MULTIPLE COMPARISONS   (Section 4.4)

1. **Bonferroni Method**.

Last time: 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.

i.e., the *simultaneous* or *family-wise* or *overall* confidence level is 90%

Analogous calculations show: If we are forming m confidence intervals, each with confidence level 1- α individually, then the *simultaneous* or *family-wise* or *overall* or *experiment-wise* confidence level will be only 1-mα.

Consequence: If we want overall level 1- α, then choose individual level 1-α/m.

This is called the **Bonferroni** method.

e.g., 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.

Bonferroni typically 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:

$$\text{se} = \sqrt{\frac{msE}{r_i}} = \sqrt{\frac{2368}{4}} = 24.33 \text{ and use}$$
t-value t(12, .99375) = 2.9345

Result: Confidence intervals have half-width 71.40 -- compare with 24.33x2.1254 = 51.71 for the individual 95% confidence intervals -- more than a third as wide.

This illustrates the reality: 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?

Pre-planned comparisons and data snooping

A ***pre-planned comparison***: Identified *before* running the experiment.

  The experiment should be designed so that items to be estimated are estimable and their variance is as small as possible.

Data-snooping**:** Looking at your data after the experiment has been performed, deciding something looks interesting, then doing a test on it.

• There's nothing inherently 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 (CI's may be too large; type II error too large)

Comment: In regression:

- Interest often in model building, not estimation or establishing causality.

- Thus less attention to multiple inference. (But model validation, using another data set, is important.)

- Some uses of regression do require attention to multiple inference (e.g., estimating more than one parameter in a regression equation).

- Bonferroni methods can be used in regression.

- "Confidence regions" in parameter space usually give tighter results.

- Unfortunately, many users of statistics aren't aware of problems with multiple comparisons.

2. General Comments on Methods for Multiple Comparisons.

- There are many methods for multiple comparison.

- All the methods that we will discuss produce confidence intervals with endpoints of the form
$$\hat{C} \pm w \; se(\hat{C}),$$
where:

  o C is the contrast or other parameter being estimated

  o $\hat{C}$ is the least squares estimate of C

  o $se(\hat{C})$ is the standard error of $\hat{C}$

  o w (the *critical coefficient*) depends on the overall confidence level α, the method, the number v of treatments, the number m of things being estimated, and on the number of error degrees of freedom.

For Bonferroni, $w = w_B = t(n\text{-}v, \alpha/(2m))$

Note: The half-width w $se(\hat{C})$ of the confidence interval is called the *minimum significant difference* (*msd*) -- 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.

- Does not depend on the number of comparisons being made

- Applies to contrasts only.

The idea:

- All contrasts are linear combination of the v-1 "treatment vs control" contrasts $\tau_2 - \tau_1, \ldots, \tau_{v\text{-}1} - \tau_1$.

- A 1- α *confidence region* for these v-1 contrasts is formed.

- This confidence region for these special contrasts determines confidence bounds for every possible contrast, independently of the number of contrasts.

Summary of utility of Scheffe method:

• Does not matter how many comparisons are made, so suitable for data snooping.

• For large m, gives shorter confidence intervals than Bonferroni.

• For m small, is "expensive insurance."

Note: Minitab 15 does not give the Scheffe method, so we won't use it in this class.

4. **Tukey Method for All Pairwise Comparisons**.

• Used for all *pairwise* contrasts τ_i - τ_j.

• Also called the Honest Significant Difference Method, since (for equal sample sizes) it depends on the distribution of the statistic

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

where $T_i = \bar{Y}_{i.} - \mu_i$. This distribution is called the *Studentized range distribution*. Like the F distribution, it has two degrees of freedom.

Critical coefficient: $w_T = q(v, n-v, \alpha)/\sqrt{2}$.

For equal sample sizes, the overall confidence level is $1-\alpha$; for unequal sample sizes, it is at least $1-\alpha$.

Note: Since this method only deals with pairwise contrasts, the standard error of τ_i - τ_j needed 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.**

• If Treatment 1 is a control, then we are likely to be interested in the treatment-versus-control contrasts $\tau_i - \tau_1$.

• Method is based on the joint distribution of the estimators $\bar{Y}_{i\bullet} - \bar{Y}_{1\bullet}$ (a type of multivariate t-distribution).

• Because the distribution is complicated, the calculation of w_D is best left to reliable software.

• 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**.

- Instead of comparing each treatment with the control group, each treatment is compared with the best of the other treatments.

- Procedure varies slightly depending on whether "best" is largest or smallest. (Minitab allows the user to check which is desired.)

- See p. 90 of textbook for details.

Summary of utility of Hsu method:

- Good for what it does.

- Not applicable to other types of comparisons.

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**.

- Various possibilities.
     See p.91 for some.

- The idea: 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.

- Use Tukey or Hsu or Scheffe at overall $\alpha = .03$ for other things of interest that arise.