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

Jona Sassenhagen Phillip Alday(?)

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/age is not significantly different for two stimulus sets/subject groups, they are considered as matched for word length/age. Such a test has extremely high failure rates and is conceptually extremely misguided. We discuss the pragmatic and philosophical futility of this procedure, present a survey showing its high prevalence, conduct a simulation study showing the high expected failure rates, and briefly discuss an alternative in the form of regression including nuisance parameters.

Introduction

A common problem in quasi-experimental research, where not all parameters of stimuli or subjects can be freely chosen by the experimenter, is the control of confounding/nuisance parameters. This is especially common in studies of language. Typically, stimuli cannot be constructed ad hoc, 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 is worse than useless and highlights common statistical misconceptions.

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 by conducting some of the tests typically also employed for the testing of research hypotheses: t-tests or ANOVA; in addition or even instead of showing e.g. descriptive statistics in the form of measures of central tendency and spread. We assume the underlying intuition is that these tests establish if two samples differ on a parameter, or 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.

- Pragmatically, tests are conservative and rejection of stimulus sets based on the (in)significance of tests both fails to detect problematic confounds, and rejects stimulus sets as confounded where proper statistical protocols could have detected a true effect.
- Philosophically, tests are employed to test hypotheses that are neither the hypotheses researchers actually care about, nor in any way related to the problem at hand.
- The preferable solution is easy and readily implemented with increasingly popular hierarchical regression models from linear mixed or Bayesian techniques.

A fundamental pragmatic problem stems from the set-up of null hypothesis significance tests (NHST). Such tests can only ever reject hypotheses, and the parameters with which they are typically performed (e.g., alpha, the probability of rejecting true hypotheses, is set to 5%) entail conservative tests, that is, tests which fail to reject many false hypotheses so as to not falsely reject correct hypotheses. Conversely, if alpha is set to such a value, beta (the probability of failing to reject false hypotheses) practically becomes low (because it is the one statistical parameter, out of effect size, sample size, alpha and beta, that is left to float) for linguistics, psychology and neuroscience (e.g. Button et al., 2013); for typical effect and sample sizes, it is rarely above 50%, meaning that truly false nulls are rejected at a rate no better than if researchers were to flip a coin. Thus, nothing is gained from not rejecting a hypothesis in this context - not rejecting the hypothesis that the difference in word length is zero is of very little relevance as the chance of detecting this difference, even where it exists, is low in typical samples, and strongly depends on sample size.

If a researcher wishes to compare two sets of stimuli on a parameter, such as word length, the sample size results from the number of words (and further depends on e.g. pairing), and alpha is typically set to .05. If a researcher wishes to be able to detect large deviations from equal word length for both groups with for example a *t*-test, the power of the test - that is, the chance of the test to detect a real group difference - depends both on what "large" means, and the sample size. Large samples (i.e., a stimulus set of 2000 words) easily reject the hypothesis even when the difference is small, but small samples (i.e. 20 pairs of supposedly well-controlled words) reject even small differences only rarely. A difference of exactly the same magnitude is detected much more readily in a large than in a small sample - and for small stimulus sets, this detection rate is extremely poor.

Furthermore, the hypothesis tested in this process is completely irrelevant. p values in NHST reflect the probability of observing evidence as or more extreme as the one given, assuming a given true effect. This "true effect" is a statement about a population, evidence a statement about the sample. For example, in an fMRI study with n = 20, observing an effect of 3 standard deviations in magnitude in the sample can reasonably give researchers some confidence in the assumption that within the assumed, but unobserved population, there is also some non-zero effect of the same sign. Thus, p values in NHST allow researchers to infer from known quantities - the sample values - to unknown quantities - the population values. However, this framework is incompatible with the aims of controlling stimulus parameters, where it is precisely the sample, not the population, that is of interest.

We assume when a researcher performs a *t*-test on the word length of his stimuli, he hopes to answer a question such as "is there a *meaningful difference* in the length of the stimuli?". However, the question actually investigated is much more subtle, and mostly orthogonal to the question at hand: "given the known difference in samples, is there no difference in the population from which was sampled?". For the context of word stimuli, this can be rephrased as "is there a bias in the procedure that selected the members of the stimulus classes so that it somehow selects words of different lengths?". While detecting a biased mechanism for stimulus selection or generation can be interesting, it is inconsequential for the analysis of an already selected or generated set of stimuli; the rare occurrence of a balanced sample

resulting from a biased sampler would be unproblematic, whereas an unbalanced sample resulting from an unbiased sampler would be problematic.

This inference of the population value is performed based on effect (and sample) size, and effect size size is close to the value of interest - "meaningful difference". However, the *true* difference is already known, and p values say nothing about to which degree this difference within the sample is meaningful - only to which degree it is generalizable to the (uninteresting) population. For the purpose of controlling confounding, no information in the p value goes beyond the values it is based on - typically, one of central tendency and one of spread. Thus, the inferential (population-focused) procedure adds nothing to the descriptive (sample-focused) procedure. It may however give researchers a false sense of security.

Randomization checks in truly experimental work

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, stimuli or subjects are typically *known* to not have been selected randomly, but by specific criteria (i.e., animate vs. inanimate words, or patients with vs. without lesions). That is, in the case of the discussion of randomization checks, experimental validity is achieved by selecting subjects (typically non-randomly) 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 are balanced 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 & Stuart 2008). However, the clinical trial literature can support experimental design choices e.g. by discussing the proper way of blocking and matching (ibid).

We assume the issue in the context of quasi-experimental researcher has also been noted and probably even included in some of the better textbooks, but are not aware of similar discussions in the psychological, linguistic or neurocognitive literature.

Prevalence and impact

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

Survey: Prevalence of the problem in B&L

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 commit the "double sin" of 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.

Survey Methods

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.

Survey Results

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.

Simulation study

We furthermore conducted a simulation study to estimate the impact of this procedure on 1. correctly detecting results where a confound in the form of influence from imbalanced stimulus selection led to a result that was significant where no significant result would have been detected without this confound, 2. incorrectly rejecting results where a confound may exist, but a true effect a. was presented and b. would have been detected regardless of the confound, 3. failing to reject results where no effect would have been detected without stimulus confounds, but one is detected with the confounding factor. Code for this simulation can be found in the appendix, but the structure of the simulation study was as follows:

- Generate results
  - generate random data for two conditions for a number of subjects
  - add a certain "real" effect to one condition
  - conduct a t-test
  - add a certain "confound" effect to one condition
  - conduct another t-test on the condition data
  - conduct a third t-test on the confound effect itself

- Evaluate results
  - if the first test rejects, the sample at hand is understood as having a real effect, without confound present.
  - if the second test rejects, the sample at hand is understood as having an effect with confound present.
  - if the third test rejects, the confound would have been detected by the procedure we criticise here.

Results of simulation study

We observe major failure rates of the described procedure, both in the form of undetected false/confounded results, and in rejecting results that would have been significant without the confound effect (see Figure 1). For example, for an effect size of .5 and a confound effect size of .3 and 20 subjects/group (in a within-subject design), ~14.5% of runs resulted in a failure to reject a result that was significant with the confound, but was not significant without; ~10.5% resulted in rejecting a result that would have been significant also without the confound; and only ~12.5% resulted in the correct rejection of results that would not have been significant without the confound effect.

We observe that the primary determinant of this procedure is the ratio of the confound effect compared to the real effect. For example, if the real effect is 0 and the confound effect is .5, over 25% correct rejections stand against 4% false rejections and 4% missed rejections.

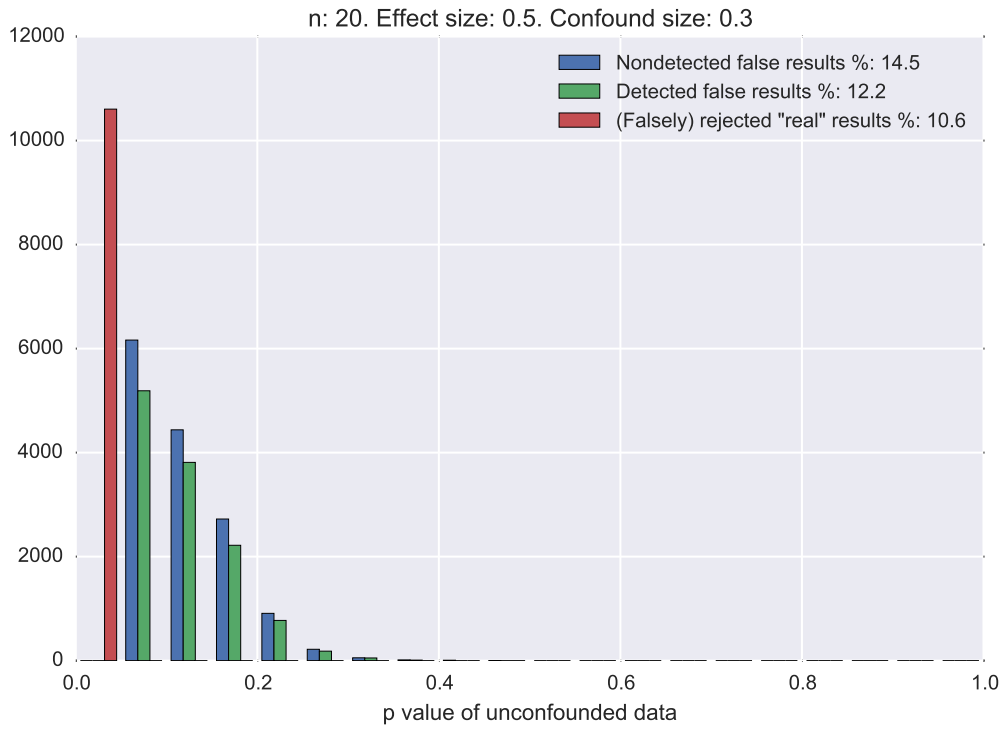

Figure 1: Simulation study results

Discussion and recommendation

In sum, NHST control of nuisance parameters is prevalent and meaningless, and can cloud true or induce false effects. Luckily, proper nuisance control (of known and measurable variables) is not complex, although it an be effortful or costly.

Researchers can still use descriptive statistics to estimate the meaningfulness of the difference in sample parameters, but p values offer no reliable, objective guideline. However, one procedure can objectively estimate the influence of a set of stimuli on the dependent variables of interest; multiple regression. 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. 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.

In a traditional ANOVA or t-test context, adding scalar covariates such as word length is impossible, but linear mixed effects (Bates et al., 2015) or Bayesian hierarchical modeling can be used to add stimulus parameters to the analysis.

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 due to strongly correlated parameters. We recommend to plot crosscorrelation matrices for all known parameters of interest to inform intuitions about dependence structures.

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 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 provides an unbiased estimate of the true effect and is unproblematic if the present collinearity is dealt with appropriately.

Common techniques for dealing with collinearity known from e.g. machine learning are not straightforwardly transferable. Regularization or its Bayesian equivalent are hard to interpret when the goal of analysis is a conservative analysis of the impact of a factor that might influence the observed variables, when the collinear parameters are usually strongly assumed to also influence these variables. The only clear-cut and fully unobjectionable way of dealing with this problem of variance inflation is collecting more data, as problems from collinearity depend both on the degree of correlation as well as on the available evidence.

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
- use high-powered samples
- add confounding parameters as covariates into the final data analysis process

Each step in this list is uncontroversial and helpful, unlike NHST of nuisance parameters.

Acknowledgements

We thank Sarah Tune for helpful discussion and Tal Linzen for bringing to our attention the randomization check literature.

References

Bates D, Maechler M, Bolker BM and Walker S (2015). *Fitting Linear Mixed-Effects Models using lme4.* ArXiv e-print; in press, Journal of Statistical Software, http://arxiv.org/abs/1406.5823.

Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013). *Power failure: why small sample size undermines the reliability of neuroscience.* Nat Rev Neurosci. 14(5):365-76. doi: 10.1038/nrn3475. Epub 2013 Apr 10.

Imai, K, King, G, Stuart, EA (2008). *Misunderstandings between experimentalists and observationalists about causal inference.* J. R. Statist. Soc. A 171, Part 2, pp. 481–502

Senn, S (1994). *Testing for baseline balance in clinical trials.* Statistics in medicine, Vol. 13.