A common misapplication of statistical inference: nuisance parameter control with null-hypothesis rejection tests

Jona Sassenhagen; Phillip Alday

February 2016

Abstract

Much experimental research on behavior and cognition rests on stimulus or subject selection where not all parameters can be fully controlled, even when attempting strict matching. For example, when contrasting patients to controls, factors such as intelligence or socioeconomic status are often correlated with patient status; when presenting word stimuli, factors such as word frequency are often correlated with primary variables of interest. One procedure very commonly employed to control for such nuisance parameter effects is conducting inferential tests on confounding stimulus or subject characteristics. For example, if word length is not significantly different for two stimulus sets, they are considered as matched for word length. Such a test has extremely high failure rates and is conceptually misguided. We show this procedure to be misguided both pragmatically and philosophically, present a survey showing its high prevalence, and briefly discuss an alternative in the form of regression including nuisance parameters.

Introduction

A common problem in brain and behavioral research, where the experimenter cannot freely determine every stimulus and participant parameter, is the control of confounding/nuisance parameters. This is especially common in studies of language. Typically, stimuli cannot be constructed out of whole cloth, but have to be chosen from existing words (which differ in many parameters); stimuli are processed by subjects in the context of a rich vocabulary; and subject populations have usually been exposed to very diverse environments and events in their acquisition of language. The basic question researchers are faced with is

then to prevent reporting e.g. an effect of word length, or bilingualism, when the effect truly stems from differences in word frequency, or socioeconomic status, which may be highly correlated with the parameters of interest. A very prevalent method we find in the literature highlights common statistical misconceptions and fails to perform the necessary control.

NHST and nuisance control

Often, researchers will attempt to demonstrate that stimuli are selected so as to concentrate their differences on the parameter of interest, i.e. reduce confounds, by conducting null-hypothesis testing such as t-tests or ANOVA on the potentially confounding parameter in addition or even instead of showing descriptive statistics in the form of measures of location and scale. The underlying intuition is that these tests establish if two conditions differ in a given parameter and serve as proof that the conditions are "equal" on that parameter.

In practice, we find insignificant tests are used as a necessary (and sometimes even sufficient) condition for accepting a stimulus set as "controlled". This approach fails dramatically on multiple levels.

- Philosophically, these tests are inferential tests being performed on closed populations and not random samples of larger populations. Statistical testing attempts to make inferences about the larger population based on randomly selected samples, but here the "samples" are not taken randomly and we are not interested in a population. For example, in a study on the effects of animacy in language processing, we do not care whether the class of animate nouns in the language is on average more frequent than the class of animate nouns. Instead, we care whether the selection of animate nouns *in our stimuli* are on average more frequent than the selection of inanimate nouns *in our stimuli*. But inferential tests answer the former question, not the latter.

- Pragmatically, this procedure does not test a hypothesis of interest. This procedure tests the null hypothesis of "the populations that these stimuli were sampled from do not differ in this feature", but what we are actually interested in is "the differences in this feature between conditions is not responsible for any observed effects". In other words (assuming that it is valid to perform tests on closed, constructed populations), this procedure tests whether the conditions differ in a certain respect, but not whether that difference actually has any meaningful influence on the result.

- Additionally, these tests carry all the usual problems of Null Hypothesis Significance Testing (cf. **REF**), including its inability to accept the null hypothesis directly. This means that even if the conditions do not differ significantly, we cannot accept the hypothesis that they do not differ; we

can only say that there is not evidence to exclude this hypothesis (which, again, is not the one we are actually interested in).

In other words, these tests are incapable of actually informing us about influence of potential confounds, but may however give researchers a false sense of security. A simple solution is to examine the descriptive measures of location and scale (e.g. mean and variance[1]) and see if the stimuli groups are "similar enough". For perceptual experiments, there may even be established discrimination thresholds below which the differences are considered indistinguishable. The preferred solution is directly examining whether these potential confounds have an influence on the results. This is accomplished by including these nuisance parameters in the statistical model and is readily implemented with increasingly popular multilevel regression models (Gelman and Hill 2006; Fox 2016), whether Bayesian or frequentist.

Randomization checks in clinical research

In the context of baseline differences between treatment and control groups in clinical trials, a similar debate has been waged (e.g. Senn 1994). The procedure is called "randomization check" as it refers to checking if assignment of subjects to treatments has truly been performed randomly. This is philosophically somewhat less misguided, but has also been determined to be pragmatically pointless. In truly experimental research such as clinical trials, the effect of treatment is the variable of interest, and true randomization can be performed with regards to the multitude of other factors that might influence results. But in the case of non-medical, quasi-experimental research (i.e. research where full control is not possible and thus confounds are unavoidable), stimuli or subjects are typically *known* to not have been selected randomly, but by specific criteria (e.g., animate vs. inanimate nouns, or patients with vs. without a particular lesion[2]). That is, in the case of medical studies with randomization checks,

[1] We note that the correct variance calculation here would be the true population variance (with n in the denominator) and not the usual Bessel's corrected estimate (with $n-1$ in the denominator) because we are not estimating the variance of a population from a sample, but rather calculating the variance of a closed, fully sampled population. Practically, as long as we are consistent between groups, it makes no difference. (The mean does not have this problem as the usual, maximum-likelihood estimator for the population mean is simply the sample mean, i.e. we use the same formula both for estimating the population mean from a sample and for computing the population mean.)

[2] One could pose the question of whether or not the patients in clinical trials have been selected randomly – after all, a patient must have a particular condition in order to take part in a particular trial. The key factor here is that the population of interest in clinical trials are "patients with condition X" and thus we often do not care if condition X correlates with condition Y because that correlation only makes sense in the context of a larger population that we are not interested in. When comparing between subgroups of a larger population, e.g. patients with or without a lesion in the larger population of speakers of a given language, we do care if condition X correlates with condition Y (e.g. having a lesion correlates with age) because these form systematic differences between subgroups within a population.

experimental validity is achieved by selecting subjects from a given population and randomizing their assignment to treatment. In our quasi-experimental case, stimuli are constructed so as to differ on one parameter which we highly expect to be correlated with other parameters, e.g. word frequency and word length, and the worry of researchers is not if assignment was random (in fact, it is known to not have been random), but if stimuli differ systematically on variables expected to impact the dependent variable of interest. Notably, randomization checks have been repeatedly shown to fail at identifying covariates that should be included into multiple regression models (Imai, King, and Stuart 2008). Nonetheless, the clinical trial literature provides important considerations for experimental design choices e.g. by discussing the proper way of blocking and matching (Imai, King, and Stuart 2008).

We assume that these issues in quasi-experimental research has been discussed some of the better statistical or ecological textbooks, but are not aware of similar discussions in the psychological, linguistic or neurocognitive literature.

Prevalence

We performed a literature survey of neurolinguistic studies to estimate the prevalence of inferential tests of nuisance parameters.

Qualitative impressions

Instances of the error can be easily found not only in recent, but also in older publications, such as this example from the 1980s:

> the two prime categories were equivalent in text frequency ([. . .] et al., 1971), and in length (both t's < 1.1

Here, the authors demonstrate in one sentence many of the fallacies underlying this procedure: both estimating a known quantity, and deducing equivalence (acceptance of the null) from a failed test (in this case, a test that leads the authors to accept the null hypothesis). To estimate how common the problem is in neurolinguistics, a high-quality neurolinguistic journal, *Brain & Language*, was investigated.

Quantitative prevalence of the problem in recent issues of *Brain & Language*

In total, 86 articles where found where researchers reported known quantities in their stimulus/materials section, and 58 (**67%**) of these reported inferential

statistics of these known values. Of these, 47 (**81%**) "accepted" the null hypothesis (i.e., implicitly assumed that stimuli or subjects were matched following a nonsignificant test). We conclude that in a large fraction of those cases where researchers published in *B&L* are concerned about nuisance parameters of experimental stimuli, they conduct meaningless tests and misinterpret the results of these tests in a potentially dangerous manner.

Discussion and recommendation

In sum, NHST control of nuisance parameters is prevalent and meaningless, based on a flawed application of statistics to an irrelevant hypothesis. Luckily, proper nuisance control (of known and measurable variables) is not complex, although it can require more effort and computer time.

Researchers should still use descriptive statistics to demonstrate the success of balancing, but beyond that p values from statistical tests on the stimulus properties offer no reliable, objective guideline. To directly and objectively estimate the influence of a set of stimuli on the dependent variables of interest, researchers should include stimulus properties in their statistical model for the data. For traditional t-tests, ANOVAs and regression models, this corresponds to using multiple regression with the stimuli properties as additional nuisance parameters. In multiple regression, all parameters are jointly estimated and the total variance is allocated over all parameters depending on their independent impact. Thus, a condition effect estimated by a model also containing nuisance parameters corresponds to the effect of condition after having controlled for nuisance parameter influence. (An additional advantage of regression models is that is possible to include continuous covariates thus avoiding issues with dichotomization. **do we need a ref here? I have one**) Importantly, to prevent p value fishing, the choice of selecting covariates to include must be made on principled grounds, and either a priori, or via unbiased model selection procedures.

Hierarchical/multilevel modeling (a.k.a mixed-effects modeling; see also Pinheiro and Bates 2000; Gelman and Hill 2006; Fox 2016) provides the necessary extension to the regression procedure for repeated-measures designs. Multilevel regression models have the additional advantage over regression of accounting for the combined variance of subjects and items in one model, which can significantly impact the pattern of observed effects (Clark 1973; Baayen, Davidson, and Bates 2008; Judd, Westfall, and Kenny 2012).

One problem in this context is that these stimulus confounds can be assumed to be correlated not only with one another and the dependent variables, but often also with the independent variables of interest (e.g., word frequency and word length correlate). This leads to model collinearity – the problem that models become hard to estimate and have biased variance estimates due to strongly correlated parameters. As such, we recommend that researchers check

the correlation of model parameters; popular software for mixed-effects models such as lme4 automatically provides a summary of correlation between fixed effects (Bates et al. 2015).

The main technique for dealing with collinearity is one that researchers traditionally already employ: attempting to balance stimulus/subject selection so that differences in nuisance parameters are minimised, e.g. via matching or blocking. That is, matching should generally still be performed in addition to multivariate estimation. However, often, they can not be entirely abolished, and significant correlations remain. Even in these cases, multiple regression is still the preferred solution if the present collinearity is dealt with appropriately.

Two solutions that go beyond the scope of this article are regularization and residualization. Regularization (whether frequentist or Bayesian) can be thought of building the additional assumption into the model that most parameter estimates should be small (close to zero) unless there is strong evidence otherwise. As such, estimates are shrunk towards zero and variables shrunk sufficiently close to zero can be thought of as not contributing towards the explanatory power of the model. In the case of highly correlated variables, one them will be shrunk towards zero as it contributes little towards the model beyond the other one. Residualization refers to regressing two correlated variables against each other and using the residuals from this regression to replace one of the variables. As the residuals are by definition the variance not explained by the correlation, the new variable is not correlated with the remaining original variable.

Calculating such complex regression models will of course require more data, as power is lost with each additional parameter being estimated. We view this as a good thing because studies in the brain and behavioral sciences are chronically underpowered (Button et al. 2013).

Thus, our recommendations for the control of nuisance parameters are:

- attempt to control stimulus parameters to a reasonable degree
- use descriptive, but not inferential statistics to guide stimulus selection
- add confounding parameters as covariates into the final data analysis process
- use high-powered samples

Each step in this list is (hopefully) uncontroversial and helpful, unlike null-hypothesis testing of stimulus balance.

Acknowledgements

We thank Sarah Tune for helpful discussion and Tal Linzen for bringing to our attention the randomization check literature. This work was supported in part by the German Research Foundation (BO 2471/3-2) and by the ERC grant (....).

References

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. "Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items." *Journal of Memory and Language* 59: 390–412.

Bates, Douglas, Martin Maechler, Benjamin M. Bolker, and Steven Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *arXiv*: 1406.5823.

Button, Katherine S, John P A Ioannidis, Claire Mokrysz, Brian A Nosek, Jonathan Flint, Emma S J Robinson, and Marcus R Munafò. 2013. "Power Failure: why Small Sample Size Undermines the Reliability of Neuroscience." *Nat Rev Neurosci* (Apr).

Clark, Herbert H. 1973. "The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research." *Journal of Verbal Learning and Verbal Behavior* 12: 335–359. doi:10.1016/S0022-5371(73)80014-3.

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. 3rd ed. Thousand Oaks, CA: Sage.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart. 2008. "Misunderstandings Between Experimentalists and Observationalists About Causal Inference." *Journal of the Royal Statistical Society A* 171, Part 2: 481–502.

Judd, Charles M., Jacob Westfall, and David A. Kenny. 2012. "Treating Stimuli as a Random Factor in Social Psychology: a New and Comprehensive Solution to a Pervasive but Largely Ignored Problem." *Journal of Personality and Social Psychology* 103 (1): 54–69. doi:10.1037/a0028347.

Pinheiro, José, and Douglas Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer New York.

Senn, Stephen. 1994. "Testing for Baseline Balance in Clinical Trials." *Statistics in Medicine* 13: 1715–1726.

Methods

Survey

The analysis was restricted to current volumes. For all articles published by *B&L* from 2011 to the 3rd issue of 2013, three raters (not blinded to the purpose of the experiment) investigated all published experimental papers (excluding reviews, simulation studies, editorials etc.). For each experiment reported in a study, the stimulus/materials sections were investigated for descriptive and inferential

statistics derived from populations that were exhaustively sampled without error. If a descriptive and/or inferential statistic (such as mean and standard deviation) were reported, the study was coded as one where the researchers were interested in a known quantity, otherwise it was discarded. If an inferential statistic (such as a p-value) was reported, the study was coded as one where researchers answered that interest with an erroneous parameter estimate, otherwise as one where researchers did not commit the error. If a statement of the form that groups were thought equivalent regarding the parameter was made, such as claims that they were "matched", "equal" or "did not differ", and this statement was backed up by a p-value greater than .05, the study was coded as "accepting the null". In cases of rater disagreement, the majority vote was registered. Representative statements from every study committing an error are presented in the appendix.