Two-way ANOVA

Last week we learned about one-way designs. They examine several levels of a single factor. For instance, we might examine the effect of StudyType (independent variable) on LanguagePerformance (dependent variable).

<table>
<thead>
<tr>
<th>Classroom study</th>
<th>Self-study</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Perhaps we think that the StudyType manipulation might only affect one gender or the other. We could do a two-way design whether a second factor, Gender.

<table>
<thead>
<tr>
<th>Female</th>
<th>Classroom study</th>
<th>Self-study</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Male</th>
<th>Classroom study</th>
<th>Self-study</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Factorial design

In factorial design the dependent variable (score on the Cambridge English Proficiency test) is sampled in every possible combination of the factors. That is, each test score is cross-classified as coming from someone who is e.g. Female AND European or Male AND Southeast Asian etc.

<table>
<thead>
<tr>
<th>Geographical location</th>
<th>Europe (1)</th>
<th>South America (2)</th>
<th>North Africa (3)</th>
<th>South East Asia (4)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male (1)</td>
<td>10</td>
<td>33</td>
<td>26</td>
<td>26</td>
<td>96</td>
</tr>
<tr>
<td>Female (2)</td>
<td>24</td>
<td>25</td>
<td>19</td>
<td>25</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>32</td>
<td>31</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>16</td>
<td>15</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Subtotal</td>
<td>99</td>
<td>127</td>
<td>116</td>
<td>105</td>
<td>447</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>16</td>
<td>25</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>13</td>
<td>32</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>23</td>
<td>20</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>20</td>
<td>15</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Subtotal</td>
<td>151</td>
<td>92</td>
<td>115</td>
<td>108</td>
<td>466</td>
</tr>
<tr>
<td>Total</td>
<td>250</td>
<td>219</td>
<td>231</td>
<td>213</td>
<td>913</td>
</tr>
</tbody>
</table>

\[ Y_{ij} = \text{total score of subjects belonging to the } i\text{-th location and } j\text{-th sex (e.g. } Y_{31} = 116) \]
\[ Y_{i.} = \text{total score of subjects at } i\text{-th location (} Y_{2.} = 219) \]
\[ Y_{.j} = \text{total score of subjects of } j\text{-th sex (} Y_{2.} = 466) \]
\[ Y_{..} = \text{grand total } = 913 \]
Interaction

Consider this made-up data from a study that manipulates two factors, A and B. The former has just two levels “1” and “2” while the latter has three levels “one”, “two” and “three”.

<table>
<thead>
<tr>
<th></th>
<th>$B_{one}$</th>
<th>$B_{two}$</th>
<th>$B_{three}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>10</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>$A_2$</td>
<td>15</td>
<td>20</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$B_{one}$</th>
<th>$B_{two}$</th>
<th>$B_{three}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>8</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>$A_2$</td>
<td>11</td>
<td>16</td>
<td>21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$B_{one}$</th>
<th>$B_{two}$</th>
<th>$B_{three}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>7</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>$A_2$</td>
<td>9</td>
<td>14</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$B_{one}$</th>
<th>$B_{two}$</th>
<th>$B_{three}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>8</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>$A_2$</td>
<td>16</td>
<td>13</td>
<td>9</td>
</tr>
</tbody>
</table>

Figure 1: Positive and negative examples of interaction

An interaction is observed when the model needs to take into account not just an additional treatment factor $\beta$ but also a multiplicative factor $\alpha_i \beta_j$ that explains how the efficacy of one factor changes in the presence of the other.

$$X_{ijk} = \mu + \alpha_i + \beta_j + \alpha_i \beta_j + \epsilon_{ijk}$$

two-factor model with interaction term

Graphically, an interaction is indicated when connected lines on a plot of the dependent variable are non-parallel. The made-up data in figure show a case of interaction and lack thereof.
> rietveld <- read.table(file = "rietveld.txt", header = T)
> interaction.plot(x.factor = rietveld$B, trace.factor = rietveld$A,
+     response = rietveld$score, ylim = c(0, 10))

![interaction plot]

**Figure 2:** A very strong (but made-up) interaction.

To use R to test the hypothesis that there is an interaction in the fake data from figure 2, include an interaction term in the model formula, for instance with a colon.

> summary(aov(score ~ A + B + A:B, data = rietveld))

```
Df  Sum Sq Mean Sq F value  Pr(>F)
A   1  7.89e-31 7.89e-31 7.89e-31    1
B   1  1.78e-30 1.78e-30 1.78e-30    1
Residuals 8     8     1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a statistically-significant interaction — Aone improves tremendously with B, whereas Atwo reverses the effect. Because the interaction is in equal and opposite directions, there are no **main effects**.
Random vs. fixed effects

In a statistical model such as $X_{ij} = \mu + \alpha_j + \epsilon_{ij}$ the factor $\alpha_j$ is a fixed effect e.g. the expected number of additional errors you will make on your Cambridge English Proficiency test if you are unlucky enough to come from South America. There is a presumption that all possible continents-of-origin are taken into account. But in a lexical decision experiment, for instance, only a subset of the possible lexical items are tested. Do observed results extend to the other words of the language? [Clark 1973] argued that the particular set of words actually tested in the experiment should be viewed as a random sample from the universe of possible words — to which we would like to generalize! Similarly, if I gauge the socio-economic status (SES) of the next 30 people who walk into the MSU union, can I make a claim about the SES of the MSU community in general? These situations call for statistical model that include a random effect.

In the simpler one-way case, we wish to compare two statistical models against each other. One model includes a term $\alpha_j$ for the random effect, the other doesn’t.

\[
X_{ij} = \mu + \epsilon_{ij} \quad \text{Model 0 “Restricted”}
\]

\[
X_{ij} = \mu + \alpha_j + \epsilon_{ij} \quad \text{Model 1 “Full”}
\]

Postulating that $\alpha$ is a random effect means that the experiment has sampled a particular selection; maybe the next experiment will get some different values. If this factor is truly random it will have nonzero variance. The hypotheses thus concern the random effect $\alpha$’s variance.

\[
H_0 \quad \sigma^2_\alpha = 0 \quad \text{Model 0}
\]

\[
H_1 \quad \sigma^2_\alpha > 0 \quad \text{Model 1}
\]

Regardless of whether a factor is fixed or random, the mean-square within $MS_w$ serves as an estimate of error variance (equation 1). I will happily furnish a copy of page 111 from Maxwell & Delaney to those who are interested in seeing the derivation.

\[
E(MS_w) = \sigma^2_\epsilon \quad (1)
\]

In just the random effects case, the mean-square between $MS_b$ will depend on the amount of variability present in our random sample of levels $j$ (equation 2).

\[
E(MS_b) = \sigma^2_\epsilon + n\sigma^2_\alpha \quad (2)
\]

By contrast, in the fixed effects case, the amount that the $F$ ratio diverges from 1.00 is a linear combination of the effect size of a level $j$ and the number of measurements at that level (equation 3).

\[
E(MS_b) = \sigma^2_\epsilon + \frac{\sum_j n_j \alpha^2_j}{a-1} \quad (3)
\]

So, the procedure for hypothesis testing does not change in the transition to fixed effects, however the reason why it works is slightly different.

\[
F = \frac{MS_b}{MS_w}
\]

If your design contains both fixed effects and random effects, it is called mixed. This would be true if you are randomly choosing participants for your study, but you are administering them fixed numbers of shots of cheap liquor. Say we want to test for a main effect of NumberOfLiquorShots on TableTennisScore, as played against a robotic opponent. Here, if we divide by $MS_w$ we would be performing a test of either the main effect of NLS or the interaction of NLS × Participant. In other words, we would get an artificially inflated $F$ score if a minority of Ping-Pong maniacs are actually helped by cheap liquor, whereas it dulls the reflexes and depresses the table tennis scores of most normal people. To correct this we need to divide by the appropriate denominator, one that reflects a model where the interaction of the fixed factor with the random factor is assumed to be zero.
VP Ellipsis example

Kobele, Kim, Hale and Runner presented the results of a study examining Voice mismatches in VP ellipsis at this year’s CUNY. The experiment employs the $2 \times 2 \times 2$ factorial design indicated in figure 3. We wanted to confirm that indeed mismatch along these two features, between the elided VP and its (over) antecedent would lead to decrease acceptability. The study used Magnitude Estimation of linguistic acceptability [Bard et al., 1996].

1. Voice match
   a. **Active-Active:**
      Jill betrayed Abby, and Matt did betray Abby, too.
   b. **Passive-Passive:**
      Abby was betrayed by Jill, and Matt was betrayed by Jill, too.

2. Voice mismatch
   a. **Active-Passive:**
      Jill betrayed Abby, and Matt was betrayed by Jill, too.
   b. **Passive-Active:**
      Abby was betrayed by Jill, and Matt did betray Abby, too.

3. Category match
   **VP-VP:** The report criticized Roy, but Kate didn’t criticize Roy.

4. Category mismatch
   a. **Noun-VP:**
      The criticism of Roy was harsh, but Kate didn’t criticize Roy.
   b. **Adjective-VP:**
      The report was critical of Roy, but Kate didn’t criticize Roy.

In this experiment, Subj is best conceived-of as a random effect. This is because, in magnitude estimation, participants use their own scale which is unaffected by any pressure to correspond to other participant’s chosen scales.

```r
> ME1.df <- read.table("me1-dataframe", header = T)
> ME1.df$Subj <- as.factor(ME1.df$Subj)
> head(ME1.df)

   Est Subj  Ellipsis Mismatch MismatchType
1 -0.5108256 1     Ellipsis Match     Voice
2  0.0000000 1     Ellipsis Match     Voice
3 -0.2231436 1     Ellipsis Match     Voice
4  0.3364722 1     Ellipsis Match   Category
5  0.1823216 1     Ellipsis Match     Voice
6  0.0000000 1     Ellipsis Match   Category
```
Figure 4: Ellipsis:Mismatch interaction

Figure 4 visually confirms the interaction: Mismatch is worse than Match but only in Ellipsis.
To communicate to R that your model incorporates a random effect, use the \texttt{Error} operator in a model formula to identify the “error strata.” Each error stratum acknowledges a different level of uncertainty; a separate denominator for the relevant $F$ test.

```r
> me1.aov <- aov(Est ~ Ellipsis * Mismatch * MismatchType + Error(Subj/(Ellipsis *
+ Mismatch * MismatchType)), data = ME1.df)
> summary(me1.aov)
```

```
Error: Subj
          Df Sum Sq Mean Sq F value Pr(>F)
Ellipsis 1 0.3132  0.3132   0.2545 0.6204
Ellipsis:Mismatch 1 0.0756  0.0756   0.0614 0.8072
Residuals  17 20.9161  1.2304

Error: Subj:Ellipsis
            Df Sum Sq  Mean Sq F value Pr(>F)
Ellipsis   1 22.8218 22.8218  50.9855 1.655e-06 ***
Mismatch  1  0.0030  0.0030   0.0066  0.9362
Ellipsis:Mismatch 1  0.4032  0.4032  0.9008  0.3559
Residuals 17  7.6094  0.4476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Subj:Mismatch
                Df Sum Sq  Mean Sq F value Pr(>F)
Mismatch      1 30.8742 30.8742 99.1812 1.644e-08 ***
MismatchType 1  9.308e-06 9.308e-06  2.990e-05 0.9957
Ellipsis:Mismatch 1  0.1670  0.1670  0.5365  0.4739
Residuals   17  5.2919  0.3113
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Subj:MismatchType
                       Df Sum Sq  Mean Sq F value Pr(>F)
MismatchType          1  0.9188  0.9188  4.4883 0.04916 *
Ellipsis:Mismatch 1  0.1006  0.1006  0.4914 0.49279
Ellipsis:Mismatch:MismatchType 1  0.1103  0.1103  0.5388 0.47292
Residuals      17  3.4801  0.2047
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Subj:Ellipsis:Mismatch
                                Df Sum Sq  Mean Sq F value Pr(>F)
Ellipsis:Mismatch       1 27.6301 27.6301 61.7403 4.658e-07 ***
Ellipsis:MismatchType 1  0.0057  0.0057  0.0128  0.9114
Ellipsis:Mismatch:MismatchType 1  0.3382  0.3382  0.7556  0.3968
Residuals          17  7.6079  0.4475
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Subj:Ellipsis:MismatchType
                            Df Sum Sq  Mean Sq F value Pr(>F)
Ellipsis:MismatchType 1  0.78774  0.78774  5.0378 0.0384 *
Mismatch:MismatchType 1  0.00496  0.00496  0.0318 0.8607
Ellipsis:Mismatch:MismatchType 1  0.00999  0.00999  0.0639 0.8035
Residuals       17  2.65823  0.15637
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
Disfluency and Conjunction example

Hale and Agafonova, following Levelt [1983] viewed speech repairs as a kind of conjunction and asked if the mode of coordination interacted with the like-category constraint. Four kinds of stimuli follow from completely crossing the two factors LIKE and WAY,

\[
\begin{array}{|c|c|c|}
\hline
\text{LIKE} & \text{same} & \text{different} \\
\text{WAY} & \text{conj} & a. & b. \\
 & \text{repair} & c. & d \\
\hline
\end{array}
\]

a. He wants to go to the movies and \([PP\text{to the mini-golf course.}]\)
b. He wants to go to the movies and \([\text{gerund mini-golfing.}]\)
c. He wants to go to the movies, I mean, to the mini-golf course.
d. He wants to go to the movies, I mean, mini-golfing.

The interaction plot suggests main effects of LIKE and WAY, as well as a bit of non-parallelism (sic).

```r
> ucp <- read.table(file = "~/Users/john/conj-repair/analysis/results-with-named-stimuli", + header = T)
> ucp[, 1] <- as.factor(ucp[, 1])
> head(ucp)
subj like way stimulus magnitude
1  0   same repair UCP8    10.0
2  0   same repair UCP12    8.5
3  0  diff repair UCP11     7.0
4  0   same repair UCP20    7.0
5  0   same conj UCP6      11.0
6  0  diff repair UCP15     6.0
> attach(ucp)
> interaction.plot(way, like, magnitude)
```
The **subjects analysis** F1 aggregates over subjects, taking the identity of the participant as a random effect.

```r
> ucps <- aggregate(ucp, by = list(subject = subj, like = like, + way = way), FUN = mean)
> rssubj <- aov(magnitude ~ like * way + Error(subject/(like * way)), + ucps)
> summary(rssubj)
```

```
Error: subject
  Df Sum Sq Mean Sq F value Pr(>F)
Residuals 86 530.36 6.17

Error: subject:like
  Df Sum Sq Mean Sq F value Pr(>F)
like 1 6.6553 6.6553 20.909 1.604e-05 ***
Residuals 86 27.3738 0.3183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subject:way
  Df Sum Sq Mean Sq F value Pr(>F)
way 1 32.400 32.400 52.515 1.719e-10 ***
Residuals 86 53.059 0.617
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: subject:like:way
  Df Sum Sq Mean Sq F value Pr(>F)
like:way 1 0.7572 0.7572 8.9439 0.003631 **
Residuals 86 7.2804 0.0847
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **items analysis** F2 aggregates over items, taking the particular selection of stimulus sentence as a random effect.

```r
> ucpi <- aggregate(ucp, by = list(item = stimulus, like = like, + way = way), FUN = mean)
> ritem <- aov(magnitude ~ like * way + Error(item/(like * way)), + ucpi)
> summary(ritem)
```

The **items analysis** F2 aggregates over items, taking the particular selection of stimulus sentence as a random effect.
Error: item

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>19</td>
<td>2.58659</td>
<td>0.13614</td>
<td></td>
</tr>
</tbody>
</table>

Error: item:like

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>like</td>
<td>1</td>
<td>1.71224</td>
<td>1.71224</td>
<td>18.847</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>1.72613</td>
<td>0.09085</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: item:way

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>way</td>
<td>1</td>
<td>7.5519</td>
<td>7.5519</td>
<td>34.086</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>4.2096</td>
<td>0.2216</td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Error: item:like:way

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>like:way</td>
<td>1</td>
<td>0.1661</td>
<td>0.1661</td>
<td>0.6512</td>
</tr>
<tr>
<td>Residuals</td>
<td>19</td>
<td>4.8467</td>
<td>0.2551</td>
<td></td>
</tr>
</tbody>
</table>

With only 20 items, no LIKE:WAY interaction was detected in the items analysis. This interaction did obtain by subjects (N = 87).

References


