Lab Exercise 2 - Introduction to Unbalanced Factorial Designs

When factorial designs are balanced, no one argues about them. Everyone agrees on how the ANOVA should be calculated and what it means. When a factorial design is unbalanced, all kinds of problems arise. Chief among them is that the independent variables are no longer independent.

Consider the odd monkeys experiment. That was a true experimental design in which subjects were randomly assigned to conditions of the experiment. In each cell of the design table there were four monkeys. The two variables were strength of reward (number of grapes the monkey received for a correct response) and strength of motivation (how long the monkey had been food deprived before the task). Those two variables had no chance to influence one another. That is, how long the monkey was food deprived was in no way related to how many grapes the monkey got for a correct response. All of that was determined by the experimental design.

Think of it another way. If we know what motivation condition the monkey was in, do we have any way to reasonably guess his reward condition? Here is monkey Fred. If I tell you that Fred was one of the strong motivation monkeys and asked you what reward condition Fred was in, you might throw up your hands and say, "How would I know?" And that's exactly the right answer.

Consider the following possibility. During the experiment, two of the monkeys in the strong motivation and low reward condition escape and run away to join the circus. Now I ask you about Fred. Fred is in the strong motivation condition, so what reward condition is Fred in? I'll give you $20 if you get it right.

Are you going to guess the low reward condition? Of course not! Because there are only two monkeys in that condition now, whereas there are four each in moderate and high reward. The smart money would be on one of those.

Knowing the animal's motivation condition gives you information about its reward condition. Maybe not a whole lot of information, but some. That means the independent variables are no longer independent, and that may or may not be bad. If we're dealing with a true experiment, it's bad! The independent variables are confounded with each other.

When the design is unbalanced, you have to ask why it's unbalanced. Is it unbalanced because nature is unbalanced, and representative sampling led to groups of different sizes, which nevertheless reflect the natural condition? Or is it unbalanced by accident?

Quasi-experiments are often unbalanced. We'll deal with that at another time. True experiments, which start balanced, as they certainly should, become unbalanced because subjects don't show up, subjects leave the experiment, subjects don't follow instructions,
equipment breaks down, etc. It's a law of experimental science: stuff happens! What do we do when our true experiment goes out of whack?

We will confine ourselves to the case of the 2x2 factorial design. Larger designs are more complicated, and my advice is let your software deal with it. The following method works properly ONLY for 2x2 designs, but it gives you some idea of how these unbalanced designs are dealt with.

M. Eysenck (1974) investigated incidental learning, which is learning that takes place when you are not trying to learn but are just exposed to the material. Subjects were given a list of 30 words and different instructions as to what to do with them (the "Instructions" variable):

- Count - count the letters in each word
- Rhyme - think of a word that rhymes with each word
- Adjective - think of an adjective that could modify the word
- Image - form a mental image of the object described by the word
- Control - the only group told that they would have to recall

Subjects were randomly assigned to those conditions. In addition, there was a second variable ("Age"). Half the subjects were college-aged people, and half were in their 50s and 60s. The DV was the number of words recalled from the list of 30. (Original source: Eysenck, M.W. (1974). Age differences in incidental learning. Developmental Psychology, 10(6), 936-941.)

Hold on, you protest. That's not a 2x2 design. That's 2x5. Correct. But we're only going to look at two of the instructions conditions: Count and Adjective. The point is, you don't have to think about the meaning of the material to count the letters. You do have to think about the meaning of the material to come up with an adjective. That is what Eysenck was interested in: how deeply the material was being processed by the subjects. Counting letters is shallow processing. Thinking about meaning is deep processing.

But you're still not happy, are you? We can't randomly assign people to age groups. True again. This is a minor problem, and ignoring it will not cause a problem with the analysis. There were 10 subjects in each of these groups, but we're going to imagine that two of the younger subjects in the Count condition got bored and started playing with their phones, and two of the older subjects in the Adjective condition had to leave the experiment to go to the bathroom.

But before we do that, let's look at the full data that were actually obtained by Eysenck, just to make sure we remember how to calculate a balanced ANOVA. A summary of the data appears on the next page. So a couple of questions from that: First, do you see any reason to be concerned about any of the assumptions? (Remember, unbalanced designs make the ANOVA more sensitive to assumptions, so they should be looked at carefully. We're not unbalanced yet, but we soon will be.) Second, can you draw an interaction plot from the cell means?
Here are some questions to see where you are in your understanding of these calculations.
1) Are the cell means weighted means or unweighted means? (Ans. There is no such thing inside the cells. They are just means.)
2) Are the marginal means weighted means or unweighted means? (Ans. Doesn't matter. This is a balanced design. Marginal means are the same no matter how you calculate them.)
3) Are the SDs in the margins the pooled SDs? (Ans. Can't be. Most of them are out of the range of the SDs they'd be pooled from.)
4) Is the SD in the lower right corner the overall pooled SD? (Ans. Same answer as for the last question.)
5) Why is the overall SD in the lower right corner larger than any of the group SDs? (Ans. Because it contains variability due to differences in the group means.)

You might as well start filling in the ANOVA summary table.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructions (B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age x Instructions (AxB)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (Within)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
You already have the total SS (SST) because I've calculated it for you. But BE SURE you know how to calculate it for yourself. Next time I may not be this generous. Total df (dfT) should also be easy to get. The error terms should also be second nature to you by now.

SSW = 18.5 + 30.0 + 109.6 + 54.0 = 212.1 (why?)
dfW = N - k = 40 - 4 = 36
MSW = SSW / dfW = 212.1 / 36 = 5.892

That means SSBetw = SST - SSW = 663.775 - 212.1 = 451.675 (why?)
And SSBetw = SSA + SSB + SSAXB (explained variability).

You learned in the previous lecture that the main effects can be calculated as if you were doing a one-way ANOVA. Sadly, that is true only when the design is balanced. When the design is unbalanced, we have to resort to the harder "means method." Do you remember it?

SS = n per group (row or column) * sum(row or column means - GM)^2

Be careful now. Remember, the main effects are in the marginal means. So when we are calculating a main effect, our "groups" are entire rows or columns. Let's calculate the rows (Age) main effect first.

SSA = 20 * (10.65 - 9.825)^2 + 20 * (9.00 - 9.825)^2 = _____________
(Hint: You should have gotten 27.225.)
dfA = rows - 1 = 2 - 1 = 1
MSA = SSA / dfA = 27.225 / 1 = 27.225

Now let's do the columns (Instructions).

(Hint: You should have gotten 378.225.)
dfB = columns - 1 = 2 - 1 = 1
MSB = SSB / dfB = 378.225 / 1 = 378.225

And now the SSAXB is obtained by subtraction.

dfAXB = dfT - dfW - dfA - dfB = 39 - 36 - 1 - 1 = 1
MSAXB = SSAXB / dfAXB = 46.225 / 1 = 46.225

The F values are MSeffect / MSW. Calculate those and put them in the table. All are being evaluated on 1 and 36 degrees of freedom. The critical F is 4.113.

Draw appropriate conclusions about each of the effects. Which is the most interesting of these effects? What are the effect sizes?
Now it's time to face the music and do the unbalanced case. Here is the data summary after we lose 4 subjects.

<table>
<thead>
<tr>
<th></th>
<th>Counting</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Younger</td>
<td>sum = 51</td>
<td>sum = 148</td>
</tr>
<tr>
<td></td>
<td>sumsq = 335</td>
<td>sumsq = 2300</td>
</tr>
<tr>
<td></td>
<td>n = 8</td>
<td>n = 10</td>
</tr>
<tr>
<td></td>
<td>mean = 6.375</td>
<td>mean = 14.8</td>
</tr>
<tr>
<td></td>
<td>SS = 9.875</td>
<td>SS = 109.6</td>
</tr>
<tr>
<td></td>
<td>SD = 1.188</td>
<td>SD = 3.490</td>
</tr>
<tr>
<td>Older</td>
<td>sum = 70</td>
<td>sum = 92</td>
</tr>
<tr>
<td></td>
<td>sumsq = 520</td>
<td>sumsq = 1094</td>
</tr>
<tr>
<td></td>
<td>n = 10</td>
<td>n = 8</td>
</tr>
<tr>
<td></td>
<td>mean = 7.0</td>
<td>mean = 11.5</td>
</tr>
<tr>
<td></td>
<td>SS = 30.0</td>
<td>SS = 36.0</td>
</tr>
<tr>
<td></td>
<td>SD = 1.826</td>
<td>SD = 2.268</td>
</tr>
</tbody>
</table>

And here is the ANOVA summary table for you to fill in as you go.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (A)</td>
<td>___</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Instructions (B)</td>
<td>___</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Age x instructions (AxB)</td>
<td>___</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
<tr>
<td>Error</td>
<td>___</td>
<td>______</td>
<td>______</td>
<td>______</td>
</tr>
</tbody>
</table>

Notice I have not included a total line, because it is not so useful in this method. The treatment and error sums of squares are NOT going to add up to SST.

We can begin by getting the SSW. There is no change in how this is calculated.

SSW = 9.875 + 30.0 + 109.6 + 36.0 = 185.475 (why is it different?)

Then dfw and MSW, no change in the calculations here either. Just different answers because we have a different number of subjects.

dfW = N - k = 36 - 4 = 32
MSW = SSW / dfw = 185.475 / 32 = 5.796

Okay, you've done the easy part. The method we are using to calculate this ANOVA is called Type III. (If there's a Type III, how much do you want to bet there are also Type I and Type II?) Type III is
sometimes also called the method of unweighted means, although the two are not identical. (They are for 2x2 designs but not for larger designs.) The idea is, we are going to artificially balance this design by calculating an effective group size, and then we are going to use that effective group size in all cells to do an ANOVA. We calculate the effective group size from all groups.

\[
ne = \frac{4}{\left(\frac{1}{8} + \frac{1}{10} + \frac{1}{8} + \frac{1}{10}\right)} = 8.889 \text{ (explain: why these values?)}
\]

Now we effectively have:

<table>
<thead>
<tr>
<th>Counting</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Younger</td>
<td></td>
</tr>
<tr>
<td>sum = 51</td>
<td>sum = 148</td>
</tr>
<tr>
<td>n = 8</td>
<td>n = 10</td>
</tr>
<tr>
<td>mean = 6.375</td>
<td>mean = 14.8</td>
</tr>
<tr>
<td>ne = 8.889</td>
<td>ne = 8.889</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Older</td>
<td></td>
</tr>
<tr>
<td>sum = 70</td>
<td>sum = 92</td>
</tr>
<tr>
<td>n = 10</td>
<td>n = 8</td>
</tr>
<tr>
<td>mean = 7.0</td>
<td>mean = 11.5</td>
</tr>
<tr>
<td>ne = 8.889</td>
<td>ne = 8.889</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>mean = 6.6875</td>
<td>mean = 13.15</td>
</tr>
<tr>
<td>ne = 17.778</td>
<td>ne = 17.778</td>
</tr>
</tbody>
</table>

The way we have calculated the cell means has not changed. Cell means are cell means. Make sure you see how they were calculated. If you were to draw an interaction plot, you would draw it using those means. However, the calculation of the marginal means is different. Notice that the mean in the first row looks like it ought to be:

\[
\frac{(51 + 148)}{(8 + 10)} = 11.056
\]

That would be the weighted marginal mean, and if that's the one we wanted, this wouldn't be called the method of unweighted means, would it? The marginal means are not the weighted means but the unweighted means, and the "number of subjects" in the rows and columns is not what's actually there, but 2 * ne. I have shown these calculations in the rows. Make sure you can get them in the columns.

The grand mean is also unweighted. Add all the cell means and divide by the number of cells. To get ne for the entire table (not really necessary), multiple ne in the cells by the number of cells.

In general, the importance of a cell to an ANOVA result depends upon its size. Notice that by artificially balancing the design the way we have, we have increased the importance of some cells and decreased that of others. Which are which?

So we've got the unexplained variability (SSW). Now we have to bite the bullet and get the explained variability (SSBetw). Above we did that by subtraction, but that doesn't work anymore because the SSs no longer sum to the total. So we have to use the "means method." (By
the way, we could have done that in the balanced case, too. It was just easier to do it by subtraction.)

\[ \text{SSBetw} = ne \times \sum (\text{cell means} - \text{GMuw})^2 \]

(I like to put ne inside the sum rather than factoring it out. There is method to that madness, but you can hold off and multiply by ne after getting the sum if you want. You should get the same answer.)

\[ \text{SSBetw} = 8.889 \times (6.375 - 9.91875)^2 + \\
8.889 \times (7.0 - 9.91875)^2 + \\
8.889 \times (14.8 - 9.91875)^2 + \\
8.889 \times (11.5 - 9.91875)^2 = \______________________ \\
\]

(Hint: You should have gotten 421.3761.)

We need this number to get the interaction SS after we get the main effects. The main effects are also gotten by this "means method."

\[ \text{SSA} = 17.778 \times (10.5875 - 9.91875)^2 + \\
17.778 \times (9.25 - 9.91875)^2 = \______________________ \\
\]

Make sure you know where those numbers come from (row marginal ne, row marginal unweighted means, and grand unweighted mean).

(Hint: You should have gotten 15.9016, rounding a bit.)

No change in the rest of it. Only the calculation of the SS is different.

\[ \text{dfA} = \text{rows} - 1 = 2 - 1 = 1 \]
\[ \text{MSA} = \frac{\text{SSA}}{\text{dfA}} = \frac{15.9016}{1} = 15.9016 \]

\[ \text{SSB} = 17.778 \times (6.6875 - 9.91875)^2 + \\
17.778 \times (13.15 - 9.91875)^2 = \______________________ \\
\]

What are the only numbers that changed in this calculation compared to the calculation of SSA?

(Hint: You should have gotten 371.2394.)

\[ \text{dfB} = \text{columns} - 1 = 2 - 1 = 1 \]
\[ \text{MSB} = \frac{\text{SSB}}{\text{dfB}} = \frac{371.2394}{1} = 371.2394 \]

Finally, the interaction term is (I'll trust you to calculate it):

\[ \text{SSAxB} = \text{SSBetw} - \text{SSA} - \text{SSB} \]
\[ \text{dfAxB} = \text{dfT} - \text{dfW} - \text{dfA} - \text{dfB} = \text{dfA} \times \text{dfB} \]
\[ \text{MSAxB} = \frac{\text{SSAxB}}{\text{dfAxB}} \]

Now the F values are obtained the same way:

\[ \text{Feffect} = \frac{\text{MSeffect}}{\text{MSW}} \]
The critical value of F is different because it is based on fewer degrees of freedom in the error term: \( F_{\text{crit}} = 4.149 \).

Draw your conclusions. How did they change after losing 4 subjects? (By the way, losing subjects is called subject attribution or subject mortality. I'll refrain from making a morbid joke about that.)

When is this Type III method the right one to use? Some people claim it should be used any time there is an unbalanced factorial design. In fact, it is the default method in many of the large statistical packages such as SPSS and SAS. (It is not the default method in R. You need to know what your software is doing!)

Other people claim the Type III method is never right. Those people are a small minority, however. (One of the R big dogs is famous for having said, "Type III, just say no.")

The truth probably lies somewhere in between. If you have a true (designed, randomized) experiment, and it has become MILDLY unbalanced BY ACCIDENT AND AT RANDOM, then I believe the Type III method is the correct one. If the design is severely unbalanced, or if it is unbalanced because that reflects the state of nature, then the Type III method is incorrect.

Consider the following. We will discuss an experiment (study, not a true experiment) in which the researcher measured two IVs, hostility and propensity to anger (as measured by psychological tests), and one DV, propensity to physical aggression (also by a test). Her design was unbalanced, because hostile vs. not hostile is different numbers of people, and same for propensity to anger. Her design was unbalanced because nature is unbalanced.

However, something more important than that is going on here. We might consider hostility a trait. Some people are hostile, some are not. Anger, on the other hand, is a state. Everybody gets angry from time to time, but it comes and goes. Some people anger more easily than others, however.

Here's Fred. Fred is a very hostile person. How easy do you think it is to make Fred angry? Whereas, here's Joe. Joe is not a hostile person. Who's more likely to become angry in a given situation, Fred or Joe?

The Type III method outlined above is designed to remove the confound (relationship) between our IVs. In the case just outlined, however, the relationship between our two IVs is not a confound. It's an effect. We don't want it removed. We want to see it! The Type III method would not allow us to do that and is, therefore, clearly wrong here.

So good news! There's still more to learn!