} means are different. To decide which means are different, we need \textbf{multiple comparison procedures}, i.e., procedures for comparing all pairs of means. Note that there are $d = g (g-1)/2$ such pairs.

We can compare $\mu_j$ and $\mu_k$ for any \textit{particular} pair of groups $j$ and $k$ as follows. Since

\[ s^2 \sim \sigma^2 \frac{\chi^2_{n-g}}{(n-g)} \]  
and

\[ \bar{y}_j - \bar{y}_k \sim N(\mu_j - \mu_k, \sigma^2 (1/n_j + 1/n_k)) \]

independently of $s^2$, it follows that
\[ \bar{y}_{j} - \bar{y}_{k} \pm t_{\alpha/2; n-g} \sqrt{s \left[ \frac{1}{n_j} + \frac{1}{n_k} \right]} \]

is a level \( 1 - \alpha \) confidence interval for \( \mu_j - \mu_k \). If the interval does not include 0 then we reject \( H_0: \mu_j - \mu_k = 0 \) in favor of \( H_1: \mu_j - \mu_k \neq 0 \). This gives a valid level \( \alpha \) test for fixed \( j, k \). We say, therefore, that the per comparison error is \( \alpha \).

The multiple comparisons method which results from performing the above test for all pairs \( j, k \), is called the least significant difference (LSD) procedure. Unfortunately, the Type I experimentwise error for LSD (i.e., the probability of finding at least one significant difference among all \( d \) comparisons, if there are no real differences) may be
much larger than $\alpha$. For a rough calculation, we can pretend that the $d$ comparisons are independent. Then the Type I experimentwise error would be $1 - (1 - \alpha)^d$, which approaches 1 as $d$ increases. So if the number of groups is large, we will be almost guaranteed to find at least one significant difference, even if the group population means are all the same.

In Section 5.1.2, Jobson describes a variety of multiple comparisons procedures which (unlike LSD) can guarantee a Type I experimentwise error of $\alpha$. Perhaps the simplest of these is the Bonferroni method. Let $A_j\ (j = 1, \ldots, d)$ denote the event that the $j$’th pair of means are declared equal
in a test such as the \( t \)-test described above, assuming that they really are equal. The Type I experimentwise error rate is 
\[ P(\bar{A}_1 \cup \bar{A}_2 \cdots \cup \bar{A}_d), \]
where \( \bar{A}_j \) is the complement of \( A_j \). Using a basic property of probability (sub-additivity), we have
\[ P(\bar{A}_1 \cup \bar{A}_2 \cdots \cup \bar{A}_d) \leq P(\bar{A}_1) + \cdots + P(\bar{A}_d). \]

Therefore, if we arrange so that the Type I per-comparison error is \( P(\bar{A}_j) = \alpha / d \), then the Type I experimentwise error will be \( \leq \alpha \).

For example, if \( g = 5 \) and \( \alpha = .05 \), we should use \( t_{.0025; n-5} \) for each individual comparison, rather than \( t_{.025; n-5} \).
Tukey’s Method

Tukey’s method, designed for the case where the \( n_j \) are equal, yields in this case an experimentwise Type I error rate of exactly \( \alpha \). The method gives confidence intervals for all \( d \) differences \( \mu_j - \mu_k \). The intervals can then be used to simultaneously test the hypotheses that \( \mu_j - \mu_k = 0 \). The intervals are given by

\[
(\bar{y}_j - \bar{y}_k) \pm q_{\alpha; g, (n-g)} s/\sqrt{n_0},
\]

where \( n_0 \) is the (common) sample size of each group, and \( q_{\alpha; g, (n-g)} \) is the \((1-\alpha)(100)\)’th percentile of the Studentized Range Distribution, given in Table 5 of Jobson, p. 605.
To establish the properties of the method, we define the *studentized range* by

\[
q_{g,(n-g)} = \max_{1 \leq j < k \leq g} \frac{|\bar{y}_j - \bar{y}_k|}{s/\sqrt{n_0}}.
\]

The Studentized Range Distribution is defined to be the distribution of \(q_{g,(n-g)}\) under the assumption that the group means are all equal.

The experimentwise Type I error rate is the probability that at least one of the intervals in Equation (1) fails to contain \(\mu_j - \mu_k\). Note that for any particular interval, failure to contain \(\mu_j - \mu_k\) occurs if and only if either
\[ \mu_j - \mu_k < \bar{y}_j - \bar{y}_k - q \alpha; g, (n-g) s / \sqrt{n_0} \]

or

\[ \mu_j - \mu_k > \bar{y}_j - \bar{y}_k + q \alpha; g, (n-g) s / \sqrt{n_0} . \]

Thus, the interval fails if and only if

\[ |\bar{y}_j - \bar{y}_k - (\mu_j - \mu_k)| > q \alpha; g, (n-g) s / \sqrt{n_0} . \]

Thus,

\[ Pr \{ \text{Type I Experimentwise Error} \} = \]

\[ Pr \left[ \frac{|\bar{y}_j - \bar{y}_k - (\mu_j - \mu_k)|}{s / \sqrt{n_0}} > q \alpha; g, (n-g) \text{ for some } j, k \right] \]

\[ = Pr \left[ \max_{1 \leq j < k \leq g} \frac{|\bar{y}_j - \bar{y}_k - (\mu_j - \mu_k)|}{s / \sqrt{n_0}} > q \alpha; g, (n-g) \right] = \alpha , \]

since the random variable in the above equation has,
by definition, the studentized range distribution with
degrees of freedom $g$ and $n-g$.

If the $n_j$ are unequal, we replace $s/\sqrt{n_0}$ in the
interval (1) by

$$s \left[ .5/n_j + .5/n_k \right]^{1/2}.$$  

This reduces to $s/\sqrt{n_0}$ when the group sizes are
equal. It can be shown that this modified version of
Tukey’s method is conservative, i.e., that the result-
ing Type I experimentwise error rate is $\leq \alpha$.

Scheffe’s Method

The Bonferroni intervals are often too wide to be
useful, especially if $g$ is large. Scheffe’s method
provides reasonably narrow confidence intervals for
all differences $\mu_j - \mu_k$ with a Type I experiment-wise error of (at most) $\alpha$. In fact, Scheffe’s method does more than this.

The difference $\mu_j - \mu_k$ is an example of a contrast. A contrast is a linear combination of the means, $L = \sum_{j=1}^{g} \lambda_j \mu_j$ where $\sum_{j=1}^{g} \lambda_j = 0$.

Scheffe’s method gives us confidence intervals for all possible contrasts, with a Type I experimentwise error of $\alpha$.

Thus, it can be safely used for a contrast which has been identified after examining the data!
In the milk example, examination of the boxplots suggests that perhaps $\mu_1 = \mu_2 = \mu_3$, but $\mu_4$ is different from the other means. We then might want to test whether the contrast $L = \mu_4 - \frac{1}{3}(\mu_1 + \mu_2 + \mu_3)$ is zero or not.

Scheffe’s interval for the contrast $L_\lambda$ is

$$I_\lambda = \sum_{j=1}^{g} \lambda_j \bar{y}_j \pm s \left\{ (g-1)F_{\alpha; (g-1), (n-g)} \sum_{j=1}^{g} \lambda_j^2/n_j \right\}^{1/2}.$$  

The intervals $I_\lambda$ cover their corresponding contrasts $L_\lambda$ with an experimentwise error of $\alpha$, as stated in the following theorem.

**Theorem:**

$$Pr \{I_\lambda \text{ contains } L_\lambda \text{ for all contrasts } \lambda \} = 1 - \alpha.$$
Proof: Define

\[ \hat{L}_\lambda = \sum_{j=1}^{g} \lambda_j \bar{y}_j, \]

\[ K = s \left[ (g-1) F_{\alpha; (g-1), (n-g)} \right]^{1/2}, \]

\[ \delta_\lambda = \left[ \sum_{j=1}^{g} \frac{\lambda_j^2}{n_j} \right]^{1/2}, \]

\[ w_j = \bar{y}_j - \bar{y} - (\mu_j - \mu), \]

where

\[ \mu = \frac{1}{n} \sum_{j=1}^{g} n_j \mu_j. \]

Note that if all the \( \mu_j \) are the same, then

\[ \sum n_j w_j^2 = \sum n_j [\bar{y}_j - \bar{y}]^2 = SSA, \]

which is distributed as \( \sigma^2 \chi^2_{g-1} \). In fact, \( \sum n_j w_j^2 \sim \sigma^2 \chi^2_{g-1} \) (independently of \( s^2 \)) regardless of the values of \( \mu_j \). The reason:
\[ \sum n_j w_j^2 \] is the SSA for an ANOVA model in which \( \{y_{ij}\} \) is replaced by \( \{y_{ij} - \mu_j\} \), for which the population group means are all equal.

We therefore find that

\[
P_r \{ I_\lambda \text{ contains } L_\lambda \text{ for all contrasts } \lambda \} 
\]

\[
= P_r \{ \hat{L}_\lambda - K \delta_\lambda \leq L_\lambda \leq \hat{L}_\lambda + K \delta_\lambda \text{ for all contrasts } \lambda \} 
\]

\[
= P_r \{ |\hat{L}_\lambda - L_\lambda|/\delta_\lambda \leq K \text{ for all contrasts } \lambda \} 
\]

\[
= P_r \{ \max_\lambda |\hat{L}_\lambda - L_\lambda|/\delta_\lambda \leq K \} 
\]

Since \( \sum_{j=1}^{g} \lambda_j = 0 \), and using the Cauchy-Schwarz inequality, we find that
\[ |\hat{L}_\lambda - L_\lambda|/\delta_\lambda = | \sum_{j=1}^{g} \lambda_j (\bar{y}_{..} - \mu_j) |/\delta_\lambda \]

\[ = | \sum_{j=1}^{g} \lambda_j [\bar{y}_{.j} - \mu_j - (\bar{y}_{..} - \mu_{..})] |/\delta_\lambda \]

\[ = | \sum_{j=1}^{g} \lambda_j w_j |/\delta_\lambda = | \sum_{j=1}^{g} \lambda_j /\sqrt{n_j} \sqrt{n_j w_j} |/\delta_\lambda \]

\[ \leq \sqrt{\sum \lambda_j^2 /n_j} \sqrt{\sum n_j w_j^2} / \delta_\lambda = \sqrt{\sum n_j w_j^2} . \]

Equality is achieved for the contrast

\[ \lambda = (n_1 w_1, \ldots, n_g w_g)' . \] Thus,

\[ \max_\lambda |\hat{L}_\lambda - L_\lambda|/\delta_\lambda = \sqrt{\sum n_j w_j^2} . \]

We conclude that

\[ Pr \{I_\lambda \text{ contains } L_\lambda \text{ for all contrasts } \lambda \} \]

\[ = Pr \{ \sqrt{\sum n_j w_j^2} \leq K \} \]
\[- 14 -
\]

\[= Pr \{ \sum n_j w_j^2 \leq K^2 \} \]

\[= Pr \left\{ \frac{\sum n_j w_j^2 / (g-1)}{s^2} \leq F_{\alpha; (g-1),(n-g)} \right\} = 1 - \alpha. \]

**Comparisons for Milk Example**

We ran the Minitab oneway ANOVA with (multiple) comparisons. We used Fisher’s [LSD] with an individual error rate of 5 percent and Tukey’s with a family error rate of 5 percent. To get Bonferroni multiple comparisons, we re-ran Fisher’s comparisons with an individual error rate of \(.05/d\) \(= .05/6 = .00833\). The critical values are \(t_{\alpha/2; n-g} = t_{.025; 116} = 1.981\) for Fisher’s [LSD], and \(t_{.00833/2; 116} = 2.684\) for Bonferroni’s, and
\[ q_{\alpha; g, (n-g)} = q_{.05; 4, 116} = 3.69 \] for Tukey’s Method.

In each case, the output gives confidence intervals for \( \mu_k - \mu_j \). A pair of means is significantly different (at level .05) if the corresponding confidence interval does not include zero.

For the LSD method, the significant pairwise differences are those between (AD1,AD4), (AD2,AD4), and (AD3,AD4). But LSD makes no adjustments for multiple comparisons.

The Bonferroni method, which does make such an adjustment, gives a critical value of 2.684 (so the intervals are wider than for LSD), and the only significant pairwise difference is between
Tukey’s method reaches the same conclusion as Bonferroni’s method, that (AD3, AD4) is the only significant difference.

The Scheffe method can be used for any contrast, but is not immediately available in Minitab. The value of $F_{\alpha; (g-1), (n-g)}$ needed here is $F_{.05; 3, 116} = 2.683$, as found either from Table 4, Jobson p. 602, or from Minitab using Calc $\rightarrow$ Probability Distributions $\rightarrow$ F $\rightarrow$ Inverse cumulative probability, Numerator Degrees of Freedom: 3, Denominator Degrees of Freedom: 116, Input Constant: .95.
In order to be significantly different under Scheffe’s method, a pair of means will need to differ by at least $22.01 \sqrt{3(2.683)(2/30)} = 16.123$, as compared to $22.01(1.981)\sqrt{2/30} = 11.258$ for LSD, and $22.01(2.684)\sqrt{2/30} = 15.253$ for Bonferroni, and $3.69(22.01)/\sqrt{30} = 14.83$ for Tukey’s method.

There are no significant pairwise differences using Scheffe. But what about the contrast

$$\mu_4 - 1/3(\mu_1 + \mu_2 + \mu_3)$$

which seemed interesting from the boxplots? (Note that data snooping is OK when we use the Scheffe method.)
We get

\[ s \left[ (g-1) F_{\alpha; (g-1), (n-g) \sum \frac{\lambda_j^2}{n_j}} \right]^{1/2} \]

\[ = 22.01 \sqrt{3(2.683)(1/30)(1+3/9)} = 13.164 \, . \]

The observed contrast is

\[ \hat{L}_\lambda = 51.03 - \frac{1}{3}(37.24 + 39.65 + 34.95) = 13.75 \, , \]

which is significantly different from zero, even though none of the pairwise differences was significant.