Navigating the Statistical Tides: An R Tutorial for the Non-Coding-Inclined

As empirical, quantitative methods continue to become more commonplace in linguistic research, linguists are increasingly employing various kinds of experimental tasks, including grammaticality judgment tasks, word/phrase-list readings, perceptual discrimination tasks, matched guise techniques, sociolinguistic interviews, surveys/questionnaires, pre-/post-tests, and corpus analyses. In order to be able to analyze the distinct kinds of data obtained from these tasks, minimally, a passive knowledge of inferential statistics and a degree of proficiency with inferential statistics software are required.

Two principal complications present themselves with respect to these requirements: (1) the set of statistical techniques deemed appropriate for specific types of data is inherently non-static, evolving as advances are made in the statistics discipline, effectively resulting in 'waves' of consensus in the Linguistics community regarding the use of particular statistical tests for linguistic data, and; (2) multiple software options exist and continue to be developed for performing inferential statistics, each with unique interfaces and capabilities that evolve alongside the very tests they were designed to perform. Thus, a linguist's endeavor to gain expertise in experimental methodologies, statistical theory, and statistical software packages is much more akin to chasing a perpetually moving target than mastering a finite set of skills.

One software package in particular, R ([3]), has arguably become the new norm for statistical analysis in the Linguistics community, boasting a free and maximally powerful open-source platform that stands in stark contrast to other competitors (e.g. SAS [4], SPSS [1], STATA [6],) that are only accessible through paid (and often University) subscription. Unfortunately, however, R's minimal user interface begs programming language, constituting a daunting and perhaps unintuitive burden for many linguists. Still, as efforts to combat this reliance on user-generated code remain somewhat restricted to Variationist Sociolinguistics (e.g. Rbrul [2], Language Variation Suite [5]), it seems increasingly likely that the newer generations of empirical linguists will be tasked with becoming proficient in R.

In this workshop, my goal is to all but eliminate R's barrier to entry, affording fuller access to statistical analyses to as wide a community of linguists as possible. Reducing user-generated coding to the absolute minimum, namely the typing out of names of independent and dependent variables of interest, I showcase a highly intuitive and step-by-step Copy-Paste guide that enables users to fully perform and interpret R analyses with linguistic data. Example datasets that illustrate each of the aforementioned types of experimental data will serve as the basis for guided tutorials of R inferential analysis, facilitating a maximum amount of 'hands-on' engagement. I opt for a maximally practical and pedagogical approach to R, restricting the set of possible statistical techniques with which to practice and gain expertise to the following prominent few:

- ANOVA / Linear Regression (with and without random effects)
- Logistic Regression / Poisson Regression (with and without random effects)

Beyond acquiring the skillset to confidently perform and interpret the results of the above statistical techniques on their own data as appropriate, attendees will gain a significant familiarity with the benefits and limitations of the linguistic community's adoption of R.

References

[1] IBM Corp. (2019). IBM SPSS Statistics for Windows/Macintosh. Armonk, NY: IBM Corp. <https://www.ibm.com/analytics/spss-statistics-software>.

[2] Johnson, Daniel Ezra (2009). Getting off the Goldvarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3(1): 359-383.

[3] R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

[4] SAS Institute Inc. (2019). SAS software. <https://www.sas.com/en_us/software/platform.html>.

[5] Scrivner, Olga & Manuel Díaz Campos (2016). Language Variation Suite: A theoretical and methodological contribution for social and linguistic data analysis. *Linguistic Society of America*. <https://languagevariationsuite.shinyapps.io/Pages/>.

[6] StataCorp (2017). Stata Statistical Software: Release 15. College Station, TX: StataCorp LLC. <https://www.stata.com/>.