1pSCa3. Experimentally elicited productions: Differences and similarities between mixed effects and ANOVA analyses

Matthew Goldrick*

*Corresponding author's address: Linguistics, Northwestern University, 2016 Sheridan Rd., Evanston, IL 60208, mattgoldrick@northwestern.edu

Currently, many experimental studies of speech production use fully counterbalanced designs to examine variation in categorical (e.g., correct/incorrect) or relatively continuous measures (e.g., reaction times, voice onset times). These data present several challenges to ANOVA analyses. Some of these issues are well known to be the speech community; for example, the non-normality of dependent variables such as proportion correct. Others have been less extensively addressed; for example, many speech studies account for participant- but not item-specific contributions to variance. I discuss the opportunities and challenges in using linear mixed effects models to address these issues.
The prototypical behavioral experiment has a fully balanced design, with an equal number of observations at each level of the factors being manipulated. For example, suppose we run a study of speech perception in noise, examining the proportion of keywords correctly identified by naive listeners. We manipulate two variables: speaker background (native, non-native) and noise level (±2dB). A balanced design would use equal numbers of native and non-native speakers, and provide listeners with equal numbers of tokens from each at each noise level. Analysis of Variance (ANOVA) has long been the gold standard for statistical analysis of such experiments. Below, I briefly review some of the well-known issues of this method for fully balanced designs. I then review how linear mixed effects models address some of these concerns.

**ISSUES WITH ANOVA FOR BALANCED DESIGNS**

**Non-Normality of Dependent Measures**

ANOVA assumes that dependent measures are normally distributed. Speech researchers have long been familiar with the fact that this assumption is false for proportions (e.g., the proportion of keywords correctly identified by listeners). To address this, many speech researchers perform ANOVAs not on raw proportions but on transformed measures (e.g., rationalized arcsine units; Studebaker, 1985). Unfortunately, such transforms fail to address the full range of issues with proportional data, particularly for observations that are near the floor or ceiling of proportions (Jaeger, 2008).

In addition to proportional dependent measures, the issue of non-normality also arises with continuous dependent measures. Reaction times are typically skewed (Baayen & Milin, 2010) as are many phonetic measures (Crystal & House, 1981). Such observations are frequently not transformed before being subject to ANOVA analysis.

**Controlling for Idiosyncratic Participant- and Item-Based Contributions to Variance**

A typical ANOVA analysis allows us to take into account the idiosyncratic contributions of different participants (treating such contributions as a random effect). For example, ANOVA could confirm in the experiment above that intelligibility differences between native and non-native speakers is robust across typical variation in intelligibility across speakers. Less frequently addressed in speech research is the fact that items can also make idiosyncratic contributions. For example, some keywords may be easier to perceive in noise; this may skew our results, particularly if some keywords can only appear in one condition but not another. (n.b. Differences between listeners could provide a third source of random variation.) In experimental psychology, this issue has been addressed by utilizing multiple ANOVA analyses that examine whether effects are robust across sources of idiosyncratic variation (subject and item analyses). Results from these can then be integrated to provide a single evaluation of the null hypothesis (e.g., min-$F'$; Clark, 1973).

**LINEAR MIXED-EFFECTS MODELS AND BALANCED DESIGNS**

Linear mixed-effects models are a growing part of the statistical analysis toolbox of studies of linguistic behavior (Baayen, Davidson, & Bates, 2008; Jaeger, 2008; see Baayen, 2008, for a textbook introduction). How does this technique address these issues with ANOVA?

**Analysis of Non-Normal Dependent Measures**

Like ANOVA, linear mixed-effects models assume dependent measures are normally distributed. They do not, in and of themselves, offer a solution to this issue. However, most studies making use of this analytic method represent proportional data using logits, the natural log odds of a response (i.e., logistic regression; see Agresti, 2013, for a textbook introduction). Utilizing this dependent measure addresses issues arising from the non-normality of proportional data. For example, unlike proportions, logits are unbounded (ranging from negative to positive infinity) and allow us to properly model the dependence between variances and means that holds for proportional data (see Jaeger, 2008, for detailed discussion). With respect to other non-normal dependent measures, linear mixed-effects models are similar to ANOVA in requiring the use of appropriate transformations. For example, the non-normality of positively skewed reaction times can be reduced through the use of logarithmic transformations (see Baayen & Milin, 2010, for detailed discussion).
Simultaneous Control of Idiosyncratic Participant- and Item-Based Contributions to Variance

Unlike traditional ANOVA techniques, linear mixed-effects models allow for simultaneous treatment of multiple crossed random effects (e.g., when all talkers produce all keywords). Instead of building separate ANOVA models to capture each effect, a single linear mixed-effects model can simultaneously capture the contributions of all random effects (e.g., including talkers, keywords, and listeners). Based on analyses of simulated datasets, this has been argued to provide more accurate estimates of the factors underlying participant performance (Baayen et al., 2008). An open issue in the field is how to best treat the contribution of random effects. Recent work analyzing simulated datasets has argued that the use of ill-specified linear mixed-effects model structures leads to an increased rate of false rejections of null hypotheses (i.e., increase in Type I error) relative to min-F ANOVA analyses (Barr, Levy, Scheepers, & Tily, 2013).

DISCUSSION

Prototypical experimental designs in speech research aim to be fully balanced. While this partially satisfies the assumption of ANOVAs, inherent properties of many dependent measures and experimental designs violate other assumptions of this analysis technique. Linear mixed-effects models and related regression techniques (properly applied) provide some relief from these issues.

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation Grant BCS0846147. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

REFERENCES