Course Business

- Midterm assignment due on Canvas next Monday at 1:30 PM
  - Unsure if an article is suitable? Can run it by me

- Add-on packages to install for today:
  - performance (may have gotten this last week)
  - emmeans

- Lab materials on Canvas
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
  - Weighted Coding
  - Unweighted Coding
- Post-Hoc Comparisons
  - Tukey Test
  - Estimated Marginal Means
  - Comparing Marginal Means
- Lab
**Unbalanced Factors**

- Sometimes, we may have **differing numbers of observations** per level.
- Possible reasons:
  - Some categories **naturally** more common.
    - e.g., college majors.
  - Categories may be equally common in the population, but we have **sampling error**.
    - e.g., ended up 60% female participants, 40% male.
  - Study was **designed** so that some conditions are more common.
    - e.g., more “control” subjects than “intervention” subjects.
  - We wanted equal numbers of observations, but lost some because of **errors or exclusion criteria**.
    - e.g., data loss due to computer problems.
    - Dropping subjects below a minimum level of performance.
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
  - Weighted Coding
    - Unweighted Coding
- Post-Hoc Comparisons
  - Tukey Test
  - Estimated Marginal Means
  - Comparing Marginal Means
- Lab
Weighted Coding

• “For the average student, does course size predict probability of graduation?”
  • Random sample of 200 Pitt undergrads
  • 5 are student athletes and 195 are not

• How can we make the intercept reflect the “average student”?
  • We could try to apply effects coding to the StudentAthlete variable by centering around the mean and getting (0.5, -0.5), but...
Weighted Coding

- An intercept at 0 would no longer correspond to the overall mean
  
  .5
  
  ATHLETE (5)
  
  Zero is here
  
  0
  
  - .475
  
  NOT ATHLETE (195)
  
  But “not athlete” is actually far more common
  
- As a scale, this would be totally unbalanced
- To fix balance, we need to assign a heavier weight to Athlete
• Change codes so the mean is 0
• \( c(0.975, -0.025) \)
• `contr.helmert.weighted()` function in my `psycholing` package will calculate this
**Weighted Coding**

- **Weighted coding**: Change the codes so that the mean is 0 again
  - Used when the imbalance reflects something *real*
  - Like Type II sums of squares

- “For the average student, does course size predict graduation rates?”
  - Average student is *not* a student athlete, and our answer to the question about an “average student” should reflect this!
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
- Weighted Coding
- Unweighted Coding
- Post-Hoc Comparisons
  - Tukey Test
  - Estimated Marginal Means
  - Comparing Marginal Means
- Lab
Unweighted Coding

• Last week we looked at *aphasia.csv*:
  • Response times (RT) in a sentence verification task
  • Effect of *SubjectType* (aphasia vs. control)
  • Effect of *SentenceType* (active vs. passive)
  • And their interaction
Unweighted Coding

• Oops! Our experiment loaded up the wrong image for one of our Passive sentences (“Groceries”)
  • It may have been sabotaged

  • `UsableItem` column is **No** for this item

• First, can we remove this from our data?
• Some possibilities:
  • `aphasia %>% filter(UsableItem == 'Yes') -> aphasia`
  • `aphasia %>% filter(UsableItem != 'No') -> aphasia2`
  • etc.
Unweighted Coding

- Oops! Our experiment loaded up the wrong image for one of our Passive sentences ("Groceries")

- Now, there’s an imbalance, but it’s an accident and not meaningful
  - In fact, we’d like to get rid of it!

```
> summary(aphasia2)

   Subject Item SubjectType SentenceType
S1    : 31  Astronaut: 30 Aphasia:465  Active :480
S11   : 31    Boy : 30
S12   : 31  Breakfast: 30
S13   : 31   Burglar : 30
S14   : 31   Cheese : 30
(Other):744 (Other) :750
```
Unweighted Coding

• Oops! Our experiment loaded up the wrong image for one of our Passive sentences (“Groceries”)

• Now, there’s an imbalance, but it’s an accident and not meaningful
  • In fact, we’d like to get rid of it!

• Retain the \((-0.5, 0.5)\) codes
  • Weights the two conditions equally—because the imbalance isn’t meaningful
  • Like Type III sums of squares
  • Probably what you want for factorial experiments
Unbalanced Factors: Summary

• Weighted coding: Change the codes so that the mean is 0
  • Use when the imbalance reflects something real
  • Can be done with `contr.helmert.weighted()`

Mean across each individual: 🧑‍🤝‍🧑‍🤝‍🧑

• Unweighted coding: Keep the codes as -0.5 and 0.5
  • Use when the imbalance is an accident that we want to eliminate

Mean of the two levels: 🧑‍🤝‍🧑‍🤝‍🧑

• With balanced factors, these are identical
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
  - Weighted Coding
  - Unweighted Coding
- Post-Hoc Comparisons
  - Tukey Test
  - Estimated Marginal Means
  - Comparing Marginal Means
- Lab
**Post-hoc Comparisons**

- Maximal model for the aphasia data was:
  - `model.Maximal <- lmer(RT ~ 1 + SentenceType * SubjectType + (1 + SentenceType|Subject) + (1 + SubjectType|Item), data = aphasia)`

- This didn’t converge:
  - `boundary (singular) fit: see ?isSingular`

```
Random effects:
Groups     Name        Variance  Std.Dev.  Corr
Item       (Intercept) 44860.91   211.804
           SentenceType  71.18    8.437   -1.00
Subject    (Intercept) 43286.19   208.053
           SentenceType  1350.57  36.750   0.18
Residual               6861.01   82.831
Number of obs: 930, groups: Item, 31; Subject, 30
```

Probably overparameterized
Post-hoc Comparisons

• Maximal model for the aphasia data was:
  ```r
  model.Maximal <- lmer(RT ~ 1 + SentenceType * SubjectType + (1 + SentenceType|Subject) + (1 + SubjectType|Item),
                       data = aphasia)
  ```

• So, let’s simplify:
  ```r
  model2 <- lmer(RT ~ 1 + SentenceType * SubjectType + (1 + SentenceType|Subject) + (1|Item),
                 data = aphasia)
  ```

• Doesn’t seem to harm model fit—\( p > .20 \)
  ```r
  anova(model.Maximal, model2)
  ```

<table>
<thead>
<tr>
<th></th>
<th>npar</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>Chisq</th>
<th>Df</th>
<th>Pr(&gt;Chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>model2</td>
<td>9</td>
<td>11214</td>
<td>11257</td>
<td>-5597.9</td>
<td>11196</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>model.Maximal</td>
<td>11</td>
<td>11216</td>
<td>11269</td>
<td>-5596.8</td>
<td>11194</td>
<td>2.2473</td>
<td>2</td>
<td>0.3251</td>
</tr>
</tbody>
</table>
Post-hoc Comparisons

- With treatment coding, we get estimates of simple-effects:

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
<th>Corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
<td>(Intercept)</td>
<td>43000</td>
<td>207.58</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td>(Intercept)</td>
<td>43286</td>
<td>208.05</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SentenceTypePassive</td>
<td>1348</td>
<td>36.72</td>
<td>0.18</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>6879</td>
<td>82.94</td>
<td></td>
</tr>
<tr>
<td>Number of obs</td>
<td></td>
<td>930</td>
<td>groups:</td>
<td>Item, 31; Subject, 30</td>
</tr>
</tbody>
</table>

Fixed effects:

|                     | Estimate | Std. Error | df | t value | Pr(>|t|) |
|---------------------|----------|------------|----|---------|---------|
| (Intercept)         | 1716.01  | 74.88      | 56.54 | 22.916  | < 2e-16 |
| SentenceTypePassive | 553.44   | 75.60      | 30.23 | 7.321   | 3.57e-08 *** |
| SubjectTypeAphasia  | 84.52    | 76.35      | 78.00 | 1.107   | 0.278   |
| SentenceTypePassive:SubjectTypeAphasia | 188.30 | 17.27 | 28.00 | 10.904 | 1.38e-11 *** |

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

  (Intr) SntnTP SbjcTA
SntncTypPss -0.465
SbjctTypAph -0.510 -0.011
SntncTP:STA -0.049 -0.114  0.095

Intercept: RT for healthy controls, active sentences

Significant RT difference for passive sentences (among healthy controls)

Not a significant RT difference for aphasics (among active sentences)

Significant special effect of aphasia + passive sentence
**Post-hoc Comparisons**

- The estimates from a model are enough to *fully describe* differences among conditions.
- With simple effects:
  - \(\text{ACTIVE, CONTROL} \approx 1716 \text{ ms}\)
  - \(\text{PASSIVE, CONTROL} \approx 2269 \text{ ms}\)

SubjectType: Aphasia, Control

SentenceType: Active, Passive

- SubjectType: Aphasia, Control
- SentenceType: Active, Passive
- ACTIVE, CONTROL: RT ≈ 1716 ms
  - Passive simple effect: +553 ms
- PASSIVE, CONTROL: RT ≈ 2269 ms
Post-hoc Comparisons

- The estimates from a model are enough to **fully describe** differences among conditions.
- With simple effects:

<table>
<thead>
<tr>
<th>Subject Type</th>
<th>Sentence Type</th>
<th>ACTIVE, APHASIA</th>
<th>RT ≈ 1801 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphasia</td>
<td>Active</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>Passive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Subject Type</th>
<th>Sentence Type</th>
<th>ACTIVE, CONTROL</th>
<th>RT ≈ 1716 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphasia</td>
<td>Active</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>Passive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

  Aphasias simple effect +85 ms

  Passive simple effect +553 ms

  PASSIVE, CONTROL
  RT ≈ 2269 ms
Post-hoc Comparisons

- The estimates from a model are enough to fully describe differences among conditions
- With simple effects:

<table>
<thead>
<tr>
<th>SubjectType</th>
<th>SentenceType</th>
<th>ACTIVE, APHASIA</th>
<th>RT ≈ 1801 ms</th>
<th>PASSIVE, APHASIA</th>
<th>RT ≈ 2542 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphasia</td>
<td>Active</td>
<td>+553 ms</td>
<td></td>
<td>+85 ms</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>Passive</td>
<td>-188 ms</td>
<td>+553 ms</td>
<td>-85 ms</td>
<td>-85 ms</td>
</tr>
</tbody>
</table>

Active, Aphasia simple effect: +553 ms
Passive, Aphasia simple effect: +85 ms
Interaction effect: +188 ms
Active, Control RT ≈ 1716 ms
Passive, Control RT ≈ 2269 ms
Post-hoc Comparisons

- But, sometimes we want to compare individual combinations (e.g., people with aphasia seeing active vs passive sentences)
- i.e., individual cells

<table>
<thead>
<tr>
<th>Subject Type</th>
<th>Sentence Type</th>
<th>ACTIVE, APHASIA</th>
<th>PASSIVE, APHASIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphasia</td>
<td>Active</td>
<td>RT ≈ 1801 ms</td>
<td></td>
</tr>
<tr>
<td>CONTROL</td>
<td>Passive</td>
<td>RT ≈ 2542 ms</td>
<td></td>
</tr>
<tr>
<td>Aphasia</td>
<td>Control</td>
<td>RT ≈ 1716 ms</td>
<td></td>
</tr>
<tr>
<td>CONTROL</td>
<td></td>
<td>RT ≈ 2269 ms</td>
<td></td>
</tr>
</tbody>
</table>
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
  - Weighted Coding
  - Unweighted Coding
  - Post-Hoc Comparisons
    - Tukey Test
      - Estimated Marginal Means
      - Comparing Marginal Means

- Lab
Post-hoc Comparisons: Tukey Test

• But, sometimes we want to compare individual combinations (e.g., people w/ aphasia seeing active vs passive sentences)
  • i.e., individual cells

• \texttt{emmeans(Model.Maximal, pairwise~SentenceType*SubjectType)}
  • Requires \texttt{emmeans} package to be loaded
    • \texttt{library(emmeans)}
  • Uses Tukey test to correct for multiple comparisons so overall $\alpha$ still = .05

• Which two cells don’t significantly differ?
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
  - Weighted Coding
  - Unweighted Coding
- Post-Hoc Comparisons
  - Tukey Test
- Estimated Marginal Means
  - Comparing Marginal Means
- Lab
Estimated Marginal Means

- `emmeans` also returns estimated means and std. errors for each cell of the design

- EMMs represent what the means of the different groups would look like if they didn’t differ in other (fixed or random) variables

<table>
<thead>
<tr>
<th>SentenceType</th>
<th>SubjectType</th>
<th>emmean</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>Control</td>
<td>1716</td>
<td>75.6</td>
<td>56.4</td>
<td>1565</td>
<td>1867</td>
</tr>
<tr>
<td>Passive</td>
<td>Control</td>
<td>2269</td>
<td>78.6</td>
<td>56.3</td>
<td>2112</td>
<td>2427</td>
</tr>
<tr>
<td>Active</td>
<td>Aphasia</td>
<td>1801</td>
<td>74.2</td>
<td>56.1</td>
<td>1652</td>
<td>1949</td>
</tr>
<tr>
<td>Passive</td>
<td>Aphasia</td>
<td>2542</td>
<td>77.1</td>
<td>56.0</td>
<td>2388</td>
<td>2697</td>
</tr>
</tbody>
</table>
Estimated Marginal Means

- As an example, let’s add *SentenceLength* as a covariate
  - Task was to read the sentence & judge whether it matches a picture, so length of sentence would plausibly affect time needed to do this

- *Not* perfectly balanced across conditions
Estimated Marginal Means

- New model & estimated marginal means:
  - `model.Length <- lmer(RT ~ 1 + SentenceType * SubjectType + SentenceLength + (SentenceType|Subject) + (1|Item), data=aphasia)
  - `emmeans(model.Length, pairwise~SentenceType*SubjectType + SentenceLength)`

<table>
<thead>
<tr>
<th>SentenceType</th>
<th>SubjectType</th>
<th>SentenceLength</th>
<th>emmean</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>Control</td>
<td></td>
<td>9.9</td>
<td>1676</td>
<td>54.3</td>
<td>28.7</td>
<td>1564</td>
</tr>
<tr>
<td>Passive</td>
<td>Control</td>
<td></td>
<td>9.9</td>
<td>2313</td>
<td>56.8</td>
<td>28.7</td>
<td>2196</td>
</tr>
<tr>
<td>Active</td>
<td>Aphasia</td>
<td></td>
<td>9.9</td>
<td>1760</td>
<td>54.3</td>
<td>28.7</td>
<td>1649</td>
</tr>
<tr>
<td>Passive</td>
<td>Aphasia</td>
<td></td>
<td>9.9</td>
<td>2585</td>
<td>56.8</td>
<td>28.7</td>
<td>2469</td>
</tr>
</tbody>
</table>

What we’d expect *if* the sentences were all of equal length

<table>
<thead>
<tr>
<th>SentenceType</th>
<th>SubjectType</th>
<th>SentenceLength</th>
<th>M</th>
<th>RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>Control</td>
<td></td>
<td>10.2</td>
<td>1716</td>
</tr>
<tr>
<td>Active</td>
<td>Aphasia</td>
<td></td>
<td>10.2</td>
<td>1801</td>
</tr>
<tr>
<td>Passive</td>
<td>Control</td>
<td></td>
<td>9.53</td>
<td>2269</td>
</tr>
<tr>
<td>Passive</td>
<td>Aphasia</td>
<td></td>
<td>9.53</td>
<td>2542</td>
</tr>
</tbody>
</table>

Our raw data (where the average sentence length differs slightly across conditions)
**Estimated Marginal Means**

- EMMs are a hypothetical
  - What the means of the different groups would look like if they didn’t differ in this covariate
  - Like *unweighted* coding in the case of missingness

- Based on our statistical model, so if our model is wrong (e.g., we picked the wrong covariate), the adjusted means will be too
  - Unlike the raw sample means
  - Need to use some caution in interpreting them
  - Be clear what you are reporting
Estimated Marginal Means

- Also possible to test whether each of these estimated cell means significantly differs from 0
  - \texttt{ls\_means(model.Length)}
  - Silly in case of RTs, but could be relevant for some other DVs (e.g., preference)

<table>
<thead>
<tr>
<th>Least Squares Means table:</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t value</th>
<th>lower</th>
<th>upper</th>
<th>Pr(&gt;\mid t\mid)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SentenceTypeActive</td>
<td>1717.838</td>
<td>38.670</td>
<td>29.4</td>
<td>44.423</td>
<td>1638.801</td>
<td>1796.876</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SentenceTypePassive</td>
<td>2448.983</td>
<td>40.440</td>
<td>29.4</td>
<td>60.559</td>
<td>2366.324</td>
<td>2531.641</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SubjectTypeControl</td>
<td>1994.074</td>
<td>55.080</td>
<td>28.4</td>
<td>36.204</td>
<td>1881.312</td>
<td>2106.837</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SubjectTypeAphasia</td>
<td>2172.747</td>
<td>55.080</td>
<td>28.4</td>
<td>39.447</td>
<td>2059.985</td>
<td>2285.509</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SentenceTypeActive:SubjectTypeControl</td>
<td>1675.578</td>
<td>54.338</td>
<td>28.7</td>
<td>30.836</td>
<td>1564.398</td>
<td>1786.757</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SentenceTypePassive:SubjectTypeControl</td>
<td>2312.571</td>
<td>56.833</td>
<td>28.7</td>
<td>40.690</td>
<td>2196.283</td>
<td>2428.860</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SentenceTypeActive:SubjectTypeAphasia</td>
<td>1760.099</td>
<td>54.338</td>
<td>28.7</td>
<td>32.392</td>
<td>1648.919</td>
<td>1871.279</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>SentenceTypePassive:SubjectTypeAphasia</td>
<td>2585.394</td>
<td>56.833</td>
<td>28.7</td>
<td>45.491</td>
<td>2469.106</td>
<td>2701.683</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
</tbody>
</table>
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
  - Weighted Coding
  - Unweighted Coding
- Post-Hoc Comparisons
  - Tukey Test
  - Estimated Marginal Means
  - Comparing Marginal Means
- Lab
Comparing Marginal Means

- `emmeans` can also test marginal means:
  - `emmeans(model2, pairwise~SubjectType)`

Effect of one variable *averaging over* the other
- e.g., aphasic participants (averaging over all sentence types) vs. controls (averaging over all sentence types)
- These are what *main effects* are testing

Results are averaged over the levels of: SentenceType
Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95

Now, include just one variable (for which we want marginal means)
Week 8.1: Post-Hoc Comparison

- Unbalanced Factors
  - Weighted Coding
  - Unweighted Coding
- Post-Hoc Comparisons
  - Tukey Test
  - Estimated Marginal Means
  - Comparing Marginal Means

Lab