Journal of Phonetics xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Special Issue: Emerging Data Analysis in Phonetic Sciences

Mixed-effects design analysis for experimental phonetics

James Kirby [a,*], Morgan Sonderegger [b,1]

[a] School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh EH8 9AD, Scotland, UK
[b] Department of Linguistics, McGill University, 1085 Dr Penfield Avenue, Montreal, Quebec H3A 1A7, Canada

ARTICLE INFO

Article history:
Received 27 September 2017
Received in revised form 19 May 2018
Accepted 30 May 2018
Available online xxxx

Keywords:
Power
Effect size
Design analysis
Incomplete neutralization

ABSTRACT

It is common practice in the statistical analysis of phonetic data to draw conclusions on the basis of statistical significance. While *p*-values reflect the probability of incorrectly concluding a null effect is real, they do not provide information about other types of error that are also important for interpreting statistical results. In this paper, we focus on three measures related to these errors. The first, *power*, reflects the likelihood of detecting an effect that in fact exists. The second and third, *Type M* and *Type S* errors, measure the extent to which estimates of the magnitude and direction of an effect are inaccurate. We then provide an example of *design analysis* (Gelman & Carlin, 2014), using data from an experimental study on German incomplete neutralization, to illustrate how power, magnitude, and sign errors vary with sample and effect size. This case study shows how the informativity of research findings can vary substantially in ways that are not always, or even usually, apparent on the basis of a *p*-value alone. We conclude by repeating three recommendations for good statistical practice in phonetics from best practices widely recommended for the social and behavioral sciences: report all results; design studies which will produce high-precision estimates; and conduct direct replications of previous findings.

1. Introduction

Statistical analysis is often used to reason about scientific questions based on a data sample, with the goal of determining "which parameter values are supported by the data and which are not" (Hoenig & Heisey, 2001, p. 4). Researchers in phonetics frequently reach such conclusions based on *significance*: the probability, or *p*-value, of obtaining an effect of the observed size (or greater), if the true effect were zero.

For example, consider a study of the effect of speech rate on Voice Onset Time (VOT) on short-lag stops in a particular language (e.g. Kessinger & Blumstein, 1997). The researcher fits a statistical model (say, a simple linear regression) in which the dependent variable is VOT, and the regression coefficient of interest β_1 is the slope of the regression line, representing an estimate of how a unit change in speech rate impacts VOT. A *t*-test is then conducted to assess whether this slope is different from zero. Judging from the literature, many researchers would conclude that there is an effect of rate if this difference is significant (i.e. if $p < 0.05$), and that if the difference is not significant ($p \geqslant 0.05$), VOT is unaffected by rate.

This focus on the *p*-value stems from a desire to avoid incorrectly rejecting the null hypothesis, when it is in fact true. This is obviously to be avoided, because we do not want to claim that an effect exists when it does not. However, *p*-values provide only limited information when interpreting studies, particularly if we are trying to interpret a study in relation to other work. To continue with the speech rate example, imagine two studies of the effect of speech rate on VOT, one of which finds a significant effect ($p \leqslant 0.05$) and one of which does not ($p > 0.05$). Given only the *p*-values, we are not in a position to assess which result is more plausible, since the *p*-value itself does not measure the probability that speech rate has a non-null effect on VOT. Moreover, the difference between the *p*-values may not itself be statistically significant (Gelman & Stern, 2006; Nieuwenhuis, Forstmann, & Wagenmakers, 2011), so we cannot even conclude that there is a meaningful difference between the two studies.

In addition to interpreting significant effects, researchers are often interested in interpreting the *lack* of a significant

* Corresponding author. Fax: +44 (0)131 651 3190.
E-mail addresses: j.kirby@ed.ac.uk (J. Kirby), morgan.sonderegger@mcgill.ca (M. Sonderegger).
[1] Fax: +1 514 398 7088.

https://doi.org/10.1016/j.wocn.2018.05.005
0095-4470/

Please cite this article in press as: Kirby, J., & Sonderegger, M. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics* (2018), https://doi.org/10.1016/j.wocn.2018.05.005

effect, a so-called "null result". The temptation is often to conclude that if a coefficient is not significantly different from zero, that it does not have an effect on the dependent variable. Concluding from non-significance that there is no effect of an experimental manipulation is a well-known statistical fallacy; the *p*-value is *not* the probability that the null hypothesis is true, but rather the probability of observing an effect of a given magnitude, or larger, assuming that the null hypothesis *is* true. In order to avoid this pitfall, it is sometimes taught, or propagated in practice, that null results cannot be interpreted at all. However, this is not strictly speaking the case: null results can sometimes give information about likely parameter values or *effect size*—arguably the central goal of data analysis—but determining whether or not this is the case requires considering information other than the *p*-value of a test statistic.

In this paper, we discuss additional quantities that can give useful and complementary information to *p*-values: the probability of rejecting the null hypothesis assuming that it is false (statistical *power*) as well as errors of magnitude and sign in estimating effect size (*Type M* and *Type S* errors: Gelman & Tuerlinckx, 2000; Gelman & Carlin, 2014). Using simulation studies based on real experimental data, we illustrate three reasons researchers in phonetics should take into account power and effect size in addition to significance:

(1) Depending on statistical power, a non-significant result can still be informative.
(2) Errors in estimates of effect size can be substantial even when *p*-values are low.
(3) Estimates of effect size improve with power, and can be robust even when *p*-values are higher than a conventional threshold, e.g. $\alpha = 0.05$.

Using a case study of so-called *incomplete neutralization* (hereafter IN), we illustrate how (1)–(3) can affect conclusions drawn with respect to two questions, which are arguably always our goal in interpreting research studies: *what can we conclude about likely values of a parameter from a single study (Q1), as well as from a body of studies (Q2)?*

This exercise provides an example of *design analysis* (or *design calculations*; Gelman & Carlin, 2014): the use of statistical tools to reason about likely outcomes (= parameter values) of replications of a study—which is generally of greater interest than the statistical analysis of a single experiment.[2] Our focus here will be on design analysis for mixed-effects regression models, because these methods have become increasingly common for phonetic data analysis, and also because they can be somewhat more technically and conceptually challenging to implement. However, we note that the basic points (1)–(3) apply to most statistical methods commonly used to analyze phonetic data, including *t*-tests, classical ANOVA, classical regressions (without random-effect terms), and GAMMs.

[2] For example, in a study of whether there is a speech rate effect on VOT for lenis stops in English, we are less interested in whether the coefficient for this effect is significantly negative ($p < 0.05$) than in what can be concluded about the *true* value of the speech rate effect. By points (2) and (3), these are not the same thing.

None of the points we raise about power and effect size are novel (see e.g. Brysbaert & Stevens, 2018; Button et al., 2013; Cohen, 1988; Colquhoun, 2014; Gelman & Carlin, 2014; Gigerenzer, Krauss, & Vitouch, 2004; Meehl, 1967; Nieuwenhuis et al., 2011; Westfall, Kenny, & Judd, 2014; Vasishth & Nicenboim, 2016; Judd, Westfall, & Kenny, 2017; Vasishth & Gelman, 2017, among others), but they are not typically addressed in interpretation of phonetic data. We believe that greater attention to these dimensions would improve the quality of phonetic research, both in terms of research design as well as interpretation. We hope the technical illustration provided in this paper will be of particular use to those researchers who are interested in performing power calculations and design analysis in the mixed-model context, but are unsure how to go about doing so.

The remainder of this paper is organized as follows. Section 2 provides some background on power, effect size, and sign and magnitude errors, including the practical issue of how to compute them. Section 3 gives a case study of incomplete neutralization in word-final German stops, focusing on points (1)–(3), in the context of interpreting individual studies (Q1) and a body of studies (Q2), using power and effect size considerations in addition to significance. Finally, in Section 4 we conclude with some more general observations and recommendations.

To facilitate the use of power and effect size error calculations in phonetic research, code and data files for carrying out all analyses in this paper, as well as further worked examples, are archived as an Open Science Foundation project (Kirby & Sonderegger, 2018a).

2. Background

In this section, we define power and effect size before turning to considerations of power calculation, magnitude and sign errors, and design analysis. While there exist large literatures on each of these topics—in particular for the behavioral and social sciences—they are not usually discussed as part of mainstream statistical analysis of phonetic data. (For psycholinguistic data on the other hand, Vasishth & Nicenboim (2016) cover similar topics, and our presentation is indebted to theirs.) We aim here to briefly summarize relevant concepts for our case study, and give relevant references where interested readers can follow up to learn more. Our case study (Section 3) provides a worked example showing one way these concepts can be applied to the analysis of phonetic data.

2.1. Power

In considering whether there is in reality an effect of a covariate or experimental manipulation, there are two essential types of errors a researcher can make: falsely concluding there is an effect when none exists (a *Type I error*, or "false positive"), or falsely concluding there is no effect when one in fact exists (a *Type II error*, or "false negative"). Type I errors are arguably more familiar, and everyday statistical practice places considerable emphasis on avoiding them. If a term is found to be statistically significant, many researchers would conclude from this that a Type I error is unlikely. The Type I error rate of a

Please cite this article in press as: Kirby, J., & Sonderegger, M. Mixed-effects design analysis for experimental phonetics. *Journal of Phonetics* (2018), https://doi.org/10.1016/j.wocn.2018.05.005