

# Multiple Comparisons & Type-I Error

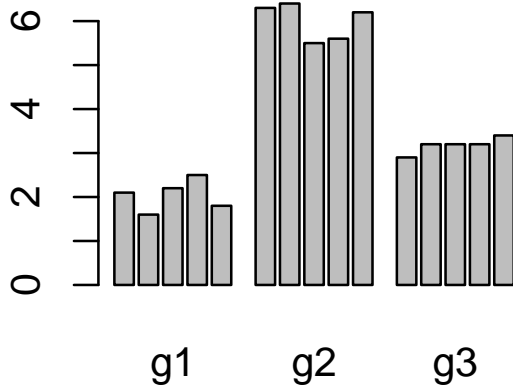
Paul Gribble

Winter, 2019

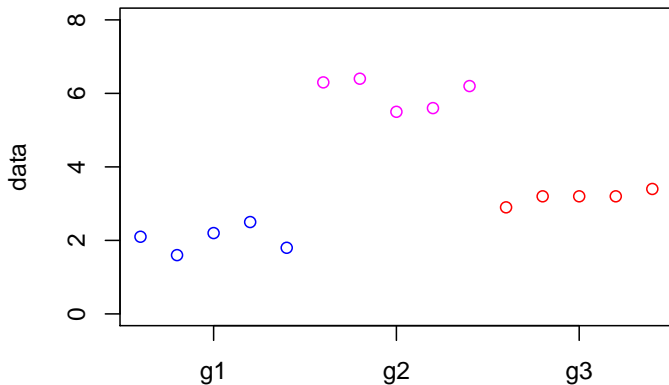
## GLM & ANOVA: an example

G1	G2	G3
2.1	6.3	2.9
1.6	6.4	3.2
2.2	5.5	3.2
2.5	5.6	3.2
1.8	6.2	3.4
means		
2.0	6.0	3.2

# GLM & ANOVA: an example

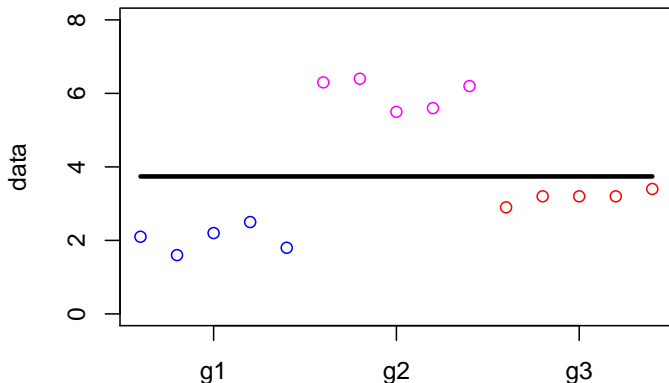


# the model comparison approach: restricted model



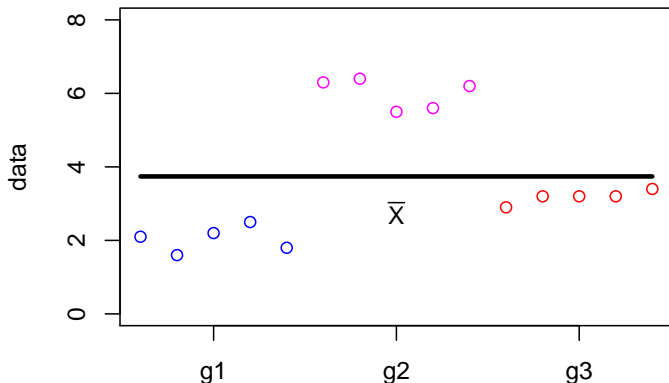
$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \quad E_r = \sum (Y_{ij} - \bar{X})^2$$

# the model comparison approach: restricted model



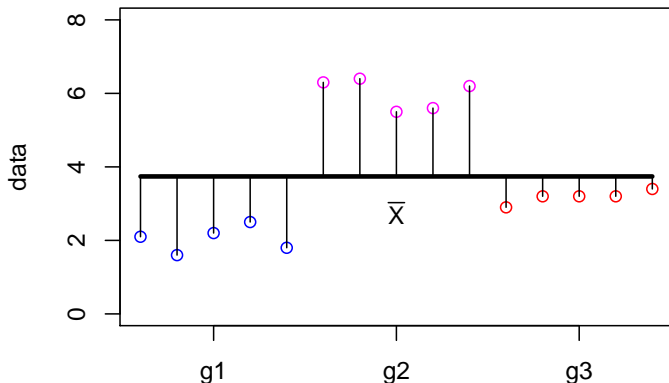
$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \quad E_r = \sum (Y_{ij} - \bar{X})^2$$

# the model comparison approach: restricted model



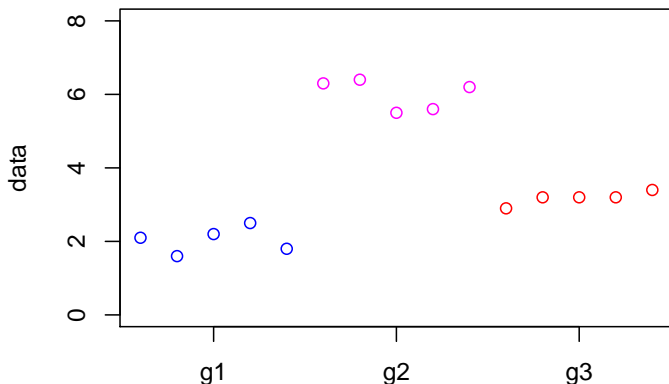
$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \quad E_r = \sum (Y_{ij} - \bar{X})^2$$

# the model comparison approach: restricted model



$$H_0 : Y_{ij} = \mu + \epsilon_{ij} \quad E_r = \sum (Y_{ij} - \bar{X})^2$$

## the model comparison approach: full model

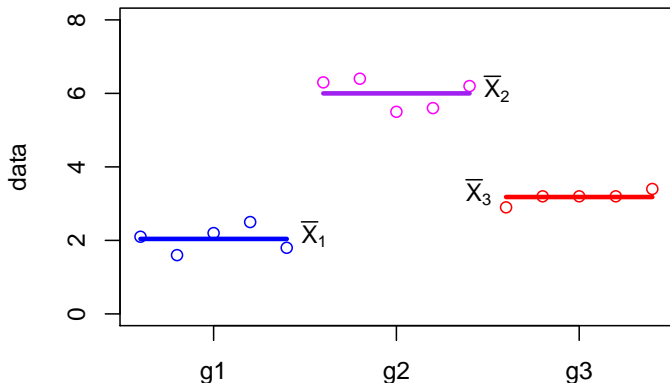


$$H_1 : Y_{ij} = \mu_j + \epsilon_{ij}$$

$$E_f = \sum (Y_{ij} - \bar{X}_j)^2$$



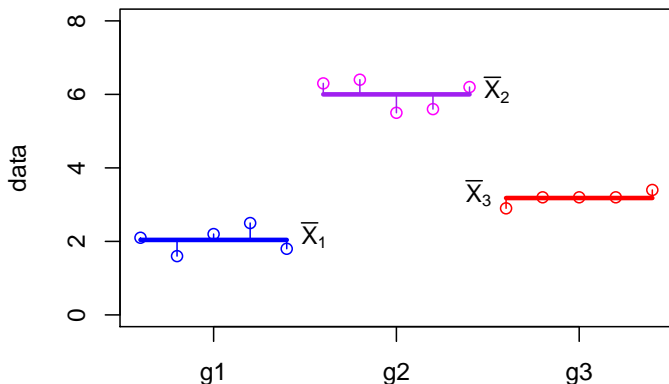
## the model comparison approach: full model



$$H_1 : Y_{ij} = \mu_j + \epsilon_{ij}$$

$$E_f = \sum (Y_{ij} - \bar{X}_j)^2$$

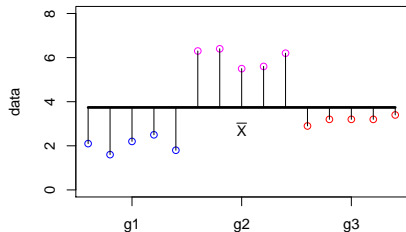
# the model comparison approach: full model



$$H_1 : Y_{ij} = \mu_j + \epsilon_{ij}$$

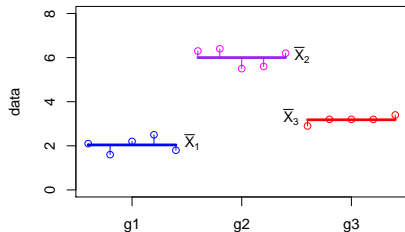
$$E_f = \sum (Y_{ij} - \bar{X}_j)^2$$

# which model has smaller error?



▶ estimate 1 parameter

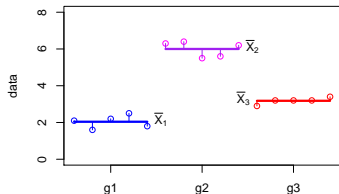
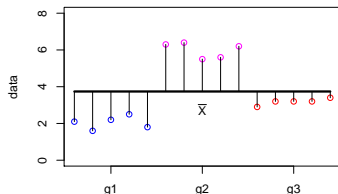
▶  $\mu$



▶ estimate 3 parameters

▶  $\mu_1, \mu_2, \mu_3$

# which model has smaller error?



- ▶ Is the reduction in error you get with the full model **worth** the extra parameters you need to estimate in  $H_1$ ?

# Testing differences between individual means

- ▶ last time we learned about one-way single-factor ANOVA
- ▶ F test of null hypothesis
  - ▶  $\mu_1 = \mu_2 = \dots = \mu_n$
- ▶ called the "omnibus test"
- ▶ omnibus test doesn't tell us *which* means are different from each other
- ▶ it *does* give us permission to start looking for differences between individual means

# Two kinds of multiple comparisons

## planned comparisons

- ▶ **in advance of looking at your results** you know which groups you want to compare
- ▶ you are restricted to performing only certain comparisons
- ▶ the comparisons must be *orthogonal* to each other

## post-hoc comparisons

- ▶ **the results dictate which means you test** (you are *chasing the biggest differences*)
- ▶ you can test as many as you like (usually)
- ▶ few (if any) restrictions on the nature of the tests you can perform
- ▶ Type-I error is controlled for by making each test more conservative

# Model comparison approach

- ▶ recall the null hypothesis & restricted model:

$$H_0 \quad : \quad \mu_1 = \mu_2 = \cdots = \mu_a$$

$$Y_{ij} = \mu + \epsilon_{ij}$$

- ▶ suppose we wanted to test a new hypothesis that only groups 1 and 2 are equal and the rest are different

$$H_0 \quad : \quad \mu_1 = \mu_2$$

$$Y_{i1} = \mu^* + \epsilon_{i1}$$

$$Y_{i2} = \mu^* + \epsilon_{i2}$$

$$Y_{ij} = \mu_j + \epsilon_{ij}, \quad \text{for } j = 3, 4, \dots, a$$

## Model comparison approach

- ▶ just as before we can compare full and restricted models by computing sums of squared errors for each (see Maxwell & Delaney for details)
- ▶ just as before we end up with an F ratio:

$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F}$$

$$E_R - E_F = \frac{n_1 n_2}{n_1 + n_2} (\bar{Y}_1 - \bar{Y}_2)^2$$

$$df_F = N - a$$

$$df_R = N - (a - 1) = N - a + 1$$

$$df_R - df_F = 1$$



# Model comparison approach

- ▶ after some more tedious algebra:

$$F = \frac{n_1 n_2 (\bar{Y}_1 - \bar{Y}_2)^2}{(n_1 + n_2) MS_W}$$

- ▶ or for equal group sizes  $n$ :

$$F = \frac{n (\bar{Y}_1 - \bar{Y}_2)^2}{2MS_W}$$

- ▶  $MS_W$  is mean-square "within" term (error term) from ANOVA output
- ▶  $df$  numerator = 1
- ▶  $df$  denominator is given in ANOVA output for  $MS_W$  term

# Model comparison approach

- ▶ so what we have now is an F test for a full versus restricted model
- ▶ full model is as before (different mean for each group)
- ▶ restricted model has same mean for groups 1 and 2, and different means for the rest
- ▶ restricted model is less restricted than the original restricted model with a single parameter (the grand mean)
- ▶ but still more restricted than full model

$$F = \frac{n(\bar{Y}_1 - \bar{Y}_2)^2}{2MS_W}$$

# Complex comparisons

- ▶ research questions often focus on pairwise comparisons
- ▶ sometimes you may have a hypothesis that concerns a difference involving more than 2 means
- ▶ e.g. 4 groups: is group 4 different than the average of the other three?

$$H_0 : \frac{1}{3} (\mu_1 + \mu_2 + \mu_3) = \mu_4$$

- ▶ we can rewrite this as:

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0$$

## Complex comparisons

$$H_0 : \frac{1}{3}\mu_1 + \frac{1}{3}\mu_2 + \frac{1}{3}\mu_3 - \mu_4 = 0$$

- ▶ this is just a linear combination of the 4 means so in general we can write:

$$H_0 : c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4 = 0$$

- ▶  $c_1$  through  $c_4$  are coefficients chosen by the experimenter to test a hypothesis of interest
- ▶ simple pairwise comparison of mean 1 vs mean 2 would be:

$$c_1 = -1$$

$$c_2 = +1$$

$$c_3 = 0$$

$$c_4 = 0$$

# Complex comparisons

an expression of the form:

$$H_0 : c_1\mu_1 + c_2\mu_2 + c_3\mu_3 + c_4\mu_4$$

is known as a "contrast" or a "complex comparison"

- ▶ linear combination of means in which *the coefficients add up to zero*
- ▶ in the general case of  $a$  groups, we can write:

$$\psi = \sum_{j=1}^a c_j \mu_j$$

# Complex comparisons

- ▶ our expression for the F test can be simplified (see M&D) to:

$$F = \frac{\psi^2}{MS_W \sum_{j=1}^a (c_j^2/n_j)}$$

where

- ▶  $df$  numerator = 1
- ▶  $df$  denominator =  $N - a$

$$H_0 : \psi = \sum_{j=1}^a c_j \mu_j = 0$$

# Complex comparisons

- ▶ some texts present contrasts not as F tests but as t-test
- ▶ when  $df$  numerator = 1, t-test is just a special case of the F-test

$$t^2 = F$$
$$t = \sqrt{F}$$

# Testing more than one contrast

- ▶ how many contrasts can we test?
- ▶ two issues:
  1. orthogonality
  2. inflation of Type-I error
- ▶ is it permissible to perform multiple tests using an  $\alpha$  level of 0.05?
  - ▶ better question: does it make sense to perform multiple tests and still assume that Type-I error rate remains at 0.05?
- ▶ does it matter if the contrasts were planned before the data were examined, or arrived at after looking at the data?



# How many contrasts?

- ▶ if  $a = 3$  there are 3 possible pairwise contrasts  
(choose(3,2))
  - ▶ 1-2, 2-3 and 1-3
  - ▶ in addition there are an infinite # of possible complex comparisons
- ▶ with an infinite # contrasts, some information will be redundant
- ▶ new question: how many contrasts can be tested without introducing redundancy?

# Non-redundant contrasts

- ▶ are these three contrasts redundant?

$$\psi_1 = \mu_1 - \mu_2$$

$$\psi_2 = \mu_1 - \mu_3$$

$$\psi_3 = \frac{1}{2}(\mu_1 + \mu_2) - \mu_3$$

- ▶ **yes**, because:

$$\psi_3 = \psi_2 - \frac{1}{2}\psi_1$$

- ▶ value of  $\psi_3$  is completely determined if we already know  $\psi_1$  and  $\psi_2$

# Non-redundant contrasts

- ▶ in general with  $a$  groups, there are  $a - 1$  contrasts without introducing redundancy
- ▶ mathematical concept for lack of redundancy is **orthogonality**
- ▶ two contrasts are **orthogonal** if:

$$\begin{aligned}\psi_1 &= \sum c_{1j}\mu_j \\ \psi_2 &= \sum c_{2j}\mu_j \\ \sum c_{1j}c_{2j} &= 0\end{aligned}$$

- ▶ or for unequal group sizes:

$$\sum c_{1j}c_{2j}/n_j = 0$$

# Orthogonal contrasts

- ▶ e.g. what about 2 contrasts  $c_1$  and  $c_2$ :
- ▶  $c_{11} = +1$ ,  $c_{12} = -1$ ,  $c_{13} = 0$
- ▶  $c_{21} = +1$ ,  $c_{22} = 0$ ,  $c_{23} = -1$
- ▶ orthogonality test:  $\sum c_{1j}c_{2j} = 0$ 
  - ▶  $(1)(1) + (-1)(0) + (0)(-1) = 1 + 0 + 0 = 1$
  - ▶ these 2 contrasts are **not** orthogonal

# Orthogonality

- ▶ who cares?
- ▶ primary implication: orthogonal contrasts provide non-overlapping information about how the groups differ
- ▶ formally: when two contrasts are orthogonal, then the two sample estimates  $\psi_1$  and  $\psi_2$  are statistically independent of one another
- ▶ each provides unique, non-overlapping information about group differences
- ▶ they are asking separate, different, distinct questions about the data

# Testing multiple comparisons

- ▶ suppose you have conducted an ANOVA on 4 groups
- ▶ suppose you want to test the following 3 contrasts:

$$\psi_1 = \mu_1 - \mu_2$$

$$\psi_2 = \frac{1}{2}(\mu_1 + \mu_2) - \mu_3$$

$$\psi_3 = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \mu_4$$

- ▶ are these orthogonal?
  - ▶  $\psi_1$ : (+1.0)(-1.0)(+0.0)(+0.0)
  - ▶  $\psi_2$ : (+0.5)(+0.5)(-1.0)(+0.0)
  - ▶  $\psi_3$ : (+0.3)(+0.3)(+0.3)(-1.0)

# Testing multiple comparisons

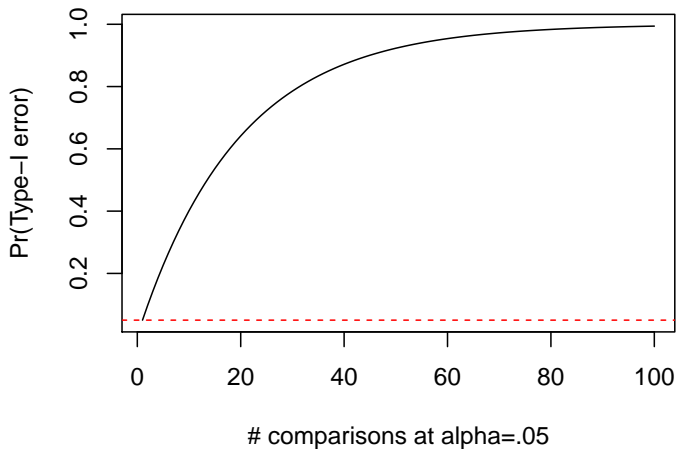
- ▶ if you test each of the three contrasts at  $\alpha = 0.05$ , what is the true Type-I error rate?
- ▶ greater than 0.05
- ▶ we are testing three contrasts **each** at the 0.05 level
- ▶ at first glance you might think true error rate should be  $(3)(0.05) = 0.15$
- ▶ close, but not quite right

# Testing multiple comparisons

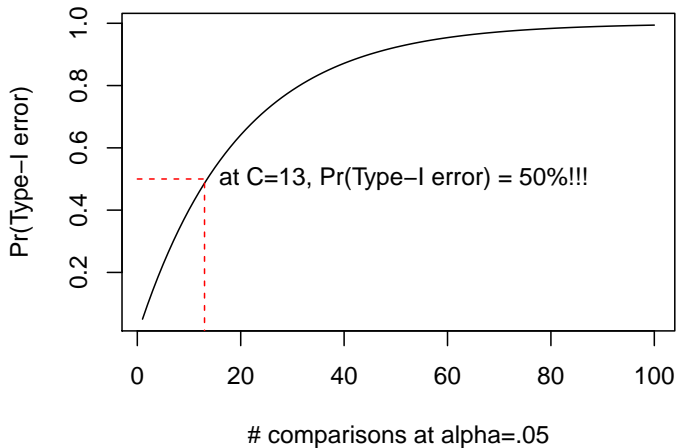
- ▶ contrasts are independent events
- ▶ probabilities don't simply sum (see M&D text)
- ▶  $\Pr(\text{at least one Type-I error}) = 1 - \Pr(\text{no Type-I errors})$
- ▶  $= 1 - (1 - \alpha)^C$
- ▶  $C$  is number of contrasts tested
- ▶ e.g. if  $\alpha = 0.05$ ,  $C = 3$ , then  $p = 0.143$
- ▶ if  $C = 10$ ,  $p = 0.40$  (**big!**)



# Testing multiple comparisons



# Testing multiple comparisons



# Testing multiple comparisons

- ▶ is this a problem?  $\Pr(\text{Type-I error}) > 0.05$  ???
- ▶ M&D text discusses some different concepts:
- ▶ error rate **per contrast**  $\alpha_{PC}$ 
  - ▶ probability that a particular contrast will be falsely declared significant
- ▶ experiment-wise error rate  $\alpha_{EW}$ 
  - ▶ probability that one or more contrasts will be falsely declared significant in an experiment
- ▶ family-wise error rate  $\alpha_{FW}$ 
  - ▶ has to do with multiple factor experiments (more later in the course)

# Testing multiple comparisons

- ▶ In our example,  $\alpha_{PC} = 0.05$
- ▶ experiment-wise error rate  $\alpha_{EW} = 0.143$
- ▶ so which error rate should be controlled at the 0.05 level?
- ▶ this is an issue "about which reasonable people differ"
  - ▶ i.e. intelligent and informed people have different approaches
- ▶ M&D suggest controlling  $\alpha_{EW}$  at the 0.05 level
- ▶ see chapter for an interesting discussion of the pros and cons of different approaches

# Methods of controlling $\alpha_{EW}$ at 0.05

- ▶ planned vs post-hoc comparisons
- ▶ 3 methods
  - ▶ Bonferroni, Tukey, Scheffe
- ▶ M&D have a flowchart (decision tree) to help you decide which procedure to use

# Planned vs Post-hoc contrasts

## 1. Planned Contrast

- ▶ a contrast that an experimenter decided to test **prior to any examination of the data**
- ▶ (i.e. the data do not influence your choice of which contrast(s) to test)

## 2. Post-Hoc Contrast

- ▶ a contrast that an experimenter decided to test only after having looked at the data
- ▶ i.e. a contrast "suggested by the data"
- ▶ e.g. following large differences you observe in your dataset

# Planned vs Post-hoc contrasts

- ▶ why is this distinction important?
- ▶ If the contrast(s) to be tested are suggested by the data, e.g. the largest differences are tested
- ▶ the sampling distribution of "differences between **any** 2 means" has a very different distribution than the "largest difference between means"
- ▶ Type-I error rate ends up being inflated if you only test the largest differences in your dataset
- ▶ M&D have a nice discussion of this in the chapter
- ▶ we will show it in R using simulations

# Multiple Planned Comparisons

- ▶ The Bonferroni adjustment is remarkable simple
- ▶ compute the F statistic and p-value for each contrast, as usual
- ▶ then instead of comparing each p-value to  $\alpha$  (e.g. 0.05), instead compare it to  $\frac{\alpha}{C}$ , where  $C$  is the total number of contrasts you will be testing
- ▶  $\alpha$  gets lowered in proportion to the number of contrasts
- ▶ each contrast is therefore more conservative
- ▶ OK for small values of  $C$  but overly conservative for large values of  $C$



# Multiple Planned Comparisons

- ▶ Holm-Bonferroni method : [https://en.wikipedia.org/wiki/HolmBonferroni\\_method](https://en.wikipedia.org/wiki/HolmBonferroni_method)
- ▶ less conservative than straight Bonferroni
- ▶ graded adjustment with larger corrections for less significant p-values
- ▶ check online for examples
- ▶ can use the `p.adjust()` function in **R**

# Multiple Planned Comparisons

- ▶ Keppel (and others) suggest a different approach
- ▶ you're allowed to test up to  $a - 1$  orthogonal planned contrasts without any adjustment of  $\alpha$
- ▶ he argues that Bonferroni correction unfairly penalizes planned orthogonal contrasts
- ▶ if contrasts are planned, orthogonal and number  $a - 1$  or fewer, then because the set of contrasts is not data-driven, and do not overlap, then there should be no need to adjust  $\alpha$  level
- ▶ overall  $\alpha$  level should be no different than that for the omnibus F test

# Post Hoc Pairwise Comparisons

- ▶ Tukey's procedure allows you to perform tests of **all possible pairwise comparisons** in an experiment and still maintain  $\alpha_{EW} = 0.05$
- ▶ the `TukeyHSD()` function in R will do this for you
- ▶ Tukey procedure makes each pairwise test more conservative
- ▶ designed to take into account the idea that data-driven tests will involve higher Type-I error rates
- ▶ there are various modifications of Tukey's procedure when sample variances are unequal or when samples sizes are unequal (see M&D)

# Post Hoc Pairwise Comparisons

- ▶ Scheffe method maintains  $\alpha_{EW}$  at 0.05 when at least some of the contrasts to be tested are complex, and suggested by the data (post-hoc)
- ▶ see M&D text for a detailed description of the method
- ▶ Scheffe method is quite conservative
- ▶ see tables 5.4 & 5.5 for comparison between methods

# Other Procedures

- ▶ Dunnett's procedure
  - ▶ useful when one of the groups is considered a control and is involved in all contrasts
- ▶ Fisher's LSD (least significant difference)
- ▶ Newman-Keuls
- ▶ see M&D text for details about these other methods

# What should I do?

- ▶ decide which approach **you** think is most reasonable, given your data and your experimental design
- ▶ be ready to defend your approach to reviewers
- ▶ be ready to use a different approach if necessary
- ▶ what's the "culture" in your lab / field / journal?

# R Code

- ▶ ANOVA using the `aov()` function in R
- ▶ computing  $F_{\text{comp}}$  manually
- ▶ using `TukeyHSD()`
- ▶ simulations of multiple comparison Type-I error rates
  - ▶ planned vs pos-hoc comparisons