**Session 1:** “**Means can be pretty mean”**

(**Important note:** The study described below is an example constructed purely for teaching purposes; any similarity between this hypothetical example and published research is entirely coincidental in nature, and unintended on my part. While the basic research question in the example is inspired by a previous study on morphological processing in native speakers of Japanese (Nakano, Ikemoto, Jacob, & Clahsen; 2016) the results reported below are entirely made up. Also, the methodological and statistical problems which the example is designed to illustrate did not occur in the previous study. In fact, the study did not include any L2 speakers at all, and was also based on a more complex experimental design, which included a number of additional control conditions)

A group of researchers investigate the processing of morphologically complex Japanese words written in Hiragana (a moraic script typically used in Japanese dictionaries, as an alternative to Kanji script) in L1 and L2 speakers of Japanese. They recruit a sample of 35 native speakers and 35 highly proficient L2 speakers of Japanese, and conduct a morphological priming experiment in which they measure lexical decision times for 28 target words, which were either preceded by morphologically a related prime word or by an unrelated control word. In accordance with the established standards for analyses in this field, they exclude data points for which the lexical decision was incorrect, because this indicates that the participant did not know the respective words. Also, to avoid a possible detrimental influence of a small number of extremely high or low values, all lexical decision times lower than 250 ms and higher than 1300 ms were considered outliers and excluded from the analyses; these exact cut-off values are in line with the values used in previous morphological priming studies.

Mean lexical decision times by condition and participant group are shown in Table 1 below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | L1 group (N=35) |  | L2 group (N=35) |
| Morphologically related | 696(94) |  | 972(188) |
|
| Unrelated control | 731(102) |  | 985(201) |
|
| **Table 1.** Mean lexical-decision times in ms (and standard deviations in brackets) by condition for L1 and L2 speakers of Japanese. | | | |

While the L1 group showed significantly faster lexical decision times in the morphologically related condition than in the unrelated condition; t1(34)=3.12; ***p<.05***; t2(27)=3.04; ***p<.05***; the L2 group did not show a difference between the two conditions; t1(18)=1.12; p>.1; t2(25)=1.49; p>.1. The authors therefore conclude that L1 and L2 speakers process morphologically complex words differently, with L1 speakers relying mainly on morphological decomposition, while L2 speakers rely mainly on whole-word-storage and retrieval in the mental lexicon.

(Please note: As any L2 study, the above experiment also contains a number of conceptual issues which would be worth discussing, such as what the L1 of the L2 speakers was, in what way they acquired Japanese, for how long they have been learning the L2, etc. For the sake of your group discussion, however, please focus on the methodological and statistical issues.)

1. Discuss within your group what exactly the above-mentioned results from the inferential tests tell you.
2. Discuss to what extent the above example is related to the issue of data trimming and means being vulnerable with regard to the influence of outliers.
3. Can you detect any fundamental flaws in the analyses?
4. Do you think the fact that the L1 group shows a significant effect, while the L2 group does not, justify the conclusion suggested by the authors?

**Bonus question**: In this study, the authors decided to conduct inferential analyses with by-participants (t1) and by-items (t2) paired-samples t-tests. Based on what you already know about linear-mixed effects models, do you think it would be possible to use these models as an alternative? How would such an analysis look like? What advantages would such an analysis have compared with the t-tests reported above?

**Extra-spicy bonus question for real stats geeks**: Do you think such an alternative analysis based on linear-mixed effects models would also help to avoid the problem of a small number of extreme outliers having a detrimental influence on the results, or does this problem also occur in this alternative analysis?