Fundamentals and applications of resampling methods for the analysis of speech production and perception data.

Crouzet, Olivier
Laboratoire de Linguistique de Nantes – LLING / UMR6310
Université de Nantes / CNRS, Chemin de la Censive du Tertre, 44312 Nantes Cedex
olivier.crouzet@univ-nantes.fr

Fundamentals: The ambition of any statistical data analysis procedure is to reach a reasonable description of the reference population from which the observations have been randomly sampled. Guaranteeing such a satisfactory description may rely on the estimation of measurement variation in the target population (e.g. through the computation of Confidence Intervals) as well as on formulating a decision concerning the theoretical / operational hypotheses (inferential tests of hypotheses) among others. Both procedures require an experimenter to estimate properties of the population, either to approximate how much the measurement is expected to vary had one selected a different random sample from this very same population, or to provide a decision as to whether one may reject the Null Hypothesis that there's no difference in the underlying population between the conditions that are being compared.

A fair amount of statistical methods (e.g. Student t-tests, ANOVA, GLM, LME, GLMM, GAM...; Faraway, 2006; Knoblauch & Maloney, 2012) rely on various types of what are called "asymptotic" results: one has to assume that the set of observations was sampled from a population whose properties (family, distribution parameters) are expected to reach a stable and mathematically defined state for sufficiently large samples (e.g. a Gaussian distribution with parameters μ –mean– and σ^2 –variance– to name only one). Transforming the data adequately may constitute an intermediate step. In any case, defining the referent population family and parameters is a strong requirement for choosing the right tool to (1) estimate how much variation may occur concerning the measurements that have been made; or (2) position the observed sample with respect to the hypothesized distribution under a Null hypothesis. Though recent approaches to statistical analysis let one model data from non-Gaussian distributions (e.g. GLM, GAM) it is still essential to identify and estimate the underlying distribution adequately.

A complementary approach to the estimation of distributional form and parameters for an underlying population is to "construct the population out of the observed sample". This general approach has been coined "resampling" and, though it was described as early as in the 1930s, it is only in the late 1980s that the increased availability of powerful personal computers could lay the basis for its practical development through various procedures (bootstrap, permutations, Monte-Carlo simulations...). Though these methods have spread in some introductory statistical textbooks (Fox, 2015) and have even given rise to entire books (e.g. Good, 2005a,b; Robert & Casella, 2010), it seems to me that our community would often take advantage at being introduced to such methods as they provide e.g. adequate treatment of small samples, of unknown distribution families and parameters for larger samples, of power analyses...

Applications: As a speech scientist working with both production and perception data, I offer to provide 3 illustrations of such approaches to the analysis of typical speech and language data: **1 —** a simulated sample from a non-Gaussian distribution (cf. Fig. 1), **2 —** data from a speech production experiment in which statistical modelling of locus equations (cf. Lindblom & Sussman, 2012) is sought (cf. Fig. 2) and **3 —** simulated data from a speech perception experiment in which the proportion of nominal classification responses would need to be modelled. *The first example* will let me discuss fundamental differences between the asymptotic and resampling approaches with a rather simple case. *The second one* will let me introduce how resampling methods can help both estimating population variation (here with linear regression estimates) and applying tests of hypotheses[1]. *The latest example* will provide insights into how one can apply these methods to categorical variables such as TRUE / FALSE responses.

All three examples will be described with base `R` (R Core Team, 2012) programming scripts in order to let the audience conceptualize what is happening in the process but examples with the dedicated `boot` (Canty & Ripley, 2016) `R` package will also be provided. All data and scripts will be made available to the participants so they can get a hands-on experience from this presentation after the workshop.

[1] Locus Equations –LEs– rely on relatively large amounts of data to be estimated and one often ends up with a small sample of LE parameters from a large sample of recordings.

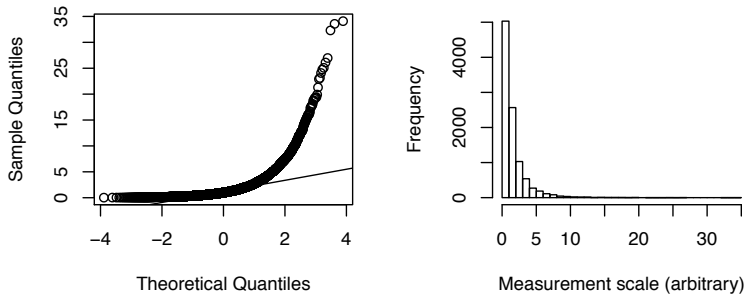

Figure 1: An illustration of a (strongly) non-Gaussian distribution from which data may be randomly sampled. The QQ-plot on the left shows strong departure from the Gaussian assumption. This distribution will be used as an example for the computation of Confidence Intervals in both the asymptotic and the resampling framework.

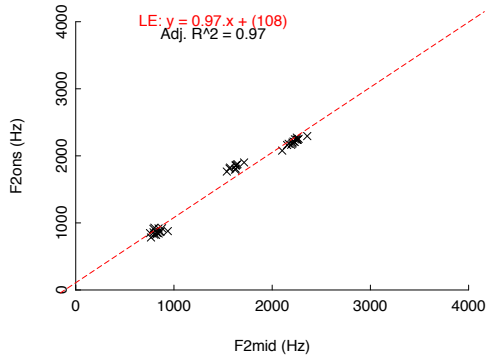

Figure 2: An illustration of the estimation of a single Locus Equation (slope and intercept) from a relatively large amount of data giving rise to a strong reduction in sample size when performing statistical analyses of LE parameters. Displayed data correspond to the production of arabic /k/ in {i, a, u} contexts by a single speaker. The presentation will show how resampling methods can be implemented to estimate (1) a confidence interval for LE slope and / or intercept; (2) whether LE estimates significantly differ between two or more experimental conditions.

References

Canty, A., & Ripley, B. D. (2016). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-18.

Faraway, J. (2006). *Extending Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. FL: Boca Raton, USA: Chapman and Hall / CRC.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, Inc, 3rd edn.

Good, P. (2005a). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer Series in Statistics, New-York, USA: Springer-Verlag Inc., 3rd edn.

Good, P. I. (2005b). *Resampling Methods: A Practical Guide to Data Analysis*. Birkhäuser, 3rd edn.

Knoblauch, K., & Maloney, L. T. (2012). *Modeling Psychophysical Data in R*. UseR!, New-York, USA: Springer-Verlag.

Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, *40*, 1–19.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, ISBN 3-900051-07-0.

Robert, C., & Casella, G. (2010). *Introducing Monte Carlo Methods with R*. UseR!, New-York, USA: Springer-Verlag.