MULTIPLE COMPARISONS (Section 4.4)

1. **Bonferroni Method**. We have seen: If we form two 95%confidence intervals for two means or two effect differences, etc., then the probability that, under repeated sampling with the same design, the procedures used will give intervals *each* containing the true mean, effect differences, etc. might only be 90% -- we have no reason to believe it must be any higher without any more information. Thus the *simultaneous* or *family-wise* or *overall* confidence level is 90%
Similarly, if we are forming m confidence intervals, each with confidence level 1- α individually, then analogous probability calculations show that the *simultaneous* or *family-wise* or *overall* or *experiment-wise* confidence level will be only 1-mα . This gives us one method of deciding what individual confidence levels to choose if we want a given overall confidence level: If we want overall level 1- α, then choose individual level 1-α/m. For example, if we are forming 5 confidence intervals and want an overall 95% confidence level, then we need to use the procedure for individual 99% confidence intervals. This method of forming simultaneous confidence intervals is called the **Bonferroni** method. It gives wide intervals.

Example: In the battery experiment, the individual 95% confidence intervals for the four means shown in the Minitab output have a Bonferroni overall confidence level 80%. If we want an overall confidence level 95% for the four confidence intervals, we need to calculate individual 98.75% confidence intervals. These would have standard error $\sqrt{\frac{msE}{r_i}} = \sqrt{\frac{2368}{4}}$ = 24.33 and use t-value t(12, .99375) = 2.9345, giving confidence intervals of half-width 71.40, in contrast to the half-width 24.33x2.1254 = 51.71 for the individual 95% confidence intervals -- more than a third as wide.

The phenomenon shown in the example is typical: To get a certain family confidence level, you will get wider confidence intervals than those formed with the individual confidence level.
A Bonferrroni approach can also be used for hypothesis tests: If you want to do m hypothesis tests on your data, and you want an overall type I error rate of α (that is, you want to have probability of falsely rejecting at least one of the null hypotheses less than α), you can achieve this by using a significance level of α/m for each test individually.

Example: Suppose the experimenter in the battery example collected the data, analyzed them, looked at the confidence intervals in the Minitab output, noticed that the estimate of the mean for the second level was largest and the estimate for the first level the second largest, and tested the null hypothesis H_0: $\mu_1 = \mu_2$. For what p-values should he reject the null hypothesis using the Bonferroni method in order to claim his result is significant at the .05 level?

This brings up the concepts of *pre-planned comparisons* and *data snooping*.

A ***pre-planned comparison*** is one identified before running the experiment. The experiment should be designed so that the items to be estimated are estimable and their variance is as small as possible.

Data-snooping occurs when you look at your data after the experiment has been performed and decide something looks interesting, then do a test on it. There is nothing wrong with data-snooping -- often interesting results are found this way. But data-snooping tests need to be done with care to obtain an honest significance level. The problem is that they usually are the result of several comparisons, not just the one formally tested. So if, for example, a Bonferroni procedure is used, you need to take into account all the other comparisons that are done informally in setting a significance level.

Summary of utility of Bonferroni methods*:*

- Not recommended for data snooping -- it's too easy to overlook comparisons that were made in deciding what to test.
- OK for pre-planned comparisons when m is small.
- Not useful when m is large -- too conservative (i.e., CI's tend to be larger than absolutely necessary, and the Type II error rate tends to be larger than necessary.

Comments:

- In regression, interest is often in model building rather than estimating parameters or establishing causality, so not as much attention is paid to multiple inference. (To have confidence in a model, one needs to validate it on an independent data set.) However, there are situations in regression where considerations of multiple inference are important (for example, if you are trying to estimate more than one parameter in a regression equation). In these situations, Bonferroni methods can be used, but there are also "confidence regions" in parameter space that usually give tighter results.
- Unfortunately, many users of statistics are not aware of the problems with multiple inference – so results are reported as statistically significant when they aren't.

2. **General Comments on Methods for Multiple Comparisons**. There are many other methods for multiple comparison. All of the methods well will discuss produce, like the Bonferroni method, confidence intervals with endpoints of the form

$$\hat{C} \pm \text{w se}(\hat{C}),$$

where C is the contrast or other parameter being estimated, $\hat{C}$ is the least squares estimate of C, se($\hat{C}$) is the standard error of $\hat{C}$, and w (called the *critical coefficient*) depends on the overall confidence level α, on the method, on the number v of treatments, on the number m of things being estimated, and on the number of error degrees of freedom. For the Bonferroni method, $w = w_B = t(n-v, \alpha/(2m))$

The half-width w se($\hat{C}$) of the confidence interval is called the *minimum significant difference* (*msd* for short), because it is the smallest value of $\hat{C}$ that will produce a confidence interval not containing 0, and hence say the contrast is significantly different from zero.

3. **Scheffe Method**. This method does not depend on the number of comparisons being made, but applies to contrasts only. The idea behind the method is that every contrast can be written as a linear combination of the v-1 "treatment vs control" contrasts $\tau_2 - \tau_1, \tau_3 - \tau_1, \ldots, \tau_{v-1} - \tau_1$. The method depends on finding a 1- α *confidence region* for these v-1 contrasts, and showing that this confidence region for these special contrasts determines confidence bounds for every possible contrast that are independent of the number of contrasts.

Summary of utility of Scheffe method:

- Does not matter how many comparisons are made, so suitable for data snooping.
- If m is large, gives shorter confidence intervals than Bonferroni.
- For m small, is "expensive insurance."

Since Minitab 15 does not give the Scheffe method, we will not use it in this class.

4. **Tukey Method for All Pairwise Comparisons**. As the name suggests, this is used for all *pairwise* contrasts $\tau_i - \tau_j$. This is sometimes called the Honest Significant Difference Method, since (for equal sample sizes) it depends on the distribution of the statistic

$$Q = \frac{\max\{T_1,\ldots,T_v\} - \min\{T_1,\ldots,T_v\}}{\sqrt{MSE/r}},$$

where $T_i = \overline{Y}_{i\cdot} - \mu_i$. This distribution is called the *Studentized range distribution*. Like the F distribution it has two degrees of freedom. The critical coefficient is $w_T = q(v, n-v, \alpha)/\sqrt{2}$. For equal sample sizes, the overall confidence level is 1-α; for unequal sample sizes, it is at least
1-α.

Note that since this method only deals with pairwise contrasts, the standard error of $\tau_i - \tau_j$ involved in the calculation of the msd is just $\sqrt{msE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$

Summary of utility of Tukey method:

- Usually gives shorter confidence intervals than either Bonferroni or Scheffe.
- In basic form can be used only for pairwise comparisons. (There is an extension to all contrasts, but it is usually not as good as Scheffe.)

Example: Battery Experiment.

5. **Dunnett Method for Treatment-Versus-Control Comparisons.** This is even more specialized than the Tukey method, but the special situation is often of interest. If we assume Treatment 1 is a control, then we are likely to be interested in the treatment-

versus-control contrasts $\tau_i - \tau_1$. This method is based on the joint distribution (a type of multivariate t-distribution) of the estimators $\overline{Y}_{i\cdot} - \overline{Y}_{1\cdot}$. Because the distribution is complicated, the calculation of w_D is best left to reliable software. However, not all software (e.g., Minitab) gives one-sided confidence intervals, which might be desired.

Summary of utility of Dunnett method:
- Best method for treatment-versus-control comparisons.
- Not applicable to other types of comparisons.

Example: Battery experiment.

6. **Hsu's Method for Multiple Comparisons with the Best Treatment**. The idea is similar to Dunnett's method, but instead of comparing each treatment with the control group, each treatment is compared with the best of the other treatments. The procedure varies slightly depending on whether "best" is largest or smallest. Minitab allows the user to check which is desired.

Summary of utility of Hsu method:
- Good for what it does.
- Not applicable to other types of comparisons.
- See p. 90 of textbook for details.

Example: Battery experiment.

7. **Other Methods**. There are many. Books have been written on the subject (e.g., Miller, Hsu). Some people have their favorites, which others argue are not good choices.

8. **Combinations of Methods**. There are various possibilities. See p.91 for some. The idea is to split α between the methods, analogous to the Bonferroni procedure.

Example: If the experiment is intended to test treatment vs control, use Dunnett with (overall) $\alpha = .02$ for that and Tukey or Hsu or Scheffe at overall $\alpha = .03$ for other things of interest that arise.