Section 3

Contrasts and Multiple Testing

The slides for this class are adapted from multiple sources:

- The main outline comes from *ANOVA and Mixed Models: A Short Introduction Using R*, by Lukas Meier (https://stat.ethz.ch/~meier/teaching/anova/index.html)
- *Introduction to Design and Analysis of Experiments*, by George W. Cobb
- Notes from prior semesters created by William Christensen and Dennis Tolley

Learning Outcomes

This section and the associated R examples and assignments achieve the following course expected learning outcomes:

- **Data Import:** Create datasets in R from space-, comma-, tab-delimited files
- **Summary Statistics:** Compute summary statistics from R datasets
- **Analyze Data:** Analyze data from 'Treatment-Control' or 'A/B' experiments using professional statistical software
- **Randomized Design:** Analyze data from completely randomized designs using professional statistical software
- **Variability:** Understand the concept of variability in data and the attempt to identify sources of that variability
- **Writing Statistical Models:** Practice writing statistical models

3. Contrasts and Multiple Testing

3.1 Contrasts

3.2 Multiple Testing

3.1 Contrasts

The ANOVA F-test tests the hypotheses

Means Model

H_0: μ_i are all equal

H_A: At least one μ_i differs from the others

Effects Model

H_0: $\alpha_i = 0, \forall i$

H_A: At least one $\alpha_i \neq 0$

In either case, rejecting the null hypothesis indicates there are differences. But where are those differences????

- We need a follow-up procedure to show where differences exist.
- We may also want to test combinations of means. We can do that with contrasts.

3.1 Contrasts

The "helicopter experiment" involves creating helicopters of different dimensions out of paper and seeing how long it takes for them to drop to the ground. There are 6 different patterns of wing length (short, medium, long) and body length (short, long).

What possible comparisons might we be interested in?

Notice: Wing Length and Body Length are factors created from specific helicopter patterns.

Helicopter Number	Wing Length	Body Length	Mean
1	Short	Short	μ_1
2	Medium	Short	μ_2
3	Long	Short	μ_3
4	Short	Long	μ_4
5	Medium	Long	μ_5
6	Long	Long	μ_6

Contrasts allow for combinations of means to be compared

A contrast is a linear combination of factor means used to test a specific hypothesis to answer a specific question.

$$H_0\colon c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t = 0$$

Where $c_1, \ldots, c_t$ are constants that define the comparisons to be made

It can also be written as

$$H_0\colon \sum_{i=1}^{g} c_i\mu_i = 0$$

In general, $\psi = \sum_{i=1}^{g} c_i\mu_i = 0$

This allows us to test $H_0\colon \psi = 0$

Using a contrast to perform a basic pairwise comparison

Suppose we want to test the following hypothesis:

H_0: $\mu_1 - \mu_2 = 0$
H_A: $\mu_1 - \mu_2 \neq 0$

In terms of $\psi = c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t$

We can define a linear contrast as

$\psi = \mu_1 - \mu_2$

Where $c_1 = 1$, $c_2 = -1$, and all other $c_i = 0$

But that is just a simple pairwise comparison. Let's look at something more complex.

More examples

Suppose a researcher wishes to test four different herbicides for killing weeds,

Trt	Chemical	Weed Type
A	2, 4-D	Broadleaf
B	Dicamba	Broadleaf
C	Glyphosate	Broad Spectrum
D	Ammonium Nonanoate	Broad Spectrum

What contrast coefficients would we use to make the following comparisons:

- 2, 4-D vs. Dicamba
- Broadleaf vs. Broad Spectrum
- Glyphosate vs. the other three

How would you construct $\psi = c_1\mu_1 + c_2\mu_2 + \cdots + c_t\mu_t$?

Hint: $c_1 = \frac{1}{2}, c_2 = \frac{1}{2}, c_3 = -\frac{1}{2}, c_4 = -\frac{1}{2}$ is the answer for one of the comparisons

Using a contrast to test main effects

Consider how originally the helicopter data were entered by numbers 1-6 to indicate the wing length and body length combinations.

How do we define the main effects of body length and wing length?

Degrees of freedom are important!

If a factor has 2 levels (df=1), then there is one contrast statement needed to test the main effect.

If a factor has 3 levels (df=2), then there are two contrast statements needed to test the main effect.

Using a contrast to test main effects

Contrast for Main Effect of Body Length

To compare the differences in body length we want to compare an average of the short-body helicopters with an average of the long-body helicopters:

Helicopter Number	Wing Length	Body Length	Mean
1	Short	Short	μ_1
2	Medium	Short	μ_2
3	Long	Short	μ_3
4	Short	Long	μ_4
5	Medium	Long	μ_5
6	Long	Long	μ_6

Contrast – Body Length (Short vs. Long):

$$\frac{\mu_1 + \mu_2 + \mu_3}{3} - \frac{\mu_4 + \mu_5 + \mu_6}{3}$$

Contrast – Body Length (Short vs. Long):

$$\psi = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3 - \mu_4 - \mu_5 - \mu_6)$$

Using a contrast to test main effects

Helicopter Number	Wing Length	Body Length	Mean
1	Short	Short	μ_1
2	Medium	Short	μ_2
3	Long	Short	μ_3
4	Short	Long	μ_4
5	Medium	Long	μ_5
6	Long	Long	μ_6

Contrast for Main Effect of Wing Length

Wing Length is more complex because it has 3 levels (df=2). We need a two-level contrast statement for testing the main effect of Wing Length.

Contrast – Wing Length (Short vs. Medium vs. Long):

$$\frac{\mu_1+\mu_4}{2} - \frac{\mu_2+\mu_5}{2} \text{ AND } \frac{\mu_2+\mu_5}{2} - \frac{\mu_3+\mu_6}{2}$$

Reject $H_0: \psi = 0$ if either of the two contrast statements produces a significant test result

$$\psi = \begin{bmatrix} .5 & -.5 & 0 & .5 & -.5 & 0 \\ 0 & .5 & -.5 & 0 & .5 & -.5 \end{bmatrix}$$

We only need to compare Short vs. Medium and Medium vs. Long. The comparison of Short vs. Long is implied by summing the two rows. < Attendance Quiz: Contrast Coefficients >

Using a contrast to test interaction effects

We can even use contrast statements to estimate and test the synergistic effect of multiple factors. This is referred to as an *interaction effect*.

What is an interaction?

Definition: There is an interaction between effects of factors A and B if the effect of factor A is dependent on the levels of factor B.

Examples of Interaction Effects

Example: Effect of Study Time and Sleep on Exam Performance

Consider a study examining how study time and sleep duration affect exam scores.

Explanatory Variables:

- Study Time (low vs. high)
- Sleep Duration (short vs. long)

Response Variable:

- Exam Score

Main Effects:

- Study Time: More study time → higher exam scores (on average)
- Sleep Duration: More sleep → higher exam scores (on average)

Interaction Effect:

- The benefit of studying more only occurs when students get enough sleep. If a student sleeps too little, even high study time doesn't lead to better performance.

Examples of Interaction Effects

Example: Drug Effectiveness and Age Group

Researchers are studying how a new drug affects *blood pressure reduction*, and they want to know whether its effectiveness depends on the *age group* of the patients.

Explanatory Variables:

- Treatment Type (Drug vs. Placebo)
- Age Group (Under 50 vs. Over 50)

Response Variable:

- Reduction in Blood Pressure

Main Effects:

- Drug: On average, the drug lowers blood pressure more than the placebo.
- Age Group: Older individuals may have higher blood pressure, so the baseline might vary.

Interaction Effect:

- The drug works very well in people under 50 but barely has any effect in those over 50.

Examples of Interaction Effects

Example: Teaching Method and Student Learning Style

A school is evaluating how different *teaching methods* affect *student test performance*, and whether the effect depends on the *student's learning style*.

Explanatory Variables:

- Teaching Method: Lecture-based vs. Hands-on
- Learning Style: Visual learner vs. Kinesthetic learner

Response Variable:

- Test Performance Score

Main Effects:

- No main effects appear due to both teaching methods and both learning styles yielding similar results on average.

Interaction Effect:

- Visual learners do better with lecture-based teaching while kinesthetic learners do better with hands-on teaching.

Using a contrast to test interaction effects

We can use contrast statements to estimate and test interaction effects.

The interaction has df=2. So, it will require two rows. The interaction means there is a difference in the levels of Wing Length at the different levels of Body Length.

- (Short vs. Medium Wing Length for Short Body Length) vs. (Short vs. Medium Wing Length for Long Body Length)
- (Medium vs. Long Wing Length for Short Body Length) vs. (Medium vs. Long Wing Length for Long Body Length)

$$\frac{\mu_1-\mu_2}{2}-\frac{\mu_4-\mu_5}{2} \text{ AND } \frac{\mu_2-\mu_3}{2}-\frac{\mu_5-\mu_6}{2}$$

$$\psi=\begin{bmatrix} .5 & -.5 & 0 & -.5 & .5 & 0 \\ 0 & .5 & -.5 & 0 & -.5 & .5 \end{bmatrix}$$

Helicopter Number	Wing Length	Body Length	Mean
1	Short	Short	μ_1
2	Medium	Short	μ_2
3	Long	Short	μ_3
4	Short	Long	μ_4
5	Medium	Long	μ_5
6	Long	Long	μ_6

Reject $H_0: \psi = 0$ if either of the two contrast statements produces a significant test result

Using a contrast to test polynomial effects

We can use contrast statements to estimate and test polynomial effects.

For example, we can test to see whether there is a linear, quadratic or cubic pattern to the means. As an example, we may suspect there to be a linear pattern in the Wing Lengths, a consistent increasing time as Wing Length increases.

The polynomial contrast coefficients are as follows:

- Linear: -1, 0, 1, -1, 0, 1
- Quadratic: 1, -2, 1, 1, -2, 1

Helicopter Number	Wing Length	Body Length	Mean
1	Short	Short	μ_1
2	Medium	Short	μ_2
3	Long	Short	μ_3
4	Short	Long	μ_4
5	Medium	Long	μ_5
6	Long	Long	μ_6

Considerations

- The number of rows in a contrast is limited to the number of degrees of freedom associated with that factor. Because we had 6 helicopter versions, we have df=5 available for contrasts. That would allow us to measure the main effect of Wing Length (2 rows), the main effect of Body Length (1 row), and the interaction effect (2 rows). Conversely, we could have chosen 5 rows of different contrasts to perform.
- The constants (contrast coefficients) we multiply the means by all must add to zero. Note that both rows in the matrix below add to zero.
- The values of the constants are such that you are testing differences of weighted averages.

$$\psi = \begin{bmatrix} .5 & -.5 & 0 & -.5 & .5 & 0 \\ 0 & .5 & -.5 & 0 & -.5 & .5 \end{bmatrix}$$

< R example: Helicopter Experiment.qmd >

3.2 Multiple Testing

A researcher would like to compare the durability of several different oil filter brands. 5 brands were tested using an ANOVA F-test. We rejected H_o implying that the mean durability is not the same for all brands. But where are the differences? The researcher decided to look at all 10 possible pairwise comparisons (Brands 1 vs. 2, Brands 1 vs. 3, Brands 1 vs. 4, …).

If the probability of making a Type-1 error for each test is $\alpha = 0.05$, what is the probability of making at least one Type-1 error among all 10 tests?

Let's take a look!

3.2 Multiple Testing

If we conduct just one test, we can set the probability of a Type-1 error to a nominal level, $\alpha = \alpha_0 = 0.05$.

However, if we conduct multiple tests with the same data from the same experiment, we are more likely to find at least one significant difference even when such differences do not actually exist.

The following equation represents the probability of committing at least one Type-1 error when conducting m tests at a nominal significance level of α_0.

$$\alpha = 1 - (1 - \alpha_0)^m$$

In our example of oil filter brands, we are making 10 comparisons. If each comparison is tested at a nominal $\alpha_0 = 0.05$ level, the probability of committing at least one Type-1 error is

$$\alpha = 1 - (1 - \alpha_0)^m = 1 - (0.95)^{10} = 0.4013$$

That's a 40% chance of making at least one Type-1 error!!!

3.2 Multiple Testing

- We need to make an adjustment if we want the probability of making at least one Type-1 error to be ≤ 0.05.

- **Comparison-wise error rate (CER):** The CER is the error rate associated with a single comparison.

- **Family-wise error rate (FWE):** The FWE is the error rate (chance of at least one Type-1 error) associated with a group or "family" of comparisons or contrasts that are to be made.

- Basic idea:
  - Instead of using standard p-value, adjust it (make it bigger) while still comparing it to $\alpha = 0.05$ to account for multiplicity.

3.2 Multiple Testing

- **False Discovery Rate (FDR)** is another error rate.
  - Represents the percentage of our H_0 rejections that are actually Type-1 errors.
  - Ex. FDR=0.20 implies that 20% of "significant findings" are not "true findings" but rather "false positives."

- $(1-\alpha)\times 100\%$ **Simultaneous confidence intervals:** The probability that all intervals simultaneously contain the true parameter is $(1-\alpha)$.

What does the F-test tell us?

- ANOVA allows us to test the equality of means/effects among levels of a factor(s)
- Example: Helicopter Experiment

$$H_0: Wing\ length\ does\ not\ significantly\ affect\ the\ the\ flight\ time$$
$$H_A: Wing\ length\ significantly\ affects\ the\ flight\ time$$

That is the same as the following hypotheses:

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_A: At\ least\ one\ \neq$$

- If we reject the null, we know there is a difference, but we don't know where that difference is
- We need to perform follow-up tests to determine where the differences are

Pairwise comparisons can show differences between pairs of means

- For a significant test of hypothesis in the ANOVA, we can follow up with pairwise comparisons
- Here are the pairwise comparisons for a significant wing length effect:

$$H_0: \mu_1 = \mu_2$$
$$H_0: \mu_1 = \mu_3$$
$$H_0: \mu_2 = \mu_3$$

We can use the following formula to test these comparisons:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MSError}} \sim t_{df\,(Error)}$$

How does this formula differ from a standard t-test for 2 means?

Pairwise comparison adjustments can help maintain accurate Type-1 error rates

Pairwise comparison adjustments: Assume g groups

METHOD	ADJUSTMENT	WHEN TO USE	SIZE
Fisher's LSD	$t_{\alpha/2,dfError}$ (no adjustment)	Screening study (you're willing to live with high familywise error rate)	Most narrow
Tukey's HSD	q/sqrt{2} (q is based on expected difference between highest group mean minus lowest group mean when in fact all g of the group means are equal—depends on g and dfError)	Looking at all pairwise (group mean) comparisons	Between Scheffe' and LSD
Scheffe'	sqrt{ $(F_{\alpha,g-1,dfError})(g-1)$ } (uses the F critical value you would use for testing group)	Post hoc comparisons (data snooping)	Widest (usually)
Bonferroni	$t_{\alpha/2k,dfError}$ (where k is the number of comparisons you have planned)	Limited number of *a priori* comparisons	Narrower than Tukey if k is small

*Dunnett's test is used to compare multiple treatments vs. a control.

Helicopter Experiment Example

Example Results

ANOVA:

```
             Df Sum Sq Mean Sq F value   Pr(>F)
WingLength    2  3.573  1.7863  78.978  < 2e-16 ***
BodyLength    1  0.034  0.0337   1.492    0.225
Team          3  0.826  0.2753  12.171 9.54e-07 ***
Residuals    89  2.013  0.0226
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There are differences!!!

Body Length

Example Results

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MSError}}$$

Long vs. Short

$$t = \frac{1.63125 - 1.59375}{\sqrt{\left(\frac{1}{48} + \frac{1}{48}\right) 0.0226}}$$

t = 1.222032

df = 89

CV = 1.98828

P-value =0.2249223

Wing Length: Short, Medium, Long

Example Results

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MSError}}$$

Medium vs. Long

$$t = \frac{1.51875 - 1.88125}{\sqrt{\left(\frac{1}{32} + \frac{1}{32}\right) 0.0226}}$$

t = -9.64526

df = 89

CV = 1.98828

P-value = 0

Wing Length: Short, Medium, Long

Example Results

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MSError}}$$

Short vs. Long

$$t = \frac{1.43750 - 1.88125}{\sqrt{\left(\frac{1}{32} + \frac{1}{32}\right) 0.0226}}$$

t = -11.8071

df = 89

CV = 1.98828

P-value = 0

Wing Length: Short, Medium, Long

Example Results

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MSError}}$$

Short vs. Medium

$$t = \frac{1.43750 - 1.51875}{\sqrt{\left(\frac{1}{32} + \frac{1}{32}\right) 0.0226}}$$

t = -2.16187

df = 89

CV = 1.98828

P-value = 0.03331162

Team: Back Row Crew, Data Titans, Factor Fury, P-value Posse

Example Results

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) MSError}}$$

Back Row Crew vs. Data Titans

$$t = \frac{1.537500 - 1.658333}{\sqrt{\left(\frac{1}{24} + \frac{1}{24}\right) 0.0226}}$$

t = -2.78434

df = 89

CV = 1.98828

P-value = 0.00655162

Other comparisons would follow a similar pattern

Due to four teams, there are $\binom{4}{2} = 6$ combinations of pairs of means to compare

Why don't research articles make adjustments for multiple comparisons or multiple tests?

Research articles do not tend to include ANY adjustments for multiple tests.

They assume the reader will make the adjustments themselves.

So, be cautious.

Also, beware of p-hacking and fishing expeditions to identify significant differences that are likely just false positives.