PSEUDOREPLICATION IN PHONETIC RESEARCH

Bodo Winter

Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Germany;
Department of Cognitive and Information Sciences, University of California, Merced, USA
bodo@bodowinter.com

ABSTRACT

This methodological paper attempts to bring the problem of pseudoreplication to the attention of the phonetic community. Pseudoreplication refers to the treatment of dependent observations as independent data points, which causes an overabundance of erroneously significant results. The relevance of this problem is demonstrated by analyses of phonetic data, and it is shown that the problem occurs frequently in the phonetic literature. Finally, simple solutions to combat pseudoreplication in the design and analysis of phonetic experiments are proposed.

Keywords: phonetic methodology, inferential statistics, pseudoreplication

1. INTRODUCTION

Pseudoreplication occurs when multiple samples from one experimental subject or one experimental stimulus are treated as independent data points in statistical analyses [5]. This problem results from the application of inferential statistics to observations that are interconnected or correlated with each other. The prevalence of pseudoreplication has been noted in ecology [5], psychology [4], neuroscience [7] and many other disciplines. Milinski mentions pseudoreplication as one of the most frequent ‘deadly sins in the study of behavior’ [9].

Phonetic research has specific methodological pitfalls that make it very easy for pseudoreplication to occur. It is a well-known fact that speech is inherently variable and that an utterance can never be spoken exactly the same way twice. This variability is the reason why many repetitions are incorporated into the designs of phonetic experiments: through many repetitions we can derive a better estimate of what the speaker “usually says” and reduce the effects of random variation. However, problems arise when inferential statistics such as t-tests and ANOVAs are used as if the repetitions were from different individuals.

The reason why repetitions cannot be treated as independent are twofold: First, there is random influence on a response that is likely to be similar across all repetitions. Everybody has slightly different, idiosyncratic ways of pronouncing words and sentences. From the perspective of an experiment, these idiosyncrasies are “random” because they cannot be controlled for, but importantly, they will be present in all of a speaker’s responses, across multiple items and multiple repetitions of items.

The second way in which repetitions are non-independent results from articulatory reduction due to repetition itself. It has been demonstrated multiple times that segmental durations tend to be shortened in repetitions e.g. [1, 10]. This is a systematic way in which repetitions are *inter*dependent. In what way do these interdependencies affect the analysis of phonetic data?

2. DEMONSTRATING THE EFFECTS OF PSEUDOREPLICATION

In the ecological literature (starting with [5]), the possible effects of pseudoreplication have been demonstrated many times. This section of the paper provides two examples of these effects that are specific to phonetics and which highlight why pseudoreplication has to be avoided.

2.1. Simulated data: a simple t-test example

In this example, the effects of pseudoreplication are demonstrated with t-tests; however, it should be pointed out that the problem is not particular to this test but also applies to more complex ones such as ANOVAs and linear mixed effects models.

Suppose we were interested in whether words of a language are longer in a focus position than in a non-focus position, and suppose (for the sake of simplicity) that we were only able to do recordings with a single speaker. Because of this we cannot use inferential statistics to make an inference on the population of speakers; however, we can make an inference on the population of words and sentences in the speaker's language if we sample many different words and sentences. We therefore perform an items analysis in which each "row" of the dataset is a single item.

In this hypothetical experiment, eight different words are put in different focus or non-focus sentence frames and paired t-tests are used to assess whether there is a significant difference in measured vowel durations (see Table 1).

Table 1: An example of a simulated dataset of vowel durations for which a chance result is obtained: the Focus condition is significantly longer than the No Focus condition: t(7)=2.6104, p=0.035

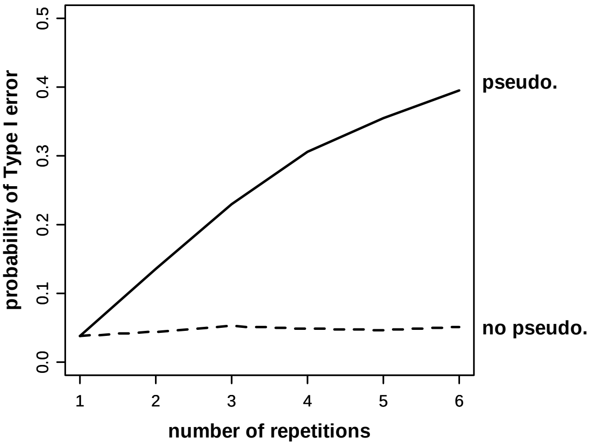
	Focus	No Focus
Word 1	173 ms	175 ms
Word 2	160 ms	137 ms
Word 3	136 ms	143 ms
Word 4	181 ms	155 ms
Word 5	152 ms	136 ms
Word 6	192 ms	161 ms
Word 7	170 ms	170 ms
Word 8	177 ms	162 ms

1000 datasets were simulated in which the vowel durations in both conditions (Focus vs. No Focus) were sampled from *the same normal distribution* with a mean of 150ms and a standard deviation of 20ms. Paired t-tests were performed on each of these datasets, and in 38 cases these indicated a significant difference even though the items in the Focus and No Focus condition stem from the same underlying distribution. The reason for these erroneously significant results (Type I errors) is that random sampling can sometimes lead to "clumping", and more high or low values end up in one of the conditions by chance alone. Such a chance result is exemplified by the simulated dataset in Table 1.

To assess the effect of repetitions on the performance of the t-test analyses, the dataset was made larger by adding repetitions, each of which was sampled from a normal distribution that was centered around the preceding value of the same word. For example, Word 1 in Table 1 has a vowel duration of 173ms and repetition 2 of Word 1 was then sampled from a normal distribution with the mean 173 and the standard deviation 20. This captures the fact that there is variation from one repetition to another, but each repetition is somewhat influenced by the preceding item of the same type. The standard deviation value of 20ms was based on the standard deviations between repetitions of actual vowel data (language: German, 11 speakers, courtesy of Heriberto Avelino). t-tests were computed *with* pseudoreplication and *without* pseudoreplication (by averaging over repetitions).

As can be seen from Fig. 1, the probability of finding a significant result even though the items were sampled from the same distribution (Type I error) drastically increases with a larger number of repetitions. With six repetitions, a significant result ($p < 0.05$) was obtained 396 times – although we *know* that there was, in fact, no difference. This means that the α-level (Type I error rate) is 0.396, thus greatly exceeding the standard level of 0.05.

Figure 1: Probability of obtaining a Type I error with a paired t-test plotted on number of repetitions (cf. [5]).

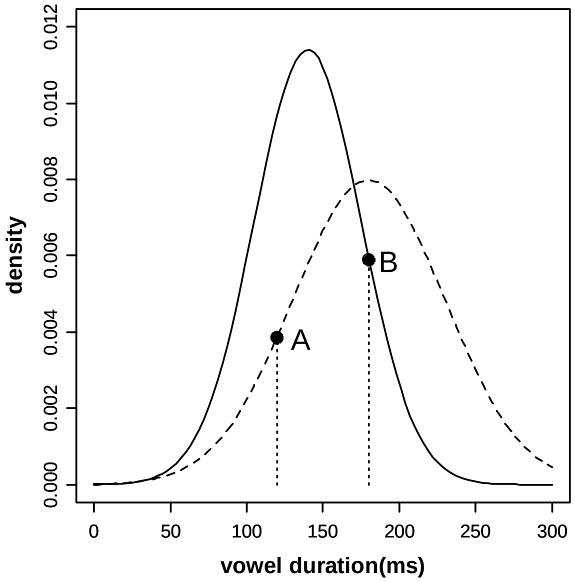


The reason for this increase in Type I error rate is simple: just by randomness alone, several large or several small values can sometimes clump together in one of the conditions – this can be expected to occur in any kind of sampling. However, if the unevenly distributed items are repeated multiple times, chance is allowed to have a larger influence on the result and spurious effects arise. On the other hand, by averaging over repetitions, pseudoreplication can be avoided and the α-level stays around the accepted value of 0.05.

2.2. Another problem with pseudoreplication

Fig. 2 depicts two hypothetical distributions for long and short vowels, one with a mean of 140ms, and one with a mean of 180ms. The x-axis depicts different duration values in milliseconds, and each value in this hypothetical example comes from a different word. There is a large amount of overlap between the distributions of the short vowel (left) and the long vowel (right), possibly reflecting a near-merger situation or an ongoing merger.

Figure 2: Two hypothetical distributions of vowel durations with sample points A and B



Let us assume we were to sample the word A from the distribution of long vowels and the word B from distribution of short vowels (Fig. 2). With many repeated samples of A and B, and by treating these samples as independent, a researcher can obtain a statistically significant result showing *that long vowels are shorter than short vowels* (see [6] p. 601. for a similar example).

The likelihood of obtaining misleading results like in this example *increases with diminishing item numbers and greater numbers of repetitions*. For example, even with three words sampled from each distribution, the researcher might be "unlucky" and select three words closer to A and three words closer to B. With many repetitions and pseudoreplication, these few words can affect the outcome of statistical analyses in a disproportionate way.

3. A BRIEF LITERATURE SURVEY

The preceding section illustrated that pseudoreplication does affect phonetic analyses. This naturally leads to the question as to how frequent pseudoreplication occurs in phonetic studies and therefore, whether this is a problem we have to worry about in phonetic research.

In the following survey of the Speech Production session of the ICPhS 2007, 36 experimental papers that used inferential statistics will be evaluated. Elevated degrees of freedom (*df*) are used as the main indicator for whether pseudoreplication appeared in a study or not (cf. [8] p. 419). For example, an analysis on 4 subjects who uttered 6 items in 2 conditions (fixed effect) can have degrees of freedom of maximum 3 for a subjects analysis (number of subjects minus number of fixed effects) and 5 for an items analysis (number of items minus number of fixed effects). Degrees of freedom such as 23 (4 subjects times 6 items minus number of fixed effects) indicate pseudoreplication in which subjects and items are conflated.

Only 26 of the experimental studies reported *df*. Of these, 16 exhibited elevated degrees of freedom (~62%), indicating pseudoreplication. 10 experimental studies did not report *df* and although these studies did not permit any certain conclusion as to whether pseudoreplication occurred, the likelihood of pseudoreplication was quite high because of very low item and subject numbers, and very high repetition numbers (some studies had up to 10 repetitions). However, it should be pointed out that not reporting *df* is a problem in itself, because it does not allow readers to assess whether statistical tests were used correctly.

Given that 60% of the studies that enabled pseudoreplication to be definitely assessed actually exhibited this phenomenon, and given that many more were very likely to exhibit it, the problem does indeed seem to be quite prevalent in phonetic research. Some studies had pseudoreplication on multiple levels (e.g. simultaneously conflating subjects, items and repetitions), thus artificially inflating sample sizes to large extents (some degrees of freedom were above 1000).

A survey to see whether pseudoreplication occurs as frequently in journal publications as in conference papers is currently in the making.

4. REMEDIES TO PSEUDOREPLICATION

Luckily, there are many ways in which pseudoreplication can be avoided and for many of the studies in the literature survey mentioned above, a simple re-analysis of the data would suffice.

To avoid pseudoreplication of items, separate subjects- and items-based analyses have to be performed [3]. For the subjects-based analysis, each subject contributes a single row in a dataset and averaging is done over items, for the items-based analysis, each item contributes a single row and averaging is done over subjects. In order to be able to conduct meaningful items analyses, some of the surveyed studies would have to increase the number of independent (unique) words or sentences presented to subjects.

An even better analysis approach without pseudoreplication would use linear mixed effects models (for a discussion of the advantages of this model type and a general introduction, (see [2]). However, mixed models only address the problem of pseudoreplication if the random and fixed effects are chosen so that all dependencies in a dataset are accounted for. For example, a repeated measures design with multiple repetitions needs at least two random effects for subjects and items and a fixed effect for repetitions.

With respect to pseudoreplication of repetitions, there are at least three possible remedies: First, repetitions can be dropped from experimental designs altogether. If the item number is large enough, an items analysis is given the chance to reach significance and random influences that affect individual responses are leveled out across different items. It should be kept in mind that other disciplines such as psycholinguistics also have to cope with a lot of variation (e.g. in reaction times), and in these disciplines precise estimates of a response are derived via large numbers of (unique) items rather than large numbers of repetitions.

Second, one can take the means across different repetitions – this needs to be done both in a subjects and in an items analysis. While this approach is valid and has the benefit of simplicity, it has the disadvantage that the information as to how much each repetition varies from any other is thrown out. It then cannot be assessed whether the phenomenon of interest changes across repetitions, for example, whether a phonetic difference between members of a contrasting pair is diminished or enhanced after several repetitions.

Third, "Repetitions" can be treated as a fixed effect in a mixed model – this is the preferred solution in this paper. This approach allows controlling for interactions between repetitions and the phenomenon of interest. Then, additional research questions such as "Does a difference between tense and lax stops get smaller with multiple repetitions?" can be answered in a quantitative way.

5. CONCLUSIONS

Pseudoreplication and the resulting increase of Type I errors render some of the interpretations we make on phonetic data less certain than they could be. Lombardi and Hurlbert ([8] p. 420) state that avoiding pseudoreplication is not simply a matter of statistical 'refinement'; it is a crucial aim of all statistical analyses.

As pointed out in section 4, pseudoreplication is easy to avoid; and by avoiding pseudoreplication in future phonetic studies, we can be more confident of our results. It is hoped that this paper brings the topic of "pseudoreplication" to the attention of the phonetic community, and that it encourages a renewed interest in experimental design and statistical analyses.

6. ACKNOWLEDGEMENTS

I thank Roger Mundry, Daniel Voigt, Sven Grawunder, Timo Röttger, Leonardo Lancia and Clive Winter for helpful comments and suggestions. I thank Heriberto Avelino for making the vowel duration data available to me. I am solely responsible for any remaining errors.

7. REFERENCES

[1] Aylett, M., Turk, A. 2004. The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language & Speech* 47, 31-56.

[2] Baayen, R.H., Davidson, D.J., Bates, D.M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59, 390-412.

[3] Clark, H.H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning & Verbal Behavior* 12, 335-359.

[4] Freeberg, T.M., Lucas, J.R. 2009. Pseudoreplication is (still) a problem. *Journal of Comparative Psychology* 123(4), 450-451.

[5] Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54(2), 187-211.

[6] Kroodsma, D. 1989. Suggested experimental designs for song playbacks. *Animal Behavior* 37, 600-609.

[7] Lazic, S.E. 2010. The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neuroscience* 11(5), 1-17.

[8] Lombardi, C.M., Hurlbert, S.H. 1996. Sunfish cognition and pseudoreplication. *Animal Behavior* 52, 419-422.

[9] Milinski, M. 1997. How to avoid seven deadly sins in the study of behavior. *Advances in the Study of Behavior* 26, 159-180.

[10] Pluymaekers, M., Ernestus, M., Baayen, H.R. 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62, 146-159.