Exercise 2: Many Labs 2 - Mixed Models Analysis

Henrik Singmann

January 2019 (updated: 2019-05-19)

Data

The file `ml2_alter_selected.csv` contains part of the data collected in the Many Labs 2 project (see: https://osf.io/8cd4r/) for replicating the fluency effect on accuracy rates for syllogisms. This effect was initially reported by Alter, Oppenheimer, Epley, and Eyre (2007).

The description of the study in the Many Labs 2 paper reads:

> Alter and colleagues (2007) investigated whether a deliberate, analytic processing style can be activated by incidental disfluency cues that suggest task difficulty. Forty-one participants attempted to solve syllogisms presented in either a hard- or easy-to-read font. The hard-to-read font served as an incidental induction of disfluency. Participants in the hard-to-read condition answered more moderately difficult syllogisms correctly (64%) than participants in the easy-to-read condition (42%; $t(39) = 2.01$, $p = 0.051$, $d = 0.64$, 95% CI [-0.004, 1.27]).

Syllogisms are logical arguments with two premises. For example:

> All dentists are golfers.
> Some dentists are not tennis players.
> What, if anything, follows

Participants have to select the correct answer (what follows with logical necessity assuming the premises listed above are true) among a set of 9 possible conclusions listed below:

1. All golfers are tennis players.
2. No tennis players are golfers.
3. All tennis players are golfers.
4. No golfers are tennis players.
5. Some golfers are not tennis players.
6. Some golfers are tennis players.
7. Some tennis players are not golfers.
8. Some tennis players are golfers.
9. Cannot reach a conclusion.

In this replication, each participant was asked to respond to 6 such syllogisms.

- The current data set contains data from three different samples that took part in Many Labs 2, as recorded in variable `source`. One group of participants was collected in `tilburgcaf` (N = 79), one group in `purkyne` (N = 140; university in the czech republic), and one group in `uniporto` (N = 35).
- `uID` contains a unique identifier for each participant.
- `syllogism` contains an identifier for each of the six syllogisms.
- `fluency` contains the between-subjects manipulation with two levels: `easy` to read font (i.e., black Myriad Web 12-point) and `difficult` to read font (10% grey italicized Myriad Web 10-point).
- `correct` contains the information whether or not the syllogism was solved correctly (= 1) or incorrectly (= 0).

Please note that the selection of samples is not random here. The samples were selected to show a specific pattern. Overall Many Labs 2 did not replicate the original effect. They report:

> [Overall,] participants in the hard-to-read condition answered a similar number of syllogisms correct (M = 1.10, SD = 0.88) as participants in the easy-to-read condition (M = 1.13, SD = 0.91; $t(2,578) = -0.79$, $p = 0.43$, $d = -0.03$, 95% CI [- 0.11, 0.05]). [...] These results do not support

the hypothesis that syllogism performance would be higher when the font is harder to read versus easier to read; the difference was slightly in the opposite direction and not distinguishable from zero ($d = -0.03$, 95% CI [-0.11, 0.05] versus original $d = 0.64$).

Also note, the correct answer for the syllogism shown above should be 5.

Tasks

```
## preparation and loading
library("tidyverse")
library("afex")
library("emmeans")
emm_options(lmer.df = "Satterthwaite")
theme_set(theme_bw(base_size = 15))
d <- read_csv("ml2_alter_selected.csv")
d <- d %>%
  mutate(
    source = factor(source),
    uID = factor(uID),
    syllogism = factor(syllogism),
    fluency = factor(fluency, levels = c("easy", "difficult"))
  )
str(d)
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1518 obs. of  5 variables:
##  $ source   : Factor w/ 3 levels "purkyne","tilburgcaf",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ uID      : Factor w/ 254 levels "1276","1277",..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ syllogism: Factor w/ 6 levels "3","4","5","6",..: 1 2 3 4 5 6 1 2 3 4 ...
##  $ fluency  : Factor w/ 2 levels "easy","difficult": 2 2 2 2 2 2 1 1 1 1 ...
##  $ correct  : num  0 0 1 0 1 0 0 0 0 0 ...
```

1. Analyse the data with a mixed model using `afex::mixed`. Use accuracy (i.e., `correct`) as dependent variable and `source` and `fluency` as independent variables (plus their interaction). Use crossed random-effects for participants (`uID`) and items (`syllogism`). For any mixed model analysis, you should start with the *maximal model justified by the design*. So your first goal is to figure out what the maximal model justified by the design is. Fit this then. Feel free to use `method = "S"` for faster fitting.
2. The maximal model should show some convergence problems (i.e., "singular fit" and/or a correlation among random components of 1.00). Reduce the random-effects structure of the model until you find a model that does not show any convergence problems.
3. Maximal and reduced models should all show a significant interaction between `source` and `fluency`. Use follow-up tests (e.g., using `emmeans`) to investigate this pattern. Does any of the samples show the pattern predicted by Alter et al. (2007)? When in doubt use the model without identifiability issues for that.
4. Make a plot of the `fluency` by `source` interaction using `afex_plot`. Spend some time thinking about how to make this plot really nice. Is the default for argument `id` really a good idea here? What is with the default argument of `error`? Remember, `afex_plot` has a vignette showing many examples: https://cran.r-project.org/package=afex/vignettes/afex_plot_introduction.html
5. The item random effect has only six levels. This is way too low for any real life situation and is only done here as an example. There are two main arguments against this.
    1. This can lead to numerical estimation problems and the absolute lower limit is somewhere around 5 or 6 level. Which is just where we are. The lme4 faq (http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html) states about that:

Treating factors with small numbers of levels as random will in the best case lead to very small and/or imprecise estimates of random effects; in the worst case it will lead to various numerical difficulties such as lack of convergence, zero variance estimates, etc..

2. Small numbers of levels for one of the random-effects grouping factor provide an upper bound for the totally attainable statistical power. This means that even if we were to increase the number of levels for one random-effects grouping factor (e.g., participants) low-numbers of levels in the other (e.g., items) will hinder our chances to find an effect, even if it is present. This is discussed extensively in Westfall, Kenny, and Judd (2014) which provide the following rather shocking example, which has even more levels than the present case:

   For example, in an experiment employing the stimuli-within-condition design [i.e., crossed-random effects with a between-item condition], [...] where the true effect size is large at $d = 0.8$, and where there are a total of eight stimuli (four stimuli per condition)-a sample size which we suspect many experimenters would consider perfectly adequate for a stimulus sample-the maximum attainable power is only about .41. However, if we just double the sample size of stimuli to a still relatively modest 16 (eight per condition), then the maximum power to detect a large effect goes up to about .78.

   Consequently, we need to use a different strategy here. Specifically, estimate a mixed-model with only random-effects for participants, but include `syllogism` as a fixed-effect that fully interacts with the other fixed effects. Again, think about what is the maximal model here. Can you estimate it? What do the results from this model tell us regarding our research question. Are there relevant effects of `syllogism`?

6. If we only have one random-effect, we might also use a simple ANOVA to analyse the data. Do so using the `afex` ANOVA functions. Does it show the same results as the mixed-model of question 4? Is there a benefit of using the mixed model over the ANOVA?

7. Which model do you think is the best model making the best argument? Why? Make a plot based on this model (using `afex_plot`) that you would put in a paper of yours (can be the same plot as above of course).

References

- Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136(4), 569-576. https://doi.org/10.1037/0096-3445.136.4.569
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143(5), 2020-2045. https://doi.org/10.1037/xge0000014
- Aldrovandi, S., Brown, G. D. A., & Wood, A. M. (2015). Social norms and rank-based nudging: Changing willingness to pay for healthy food. *Journal of Experimental Psychology: Applied*, 21(3), 242-254. https://doi.org/10.1037/xap0000048