

Contrasts

- Howell: Chapter 12
- Other books...
- Last modified: 11/2/04

Multiple comparisons

- Suppose we want to make the following comparisons:

$$\mu_1 = \mu_2$$

$$\mu_3 = \mu_4$$

$$\mu_1 = \mu_3$$

- Suppose that each test uses an α level of .05
 - *Error rate per comparison: .05*
- How likely is it that we will obtain at least one significant result in our study?
 - How probable is it that we will make at least one Type I error?

Familywise error rate

- This probability is the *familywise* error rate

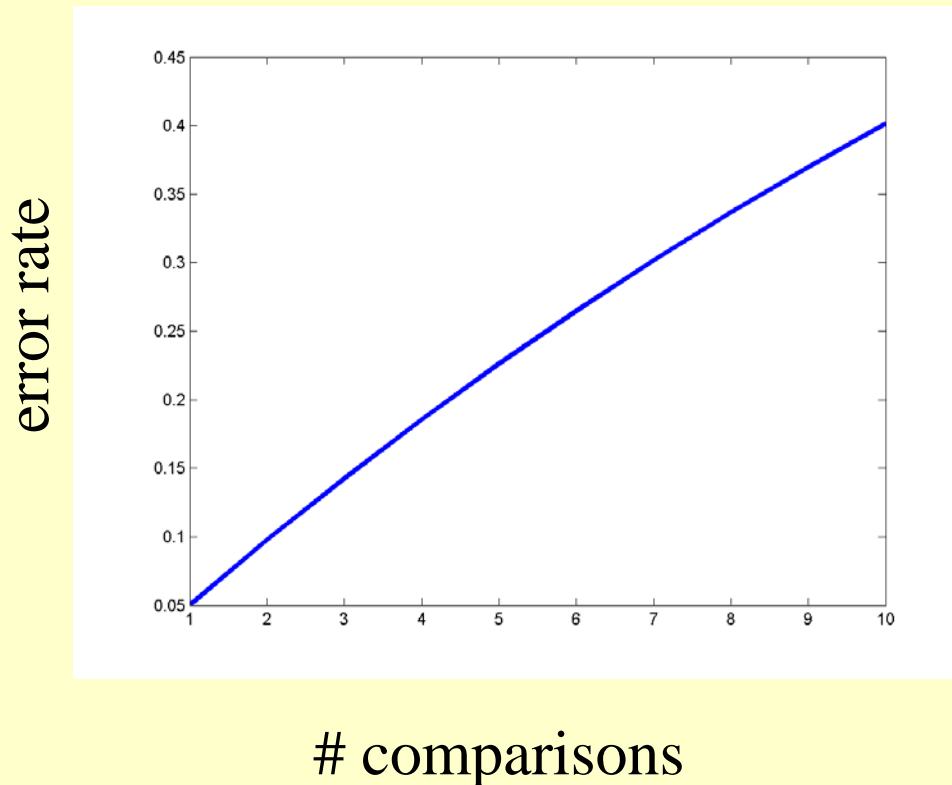
$P(\text{at least one Type I error}) = 1 - P(\text{no Type I errors})$

$$= 1 - \underbrace{(1 - \alpha) \times (1 - \alpha) \times \dots \times (1 - \alpha)}_C = 1 - (1 - \alpha)^C$$

- Many multiple comparisons procedures attempt to control the familywise error rate

Familywise error rate

- When α is the typical value of 0.05, the familywise error rate will grow as follows



Null hypothesis

- In ANOVA, the null hypothesis is

$$\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

- *Rarely* are we really interested in the “complete” null hypothesis
- Sometimes, we might be interested in testing a *few* hypotheses
- Sometimes we end up testing *all pairwise means*
 - $a(a - 1)/2$ comparisons

Contrast types

- ***Planned contrast***: contrast that the experimenter decided to test prior to any examination of the data
- ***Post hoc contrast***: contrast that the experimenter decided to test only after having observed some or all of the data
 - Post hoc contrast is said to be *suggested* by the data
- ***Why*** is the distinction between the two types of contrast important?

Example

- Suppose that we obtain the following means:

$$\bar{Y}_1 = 50, \bar{Y}_2 = 44, \bar{Y}_3 = 52, \text{ and } \bar{Y}_4 = 60$$

- Consider the contrast $\mu_2 - \mu_4$ when all the population means are **equal**: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
- If comparison of groups 2 and 4 was planned and $\alpha = 0.05$ is used, 5 out of every 100 times the experiment would be conducted, the contrast would be statistically significant
 - Type I error

Example

- If this contrast had not been planned, how would things change?
- Suppose the experiment were repeated with
$$\bar{Y}_1 = 46, \bar{Y}_2 = 57, \bar{Y}_3 = 49, \text{ and } \bar{Y}_4 = 54$$
- These results suggest that groups 1 and 2 be compared as they have the largest difference
 - Usual goal of hypothesis testing is to “obtain” statistically significant result

Example

- The result would be that the probability of committing a Type I error would *greatly exceed* 0.05
 - The sampling distribution of $\bar{Y}_k - \bar{Y}_l$ is very *different* from $\bar{Y}_{\max} - \bar{Y}_{\min}$ where Y_{\max} and Y_{\min} are associated with the largest and smallest sample means
 - The critical value of the F distribution that provides an α level of .05 for $\bar{Y}_k - \bar{Y}_l$ is too *small* for judging the significance of $\bar{Y}_{\max} - \bar{Y}_{\min}$
- *It matters greatly whether a contrast is planned or is selected post hoc*

Significance of F

- Without a significant group effect, individual comparisons are considered *inappropriate* by some
- In fact, error rates of some multiple comparison tests require overall significance
 - Fisher's least significance difference test
- But, the logic of most post hoc tests does *not* require overall significance before making specific comp.
- Note that the overall F *dilutes* differences among groups
 - *Dilutes* the F when several group means are equal to each other but different from some other mean

Significance of F

- In fact, an F test may be thought of as the average pairwise t^2 (in the case of equal n 's and equal variances)
- In general, requiring overall significance will actually change the familywise error rate, making the multiple-comparison tests *conservative*
 - Significance levels were established without regard for F
- Wilcox: “there seems to be little reason for applying the F test at all”
- But this is not the traditional view...

Multiple planned comparisons

- *Bonferroni* inequality

$$1 - (1 - \alpha)^C \leq C\alpha, \quad 0 \leq \alpha \leq 1$$

- Remember that

$$P(\text{at least one Type I error}) = 1 - (1 - \alpha)^C$$

- Such that

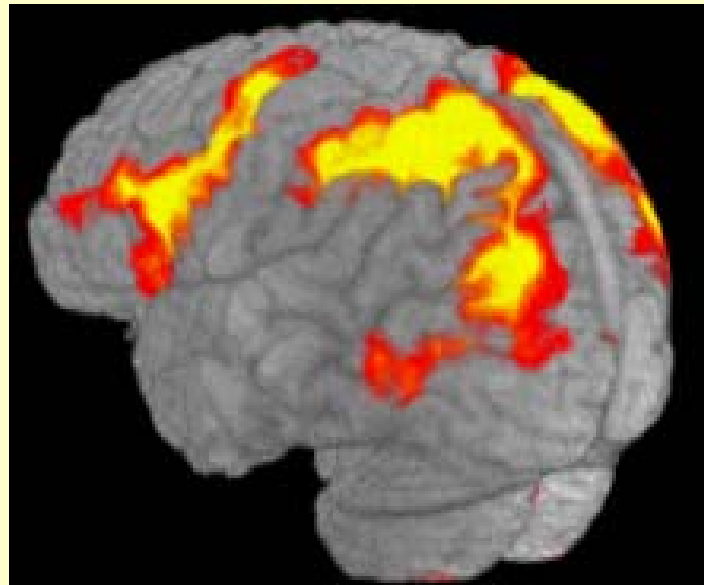
$$P(\text{at least one Type I error}) \leq C\alpha$$

- Thus, by setting $\alpha = .05/C$

$$P(\text{at least one Type I error}) \leq .05$$

Bonferroni's method

- Bonferroni's *method*: set $\alpha = \alpha' / C$, where α' is the level adopted for a single test
- Example: imaging: 50,000 to 100,000 comparisons



Bonferroni's method

- Bonferroni's method precisely controls for the probability of a Type I error for the experiment
 - Widely used, not only in the context of ANOVA
- Cost: As the number of tests increases, it becomes more difficult to detect an individual *true* effect
 - Each individual hypothesis is tested at the α'/C level
- Some have argued that this puts each hypothesis test at an unfair *disadvantage*

Post-hoc comparisons

- Suppose a researcher is interested in the following comparisons: μ_1 vs. μ_2 , μ_2 vs. μ_3 , and μ_3 vs. μ_4
- As long as these comparisons have been made a priori, α for the entire experiment can be maintained at .05 by using a level of .05/3
- This is the case even though there are a total of 6 pairs [$4(4 - 1)/2$], as long as the other comparisons are *ignored!*
- *Looking at the data and changing one's mind is equivalent to doing all of the comparisons and adopting an improper familywise α*

Studentized range statistic

- In general, when the number of comparisons C is $a(a - 1)/2$, the Bonferroni approach is usually **not as powerful** as other special-purpose techniques that have been developed specifically for pairwise comparisons
- Many post hoc tests employ the *Studentized range statistic*

$$q_r = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\sqrt{\frac{MSE}{n}}}$$

- Where \bar{Y}_{\max} and \bar{Y}_{\min} are the largest and smallest sample means

Example

- To use q , we first rank the means from smallest to largest

$$\begin{array}{ccccc} \bar{Y}_1 & \bar{Y}_2 & \bar{Y}_3 & \bar{Y}_4 & \bar{Y}_5 \\ 4 & 10 & 11 & 24 & 29 \end{array}$$

- Suppose also that $n = 8$, $df_{\text{error}} = 35$, and $MSE = 32$

$$q_5 = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\sqrt{\frac{MSE}{n}}} = \frac{29 - 4}{\sqrt{\frac{32}{8}}} = \frac{25}{\sqrt{4}} = 12.5$$

Example

- In general, tables for q exist that take into account the *number of steps* of the means
- In the present case: $q_{0.05}(5, df_{\text{error}} = 35) = 4.07$
- Since $q_5 = 12.5 > 4.07$, we reject the null hypothesis and conclude that there is a significant difference *between the largest and smallest* means

Example

- It is useful to solve for the *smallest* difference that would be *significant* and compare with actual difference

$$\bar{Y}_{\max} - \bar{Y}_{\min} = q_{\alpha}(r, df_{\text{error}}) \sqrt{\frac{MSE}{n}}$$

- In the present case

$$\bar{Y}_{\max} - \bar{Y}_{\min} = 4.07 \sqrt{\frac{32}{8}} = 8.14$$

- A difference in means equal to or greater than 8.14 would be deemed *significant*

Multiple comparisons

- Results from multiple comparisons

$$\begin{array}{ccc} \mathbf{A} & & \mathbf{B} & & \mathbf{C} \\ \hline & & & & \hline \end{array}$$

- In general, if $A = B$, $B = C$, then $A = C$...
- Remember that the situation here is *probabilistic*
- *Failure to reject H_0 does not mean that they are equal*
 - They are not sufficiently different for us to assert that they are different
 - Although we don't have evidence that A and B differ or that B and C differ, we *do* have evidence that A and C differ

Linear contrasts

- Pairwise comparisons of means are a special case of linear contrasts
- Comparison of one group with another group with general weights

$$\mu_1 = \mu_2$$

$$\mu_3 = \mu_4$$

$$\mu_1 = (\mu_3 + \mu_4) / 2$$

- Linear contrasts take the form of a linear combination of the means

$$L = a_1 \bar{Y}_1 + a_2 \bar{Y}_2 + \dots + a_k \bar{Y}_k$$

$$= \sum a_j \bar{Y}_j$$

Linear contrasts

- Restriction

$$\sum a_j = 0$$

- Suppose we had three means and wanted to compare the first and the second only. This could be done by having $a_1 = 1$, $a_2 = -1$, and $a_3 = 0$

$$L = (1)\bar{Y}_1 + (-1)\bar{Y}_2 + 0\bar{Y}_3 = \bar{Y}_1 - \bar{Y}_2$$

Linear contrasts

- Linear contrasts allow us to express the sum of squared differences between the *means* of sets of treatments

$$SS_{\text{contrast}} = \frac{nL^2}{\sum a_j^2} = \frac{n\left(\sum a_j \bar{Y}_j\right)^2}{\sum a_j^2}$$

- Formula assumes equal sample sizes

Example

- Suppose we had $\bar{Y}_1 = 1.5, \bar{Y}_2 = 2.0, \bar{Y}_3 = 3.0, n = 10$

- We can easily compute

$$\begin{aligned}SS_{\text{treat}} &= n \sum (\bar{Y}_j - \bar{Y})^2 \\ &= 10 \left((1.5 - 2.17)^2 + (2 - 2.17)^2 + (3 - 2.17)^2 \right) \\ &= 10(0.44 + 0.028 + 0.694) = 11.667\end{aligned}$$

- Let's say we wanted to compare the average of treatments 1 and 2 to treatment 3. Thus

$$L = \sum a_j \bar{Y}_j = (1)1.5 + (1)2.0 + (-2)3.0 = -2.5$$

$$SS_{\text{contrast}_1} = \frac{nL^2}{\sum a_j^2} = \frac{10(-2.5)^2}{6} = 10.417$$

Example

- Now suppose we wanted to compare groups 1 and 2:

$$L = \sum a_j \bar{Y}_j = (1)1.5 + (-1)2.0 + (0)3.0 = -0.5$$

$$SS_{\text{contrast}_2} = \frac{nL^2}{\sum a_j^2} = \frac{10(-0.5)^2}{6} = 1.25$$

- Note that the sum of the two SS_{contrast} is

$$SS_{\text{treat}} = SS_{\text{contrast}_1} + SS_{\text{contrast}_2}$$

$$11.667 = 10.417 + 1.25$$

- In this case, we can say that the contrasts *completely partition* SS_{treat}

F test for contrast

- Note that the *absolute* value of the contrast weights does not matter

Means: Y_1 Y_2 Y_3 Y_4 Y_5

$$a_j: \begin{cases} 2 & 2 & 2 & -3 & -3 \\ 1 & 1 & 1 & -1.5 & -1.5 \end{cases}$$

- The *significance* of a contrast can be tested with an *F* test

$$F = \frac{MS_{\text{contrast}}}{MS_{\text{error}}} = \frac{nL^2 / \sum a_j^2}{MS_{\text{error}}}$$

Contrasts

- The square root of the F for a simple contrast ($a_1 = 1, a_2 = -1$) is the same value obtained in a t test
 - t tests are special cases of linear contrasts
- Note however that if you run several contrasts, the familywise error will be much larger than α
 - Bonferroni correction can be used
 - Run fewer contrasts! Only as many as needed
 - If they were really a priori, probably ok without correction

Orthogonal contrasts

- Some contrasts are independent of one another, while others “share” information
- Independent contrasts are called *orthogonal*

$$\sum a_j = 0$$

$$\sum a_j b_j = 0$$

contrasts given by a_j and b_j are orthogonal

Orthogonal contrasts

- It is useful to have orthogonal contrasts because they **exactly partition** SS_{treat}
- However, it is not strictly necessary that all contrasts be orthogonal
 - But remember that in this case the contrasts do **not** convey independent information

Degrees of freedom

- For a contrast, we have

$$F = \frac{MS_{\text{contrast}}}{MS_{\text{error}}}$$

- What are the *numerator* degrees of freedom?
- Consider df for SS_{contrast}
- A contrast always compares two quantities

$$\mu_1 = \mu_2$$

$$\mu_3 = \mu_4$$

$$\mu_1 = (\mu_3 + \mu_4) / 2$$

Degrees of freedom

- Thus SS_{contrast} has $df = 1$
- Another way to think of the df is that the F for a contrast can in fact be written in the usual way

$$F = \frac{(SSE(R) - SSE(F)) / (df_R - df_F)}{SSE(F) / df_F}$$

- To determine df_R , we must determine the number of *independent* parameters in the **restricted model** which is associated with the contrast
- Consider the following null hypothesis:

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

Degrees of freedom

- The corresponding *restricted* model is

$$Y_{ij} = \mu_j + \varepsilon_{ij}$$

where $1/3\mu_1 + 1/3\mu_2 + 1/3\mu_3 - \mu_4 = 0$

- Model has 4 parameters but only 3 are *independent*
- In the general case for a groups, we would have $a - 1$ independent parameters
- Thus

$$\begin{aligned} df_R - df_F &= [N - (a - 1)] - (N - a) \\ &= 1 \end{aligned}$$