Chapter 9: Multiple Comparisons and Contrasts

1. The Multiple Testing Problem
2. Tukey’s (HSD) Method for Pairwise Comparisons
3. Bonferroni’s Method
4. Confidence Intervals for Pairwise Comparisons
5. Comparison of Methods
6. Dunnett’s Method for Comparisons against a Control
7. Power for Pairwise Comparisons after one-way ANOVA
8. Other multiple testing methods (including FDR)
9. Linear contrasts

Examples:

1. Clover Example: Multiple Comparisons
2. Wheat Example: Contrasts
3. Multiple Testing Simulation

1. The Multiple Testing Problem

Clover Example: (Steele and Torrie) Nitrogen (N) content of Red Clover inoculated with combinations of *Rhizobium trifolii* and of *Rhizobium meliloti* (bacteria) strains. Five pots of each of six treatments completely randomized in a greenhouse experiment (t = 6 trts, n = 5 pots/trt, n_T=30).

Based on the one-way ANOVA F-test (F = 14.37, p-value < 0.0001), we Reject H0 and conclude not all the means are the same.

Which means are different?
In R, we can run all pairwise comparisons using `emmeans(OneWayFit, pairwise ~ strain, adjust = "none")`. Using `adjust = "none"` these are unadjusted pairwise comparisons. This means we are not correcting for multiple testing.

The Problem: Running all pairwise comparisons involves 15 tests for this example. Prob(1 or more type I errors) >> 0.05

Comparisonwise vs Experimentwise Error Rate

A **false rejection (or type I error)** is when we Reject H0 when H0 is really true.

The **comparisonwise error rate (CER)** is the probability of a false rejection on a single test. This is what we have focused on so far.
Example: A single pairwise comparison (ex: H_0: μ_1-μ_2 = 0).

The **experimentwise error rate (EER)** is the probability of having at least one false rejection in the group of tests from a single experiment.
The “experiment” is a single one-way ANOVA analysis.
Example: For the Clover data with t=6 treatments, we have 15 pairwise comparisons.

Multiple Testing Simulation

Generate data with true H0: $\mu_1 = \mu_2 = \ldots = \mu_t$
n = 10 observations per group
t = 5 or 10 treatment groups
1000 runs of the simulation for each scenario
Note: 1 run = 1 ANOVA = 1 “experiment”

With t = 5 treatment groups, the observed EER is 29%.
With t = 10 treatment groups, the observed EER is 61%.

In other words, 29 - 61% of experiments in which treatments are really the same will find evidence of at least one difference using unadjusted comparisons.

The scientific community is generally unwilling to accept such high error rates. How do we get control of the experimentwise error rate?

Fisher’s F-Protected LSD

Recall LSD = least significant difference = unadjusted comparisons.

Strategy:

If the F-test p-value < α, run further (unadjusted) pairwise comparisons

If the F-test ≥ α, STOP and DO NOT compare means.

This is called the F-protected LSD method (FLSD).

Note: FLSD controls the EER under the complete null.

If H_0 is true, then F test p-value < 0.05 5% of the time. Only then we will have the opportunity to make type I errors in that experiment.

Problem with Fisher's F-Protected LSD

Example: $t=11$, $n=10$, $\alpha=0.05$, $\sigma=1$

Suppose $\mu_1= \mu_2 ..... =\mu_{10} = 0$, $\mu_{11}=1{,}000{,}000$

This is a "partial null hypothesis"; some treatments are equal, some are not. Many other partial null hypotheses are possible.

Here not all rejections are false.

Now use the FLSD method:

1. The overall F test p-value < α almost certainly.
2. The LSD (unadjusted) method will then be used to compare the 10 other treatments to each other. From simulation, EER is approximately 61%.

This situation seems far-fetched, but the control (or some other) treatment is often very different from the others. The F-protection is "broken" by the presence of this very different treatment.

FLSD controls EER under complete null, but not under partial null.

2. Tukey's (HSD) Method

John Tukey proposed a method that can be used to run all pairwise comparisons while controlling maximum EER. That is, the experimentwise error rate that would occur under the worst case of all possible partial null hypotheses.

This method is based on a Honestly Significant Difference (HSD) rather than the LSD value.

$$TukeyME = HSD = q_a(t, \mathrm{df})\sqrt{\frac{\mathrm{s}_{\mathrm{W}}^2}{\mathrm{n}}} = q_a(t, \mathrm{df})\sqrt{\frac{\mathrm{MSResid}}{\mathrm{n}}}$$

Values of $q_\alpha(t,df)$ are given in Table 10 for $\alpha=0.05$ and $\alpha=0.01$ where t = number of treatments, df = dfResid.
To find the Tukey q table value using R:

```
qtukey((1-alpha), t, dfResid)
```

Return to the Clover example:

t = 6 treatments,
n = 5 observations per treatment
dfResid = 30 – 6 = 24
`qt(0.975, df = 24)` = 2.064 (or use Table 2)
`qtukey(0.95, 6, df = 24)` = 4.37 (or use Table 10)

$$s_W^2 = MS\,\mathrm{Re}\,sid = \hat{\sigma}^2 = 11.79$$

$$UnadjME = LSD = t_{\alpha/2}\sqrt{s_W^2\,\frac{2}{n}} = 2.064\sqrt{11.79\,\frac{2}{5}} = 4.5$$

$$TukeyME = HSD = q_\alpha\sqrt{\frac{s_W^2}{n}} = 4.37\sqrt{\frac{11.79}{5}} = 6.7$$

Tukey ME > Unadj ME, hence we will find evidence of fewer differences (or weaker evidence of differences) using Tukey's method.

Comments about Tukey's Method

1. Tukey's method controls maximum EER, without the need for F-protection. So you can consider Tukey adjusted pairwise comparisons regardless of F-test results.
2. It is possible to have F test p-value < α, but have no evidence of differences from pairwise comparisons after Tukey adjustment.
3. The HSD value from the previous slide is an (adjusted) ME for pairwise comparisons. It can be used to construct CI for pairwise comparisons.
4. When sample sizes are equal, use this just like an LSD value. A difference for Trts i and j is declared when $|\bar{y}_i - \bar{y}_j| > HSD$ (because 0 will not be contained in the CI). Order the means and underline the means that are <u>not</u> different or use `CLD()` from `emmeans`.
5. In R, we will calculate Tukey adjusted p-values using the `emmeans` package (See next slide or **Clover** Example.)

Tukey Adjustment using R

For any of these methods, we start by fitting the model:

```
OneWayFit <- lm(N ~ Strain, data = clover)
```

1. **emmeans()** from the emmeans package. Note that the Tukey adjustment is done by default.
   ```
    > library(emmeans)
    > emmeans(OneWayFit, pairwise ~ Strain)
   ```
2. There is a base function TukeyHSD(), but unlike option above, this does not extend to more complicated models.
3. Multiple comparisons can also be done using the multcomp package, but we will not use this option.

See the “**Clover Example**”.

3. Bonferroni's Method

Bonferroni's method will control the experimentwise error rate for any set of m tests (not restricted to pairwise comparisons).

Let α_E represent the experimentwise error rate (EER) and α_I represent the comparisonwise (or individual) error rate (CER).
Bonferroni's inequality states that $\alpha_E \leq m\alpha_I$.

Hence if we want to control the experimentwise error rate at α_E at a fixed level α, we need to use α_I/m for each of the m tests.

Hence to incorporate the Bonferroni adjustment with m tests:

1. Calculate a Bonferroni adjusted ME using α/m (instead of α).
2. Calculate Bonferroni adjusted p-values by multiplying the (unadjusted) p-values by m. If p-values come out to be greater than 1, just report a value of 1. Or use `emmeans()` with `adjust = "bonferroni"`

Return to the Clover example:

t = 6 treatments ->
m = 15 tests of pairwise comparisons
n = 5 observations per treatment
dfResid = 30 – 6 = 24
`qt(1-(0.05/(2*15)), df = 24)` = 3.258

$$s_W^2 = MS\,\mathrm{Re}\,sid = \hat{\sigma}^2 = 11.79$$

$$BonME = t_{\alpha/(2*m)}\sqrt{s_W^2\frac{2}{n}} = 3.258\sqrt{11.79\frac{2}{5}} = 7.1$$

Bon ME > Tukey ME, hence we will find evidence of fewer differences (or weaker evidence of differences) using Bonferroni’s method.

Comments about Bonferonni's Method

1. Both Tukey and Bonferonni methods control the EER. However, Bonferonni is more conservative and hence will yield evidence of fewer differences (or weaker evidence of differences). (See the Multiple Testing Simulation.)
2. Hence, **Bonferonni's method is NOT commonly used for pairwise comparisons.**
3. However, it is still a handy test to be aware of because (1) it can be used for any set of m tests (not just pairwise comparisons after ANOVA) and (2) it is very easy to implement "by hand".
4. In R, we will calculate Tukey adjusted p-values using the `emmeans` package and function with `adjust = "bonferroni"`. (See **Clover** Example.)

4. Confidence Intervals for All Pairwise Comparisons

For unadjusted (one at a time) confidence intervals, the error rate is α per interval (a comparisonwise error rate). The probability that there is at least one error in the collection of intervals (experimentwise error rate) is much higher.

Simultaneous (or adjusted) confidence intervals for differences between treatment means can also be computed using the Tukey's HSD method. These intervals have an error rate of α for all the intervals simultaneously. The probability that all these intervals simultaneously contain the true differences is 0.95.

Simultaneous CIs for Pairwise Comparisons

In R, using the emmeans package, simultaneous confidence intervals can be constructed using the `confint()` function and plotted using the `plot()` function.

See the "**Clover Example**".

Simultaneous confidence intervals are much wider than Unadjusted LSD intervals, because they have the simultaneous coverage property that:

Prob(all intervals contain the true differences) ≥ 0.95

Frequency interpretation: Tukey HSD intervals are calculated using a method with the property that under repetition of the experiment, in 95% of the experiments all of the Tukey intervals for that experiment will contain the true mean differences.

Adjustments for Unequal Sample Sizes

1. The **unadjusted (LSD)** method, as previously described allows unequal sample sizes:

$$\bar{y}_{i.} - \bar{y}_{j.} \pm t_{\alpha/2} s_W \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

2. The **Tukey (HSD)** method can be similarly adjusted for unequal sample sizes:

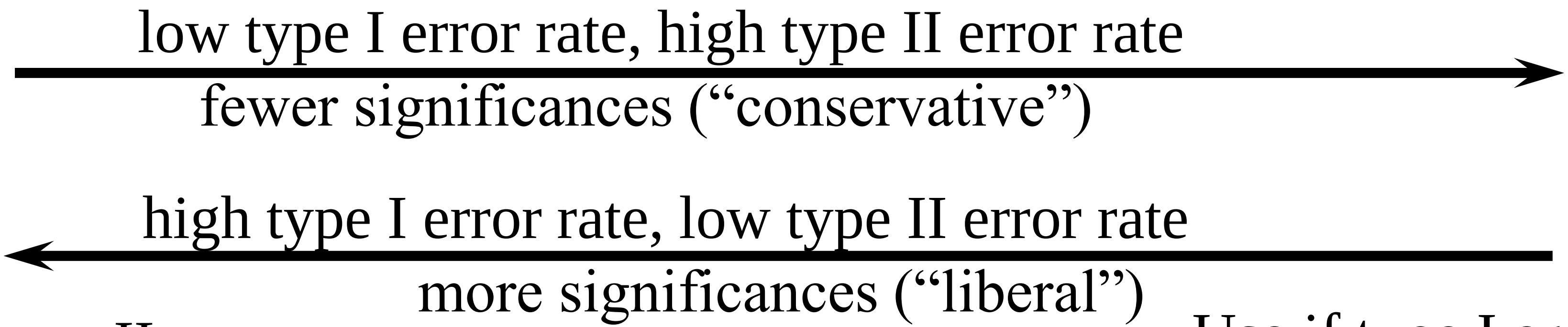
$$\bar{y}_{i.} - \bar{y}_{j.} \pm q_{\alpha}(t, \mathrm{df}) \sqrt{\frac{s_W^2}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

This is called the **Tukey-Kramer** interval. It is approximate, but has been shown to be conservative (fewer errors than claimed).

5. Comparison of Methods

Asking which method is the best is really asking which is the correct α. The correct α, as usual, depends on the relative seriousness of type I versus type II errors. Below are the ME values for the clover data.

LSD/ Unadjusted	HSD/ Tukey	Bonferroni
4.5	6.7	7.1



low type I error rate, high type II error rate

fewer significances ("conservative")

high type I error rate, low type II error rate

more significances ("liberal")

Use if type II error is very serious

Use if type I error is very serious

Comments about Multiple Testing Adjustments

1. Using a multiple testing adjustment yields higher p-values (or wider confidence intervals) than unadjusted.
2. Using a multiple testing adjustment, running more tests yields higher p-values (or wider confidence intervals) than unadjusted.
3. It is usually not as important which method we use as it is that we correctly interpret the method we do use. A common (but serious) mistake is to use a method that controls CER, but interpret it as if it controls EER.
4. Different subject areas have different expectations about correcting for multiple testing.
5. Tukey adjustment is very common. Tukey controls the maximum EER, is simple and many people are familiar with it. It is slightly conservative.
6. Bonferroni is too conservative for running all pairwise comparisons. In other situations, it can be a good choice.

Return to the Multiple Testing Simulation

Generate data with true H0: $\mu_1 = \mu_2 = \ldots = \mu_t$
n = 10 observations per group
t = 5 or 10 treatment groups
1000 runs of the simulation for each scenario

Unadjusted: CER ≈ 0.05; EER >> 0.05
With t = 5 treatment groups, the observed EER is 29%.
With t = 10 treatment groups, the observed EER is 61%.

Tukey: CER << 0.05; EER ≈ 0.05
With t = 5 treatment groups, the observed EER is 5.5%.
With t = 10 treatment groups, the observed EER is 4.0%.

Bonferroni: CER << 0.05; EER < 0.05
With t = 5 treatment groups, the observed EER is 3.9%.
With t = 10 treatment groups, the observed EER is 2.8%

Conclusions from Multiple Testing Simulation

Recall that CER = comparison-wise error rate, EER = experiment-wise error rate.

Unadjusted method controls CER but not EER.

Tukey’s method controls EER by reducing the CER.

Bonferroni’s method over-controls the EER. Hence we say this method is “conservative”.

6. Dunnett’s Method

In some cases, the primary objective of an experiment is to compare a single “control” treatment to all other (“active”) treatments.
Dunnett’s method is designed to control the experimentwise error rate in the family of comparisons of individual means vs control.
Note that this is a subset of all pairwise comparisons, since we only consider t-1 tests.

$$DunnettME = t_{Dunnett}\sqrt{\frac{2s_W^2}{n}}$$

where df = dfResid and $t_{Dunnett}$ is from Table 11.

For the clover data (dfResid = 24):

$$DunnettME = 2.76\sqrt{\frac{2(11.79)}{5}} = 5.99$$

Comments about Dunnett's Method:

In R, Dunnett's method can be run using `emmeans`:

```
emmeans(OneWayFit, dunnett ~ Trt)
```

Notes: The first group is used as "control". Can reorder factor levels so that the control group of interest is first. See **"Clover Example".**

1. For comparison: DunnettME = 5.85 is smaller than the Tukey's HSD = 6.71, but larger than the Unadjusted LSD = 4.5.
2. Dunnett's method increases power (compared to Tukey) by reducing the number of tests.
3. Tukey's method can be used when we are interested in comparing "active" treatments verses "control". However, since Tukey controls EER when comparing all pairs of treatments, it will have unnecessarily low power. We are "paying a price" for testing comparisons that aren't really of interest.

7. Power for Pairwise Comparisons after one-way ANOVA

Using Lenth's online power calculator, we can calculate power corresponding to pairwise comparisons after one-way ANOVA. Power can be calculated accounting for Tukey and Dunnett adjustments.

Example: Investigators plan to compare t = 6 groups with n = 8 observations per group. They want to calculate power to detect a meaningful (or conjectured) difference between means of 2. They conjecture the within-group standard deviation will be 1.2.
Note: In Lenth, set "contrast coefficients" to 1 -1 to represent a pairwise comparison.

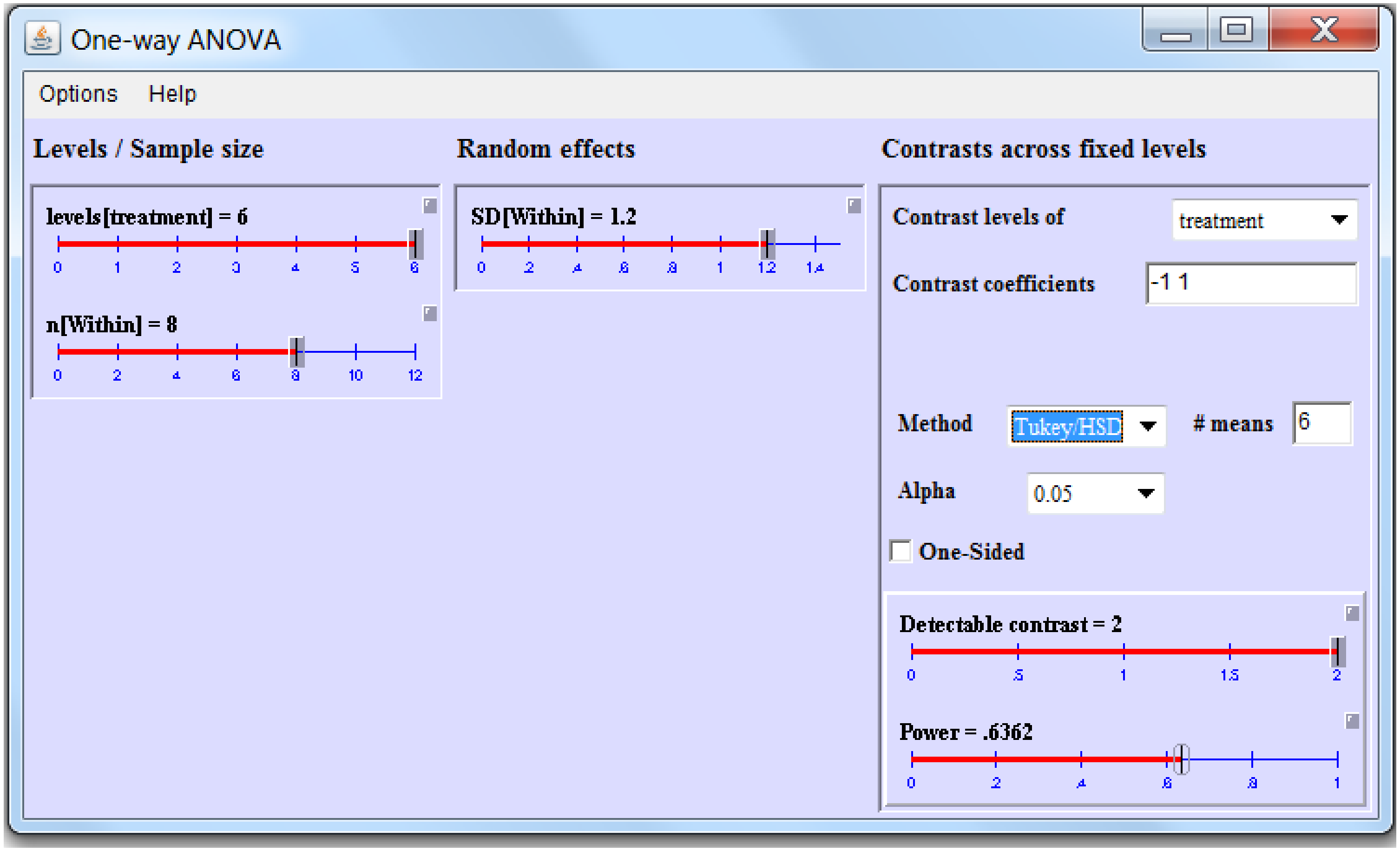
Using One-way ANOVA with Method = Tukey/HSD, power is found to be 0.6362.
Using One-way ANOVA with Method = t (=unadjusted), power is found to be 0.9026.
For comparison, using a two-sample t-test, power is found to be 0.9125.

http://homepage.stat.uiowa.edu/~rlenth/Power/

* Choose Balanced ANOVA, then Built-in models = “One Way ANOVA”, Study the Power of = “Differences/Contrasts”.

t =
# trts

n =
n per
group



Levels / Sample size

levels[treatment] = 6

n[Within] = 8

Random effects

SD[Within] = 1.2

Contrasts across fixed levels

Contrast levels of: treatment

Contrast coefficients: -1 1

Method: Tukey/HSD # means: 6

Alpha: 0.05

□ One-Sided

Detectable contrast = 2

Power = .6362

contrast
coefs

conj
difference
between
means

sd within group
= sd within

8. Other Multiple Testing Methods available in R

From glht() multcomp package:

- Bonferroni
- Dunnett
- Holm
- Shaffer
- Tukey
- Westfall

From p.adjust():

- holm
- hochberg
- hommel
- bonferroni
- BH = fdr (Benjamini & Hochberg)
- BY (Benjamini & Yekutieli)

Large Scale Multiple Testing

So far in these notes, we have focused on “classical” multiple testing scenarios. Pairwise comparisons of means from ANOVA is an example.

In large scale multiple testing scenarios (hundreds or thousands of tests), different methods are used which tend to focus on the false discovery rate (FDR).

FDR = (# False Rejections)/ (Total # of Rejections)

For example, suppose a genomics experiment was conducted and 100 genes were identified as “differentially expressed” with a FDR of 5%. Then we would expect 5/100 of our differentially expressed genes to be “false positives”.

Large Scale Multiple Testing continued

Notice that the idea of FDR (proportion of rejections that are false) is very different from the idea of EER (proportion of experiments with at least one false rejection).

FDR methods will be considerably more lenient (meaning more false rejections) but have higher power to detect a difference for a particular comparison. Again, this is a trade off.

Most common FDR method is Benjamini-Hochberg but there are MANY other methods for controlling the FDR.

9. Linear Contrasts in the One-way ANOVA

In many cases, ANOVA and pairwise comparisons of means will address all research questions. However, in some cases contrasts are required to address additional comparisons of interest.

A **linear contrast** (usually called just a **contrast**) is a comparison of treatment means – often a comparison of the averages between groups of means.

A contrast is written as:

$l = a_1\mu_1 + a_2\mu_2 + \cdots + a_t\mu_t = \sum_{i=1}^{t} a_i\mu_i$

The contrast is described by a list of coefficients $(a_1, a_2, \ldots a_t)$.

Some a_i's may be zero!

A linear contrast has the property that the sum of the contrast coefficients (the sum of the a_i's) is zero: $\sum_{i=1}^{t} a_i = 0$.

To estimate, just substitute sample means as estimates of the μ_ι's:

$\hat{l} = a_1\overline{y_1} + a_2\overline{y_2} + \cdots + a_t\overline{y}_t = \sum_{i=1}^{t} a_i\overline{y}_i$

Wheat Contrasts Example: An experiment is performed to compare yield for t = four varieties (A, B, C, D) of wheat. Varieties A and B are similar in that they are classified as “resistant” to a particular disease. Varieties C and D are classified as “susceptible”.

Sample means: A = 5.54 , B = 5.16, C = 4.06 , D = 7.42

Contrast Example #1: Compare variety A vs B

H_0: μ_A - μ_B = 0 or $1*\mu_A – 1*\mu_B + 0*\mu_c + 0*\mu_D = 0$

Coefficients (a’s): (1, -1, 0, 0)

Check Sum to Zero: +1 -1 + 0 + 0 = 0

Estimate = $\overline{y_A} - \overline{y_B}$ = 5.54 – 5.16 = 0.38

*Note we already get an estimate and test of this comparison using `emmeans( , pairwise ~ variety)`!

Contrast Example #2: Compare average of resistant varieties (A, B) versus average of susceptible varieties (C, D).

$$H_0: \frac{\mu_A + \mu_B}{2} - \frac{\mu_C + \mu_D}{2} = 0$$

$$l = \frac{1}{2}\mu_A + \frac{1}{2}\mu_B - \frac{1}{2}\mu_C - \frac{1}{2}\mu_D$$

Coefficients (a's): (0.5, 0.5, -0.5, -0.5)

Check Sum to Zero: +0.5 +0.5 – 0.5 -0.5 =0

Estimate = $0.5\overline{y_A} + 0.5\overline{y_B} - 0.5\overline{y_C} - 0.5\overline{y_D}$

= 0.5* 5.54 + 0.5*5.16 – 0.5*4.06 - 0.5* 7.42 = -0.39

*This comparison is <u>not</u> included in the pairwise comparisons!

SE for Linear Contrasts

We have seen how to set up contrasts (by specifying coefficients) and estimate them (using sample means).

However, any tests or confidence intervals require first calculating the standard error of *l*.

Since a contrast involves sums and differences of independent normal means, it is itself normally distributed, and hypotheses about contrasts can be tested using a t-test, or equivalently, an F-test (remember t^2=F).

$$SE(\hat{l}) = \sqrt{s_W^2 \sum_{i=1}^{t} \frac{a_i^2}{n_i}}$$

If n's are equal, then $SE(\hat{l}) = \sqrt{\frac{s_W^2}{n} \sum_{i=1}^{t} a_i^2}$

CI and Test for Linear Contrasts

Let $l = a_1\mu_1 + a_2\mu_2 + \ldots + a_t\mu_t$

Then $\hat{l} = a_1\bar{y}_{1.} + a_2\bar{y}_{2.} + \ldots + a_t\bar{y}_{t.}$

A (1-α)x100% **Confidence Interval** for a contrast *l:*

$$\hat{l} \pm t_{\alpha/2} \sqrt{s_W^2 \sum_{i=1}^{t} \frac{a_i^2}{n_i}}$$

Hypothesis Test for a contrast l:

$$H_0 : l = l_0 \quad \text{vs} \quad H_A : l \neq l_0$$

$$t = \frac{\hat{l} - l_0}{\sqrt{s_W^2 \sum_{i=1}^{t} \frac{a_i^2}{n_i}}}$$

Reject H0 if $|t| > t_{\alpha/2}$

NOTES: (1) df=dfResid=n_T-t (2) s_W^2 = MSResid.

Notes on Linear Contrasts

1. In many cases, pairwise comparisons of means answers all of the relevant research questions. Occasionally, a specific contrast is needed to answer the research question.
2. Pairwise (unadjusted) comparisons of 2 means is a special case of a contrast.
3. In R, contrasts can be estimated and tested using the `contrast()` function within the `emmeans` package. See “**Wheat Example”.**
4. Watch out for the ordering of the grouping variable! An easy way to check to make sure the coefficients have been chosen/stated correctly is to plug in the sample means and confirm the estimate returned by `contrast().`

Multiple Comparisons for *a priori* contrasts

A linear contrast is ***a priori*** if it was selected by the experimenter as being important for testing at the outset of the experiment (before looking at the data).

It is usually considered acceptable to test a small number of *a priori* contrasts without making any multiple comparison adjustments.

For larger groups of *a priori* contrasts, use Bonferroni's method.
You will pay a penalty for testing a lot of contrasts since Bonferroni adjusted p-values are calculated by multiplying the unadjusted p-values by the number of tests.

Scheffe's method for *a posteriori* contrasts

A linear contrast is ***a posteriori*** if it was selected by the experimenter after looking at the data.

Scheffe's method can be used for controlling maximum EER when testing a large number of *a priori* contrasts, or testing any number of *a posteriori* contrasts. Scheffe's method is even more conservative than Bonferroni's method.

We will not cover the details of this method in theses notes!

Steel and Torrie call the process of searching the means for contrasts that show evidence of differences "Data Dredging". Scheffe's method is designed for such contrasts.

Power for a Contrast

Example: Consider an experiment designed to evaluate the effect of pesticide spraying on a brain hormone of a particular species of bird. There are two experimental areas (Spray, Control) and *n* birds are to be sampled at each of two times: Pre and Post spraying. It is expected that the within group standard deviation will be $\sigma = 2$.

μ_1 = mean for Control area Pre-spraying

μ_2 = mean for Control area Post-spraying

μ_3 = mean for Spray area Pre-spraying

μ_4 = mean for Spray area Post-spraying

Investigators are interested in an “interaction” contrast. Specifically:

$l = (\mu_1 - \mu_2) – (\mu_3 - \mu_4) = \mu_1 - \mu_2 - \mu_3 + \mu_4$

They conjecture that the true difference is: 10 – 8 – 11 + 7 = -2.

Using Lenth (next slide), we find that *n* = 43 per group corresponds to power = 0.903.

A Note about this Example:

In STAT512, we will be able to test this interaction contrast directly by running a two-way ANOVA. The result will be the same as the one-way ANOVA approach.

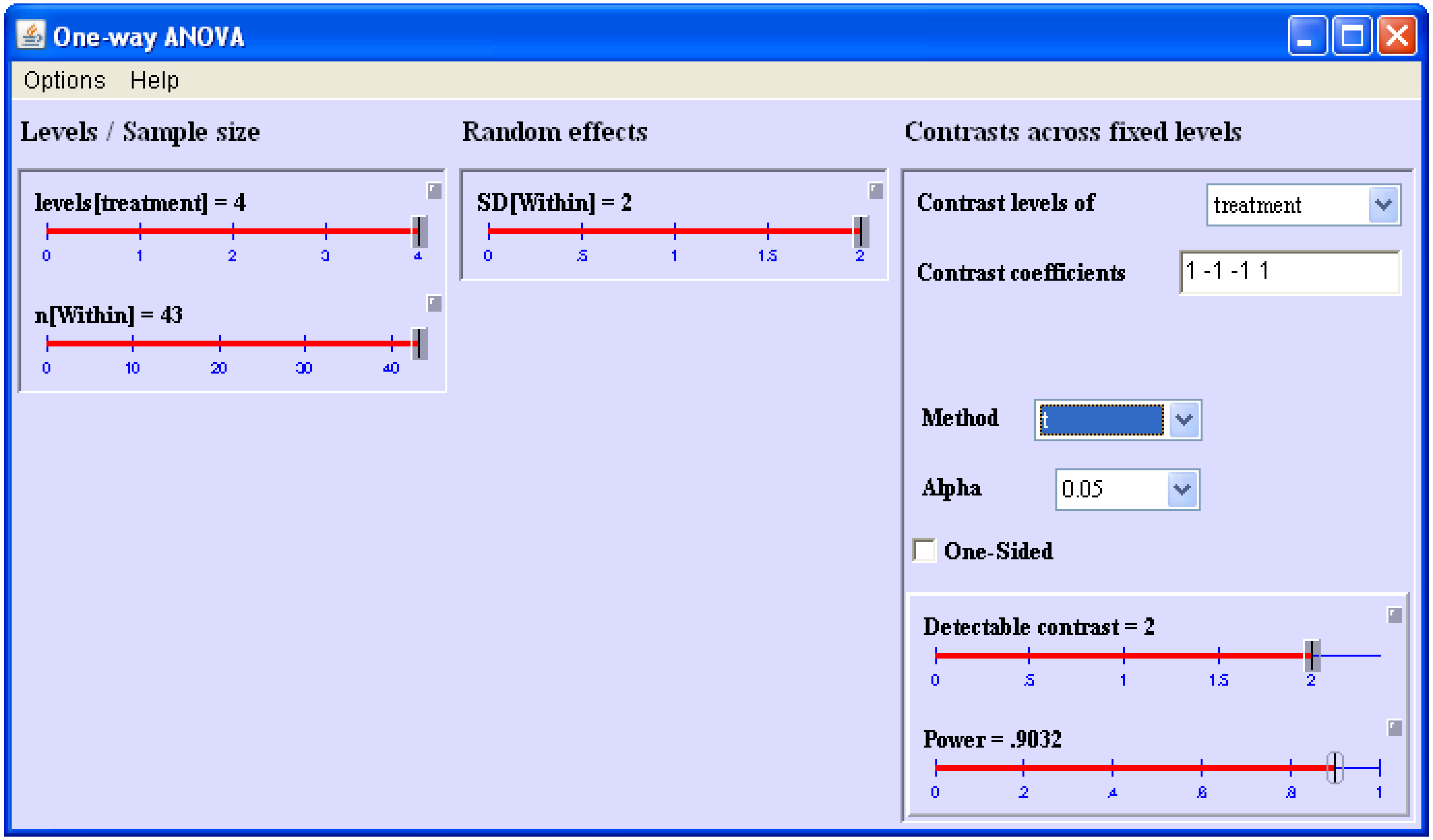
Power for a Contrast using Lenth

http://homepage.stat.uiowa.edu/~rlenth/Power/

* Choose Balanced ANOVA, then Built-in models = “One Way ANOVA”, Study the Power of = “Differences/Contrasts”.

t =
# trts

n =
n per
group



Levels / Sample size

levels[treatment] = 4

n[Within] = 43

Random effects

SD[Within] = 2

Contrasts across fixed levels

Contrast levels of: treatment

Contrast coefficients: 1 -1 -1 1

Method: t

Alpha: 0.05

□ One-Sided

Detectable contrast = 2

Power = .9032

contrast
coefs

conj
value of
contrast

sd within group
= sd within