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

Jona Sassenhagen Phillip M. Alday

May 2016

Abstract

Experimental research on behavior and cognition frequently rests on stimulus or subject selection where not all characteristics 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. Similarly, when presenting word stimuli, factors such as word frequency are often correlated with primary factors of interest. One procedure very commonly employed to control for such nuisance 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 high error rates and is conceptually misguided. It reflects a common misunderstanding of statistical tests: interpreting significance not to refer to inference about a particular population parameter, but about 1. the sample in question, 2. the *practical relevance* of a sample difference (so that a nonsignificant test is taken to indicate evidence for the absence of relevant differences). We show inferential testing to be inappropriate both pragmatically and philosophically, present a survey showing its high prevalence, and briefly discuss an alternative in the form of regression including nuisance factors.

Introduction

Methods sections in many issues of *Brain & Language* and similar journals feature sentences such as

> Animate and inanimate words chosen as stimulus materials did not differ in word frequency ($p > 0.05$).
>
> Controls and aphasics did not differ in age ($p > 0.05$).

In the following, we discuss the inappropriateness of this practice. A common problem in brain and behavioral research, where the experimenter cannot freely determine every stimulus and participant characteristic, is the control of confounding/nuisance variables. This is especially common in studies of language. Typically, word stimuli cannot be constructed out of whole cloth, but must be chosen from existing words (which differ in many aspects). 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. A similar problem exists, for example, when comparing controls to specific populations, such as bilingual individuals or slow readers. The basic problem 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 correlated with the variable of interest. A prevalent method we find in the literature, namely inferential null hypothesis significance testing (NHST) of stimuli, fails to perform the necessary control.

NHST and nuisance control

Often, researchers will attempt to demonstrate that stimuli or participants are selected so as to concentrate their differences on the factor of interest, i.e. reduce confounds, by conducting null-hypothesis testing such as t-tests or ANOVA on the potential confound in addition to or even instead of showing descriptive statistics in the form of measures of location and scale. The underlying intuition is that these tests establish whether two conditions differ in a given aspect and serve as proof that the conditions are "equal" on it. This is, in turn, based on the common, but incorrect intuition that significance in NHST establishes that a contrast shows a *meaningful effect*, and the related issue that non-significant tests indicate the absence of meaningful effects.

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

- Philosophically, these tests are inferential tests being performed on closed populations, not random samples of larger populations. Statistical testing attempts to make inferences about the larger population based on randomly selected samples. Here, the "samples" are not taken randomly, and we are not interested in the population they are drawn from, but in the stimuli or subjects themselves. 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 inanimate 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. Tests refer to the population of stimuli that will largely not be used, or the population of subjects that will not be investigated in the study.

- Pragmatically, beyond being inappropriate, 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, this procedure tests whether the conditions differ in a certain respect to a measurable degree, 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. Cohen 1992), 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 enough evidence to exclude this hypothesis (which we are not actually interested in). In typical contexts (e.g. setting the Type I rate to the conventional 5% level), the power to reject the null hypothesis of no differences is low (Button et al. 2013) due to a small number of items, meaning that even comparatively large differences may be undetected, while in larger sets, even trivially small differences may be rejected. Especially with small samples (e.g., 10 subjects per group, or 20 items per condition), the probability of detecting moderate confound effects is thus low – even if there are substantial differences, tests will not reject the null hypothesis, and stimulus sets might be accepted as being balanced based on a test with a low probability of rejecting even moderately imbalanced samples of such a size.

In other words, these tests are incapable of actually informing us about the influence of potential confounds, but may give researchers a false sense of security. This inferential stage offers no benefit beyond examining the descriptive measures of location and scale (e.g. mean and standard deviation) and determining 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 to what extent these potential confounds have an influence on the results, such as by including these confounds in the statistical model. This is often readily implemented via multiple regression, particularly "mixed-effect" approaches (Gelman and Hill 2006; Fox 2016).

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 a "randomization check" as it refers to checking if assignment of subjects to

treatments has truly been performed randomly. Philosophically, this is somewhat less misguided, but has also been determined to be pragmatically inappropriate. 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). That is, in the case of medical studies with randomization checks, experimental validity is achieved by selecting subjects from a given population and randomizing their assignment to treatment. In the case of studies in the brain and behavioral sciences, stimuli are constructed so as to differ on one factor which we highly expect to be correlated with others, e.g. word frequency and word length, and the concern of researchers is not if assignment was random (in fact, it is known to not have been random), but if stimuli differ systematically on factors expected to impact the dependent variable of interest. We are not aware of similar discussions in the psychological, linguistic or neurocognitive literature. Nonetheless, the clinical trial literature provides important considerations for experimental design choices, e.g. the proper way of blocking and matching (Imai, King, and Stuart 2008), and can thus inform preparing stimulus sets or participant groups even for non-clinical experiments.

Prevalence

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

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 [...], and in length (both t's < 1.1)

Here, the authors deduce equivalence (acceptance of the null) from a failed test (i.e. a test where the null cannot be rejected), with regards to the population of stimuli they did not present rather than the sample at hand. 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 were found where researchers reported known quantities (e.g. perfectly measurable characteristics of a fixed set of stimuli) 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 confounds of subject groups or experimental stimuli, they conduct inappropriate tests and misinterpret the results of these tests in a potentially misleading manner.

Representative statements from every study committing an error as well as further details on the precise survey methodology are available online at https://github.com/jona-sassenhagen/statfail.

Simulation

In addition to our survey of the literature, we also performed a simulation to examine the impact of inferential tests on confounding variables. In our simulation, we examined the role of:

1. **manipulation effect size:** the effect size from the manipulation of interest
2. **confound feature size:** the measured size of the confounding feature
3. **confound feature-effect correlation:** the correlation between that measured size the and the effect size of the feature

The last two emphasize the pragmatic problem the procedure: inferential tests on group attributes (e.g. word frequency vs. condition in language studies) examine difference in the feature and not the impact of that (difference in the) feature on the outcome. In our simulation, this is equivalent to assuming that the measured feature difference exactly correlates with the impact that feature has on the outcome, which is a fairly strong assumption.

The results of this simulation for several static values of the above factors are available online on RPubs (http://rpubs.com/palday/statfail), while an interactive version is available online at ShinyApps (https://palday.shinyapps.io/statfail/). All source code (in R) is available on GitHub (https://github.com/jona-sassenhagen/statfail), including the ability to run the simulation on a local computer.

In particular, we find that when the correlation between feature size and feature effect is not perfect, testing covariates can lead to false rejections of manipulations as "confounded" in 50% or more of studies for even large effects (20 items each

for 2 groups, Cohen's $d = 2$ for manipulation, Pearson's $r = 0.75$ between feature and its effect; rejection rate $> 60\%$ for Cohen's $d = 1$ for feature attribute, rejetion rate $= 75\%$ for Cohen's $d = 2$ for feature attribute), *even when the confounding covariate was not significant in multiple regression.*

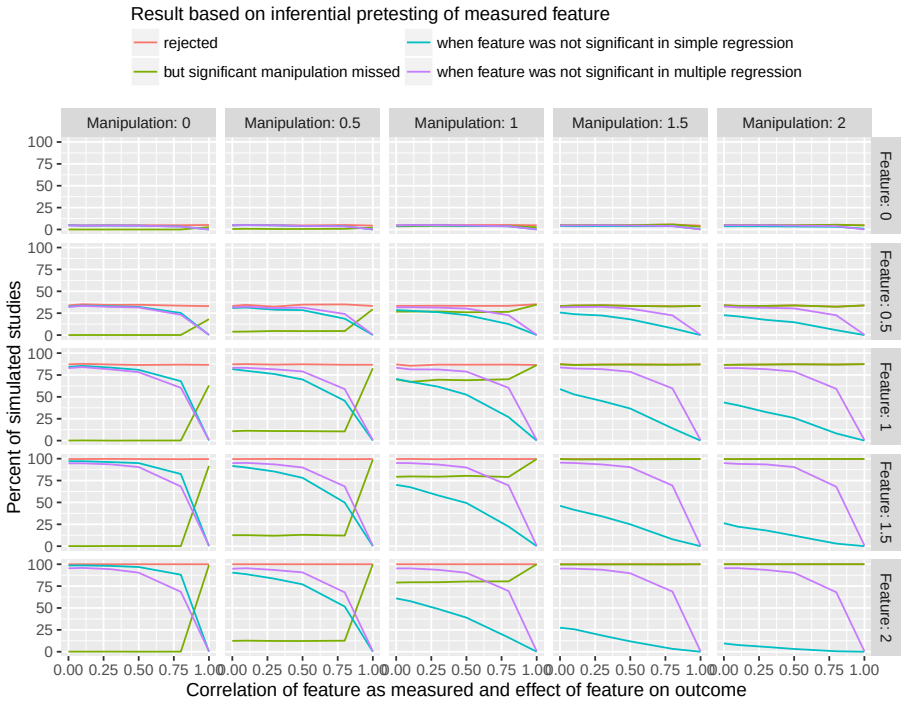

Figure 1: Simulation results for rejections. The difference in feature size and the effect size of the manipulation are given in Cohen's d. The feature size is not the size of its effect, but rather the measured difference in the feature itself. The impact of the feature is determined by the correlation between its measured size and effect size. Rejections where the significant manipulation was missed are an example of throwing the baby out with the bathwater. Rejections where the confounding feature was not a significant predictor in simple or multiple regression are arguably rejections where there is no effect of the feature.

Discussion and recommendation

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

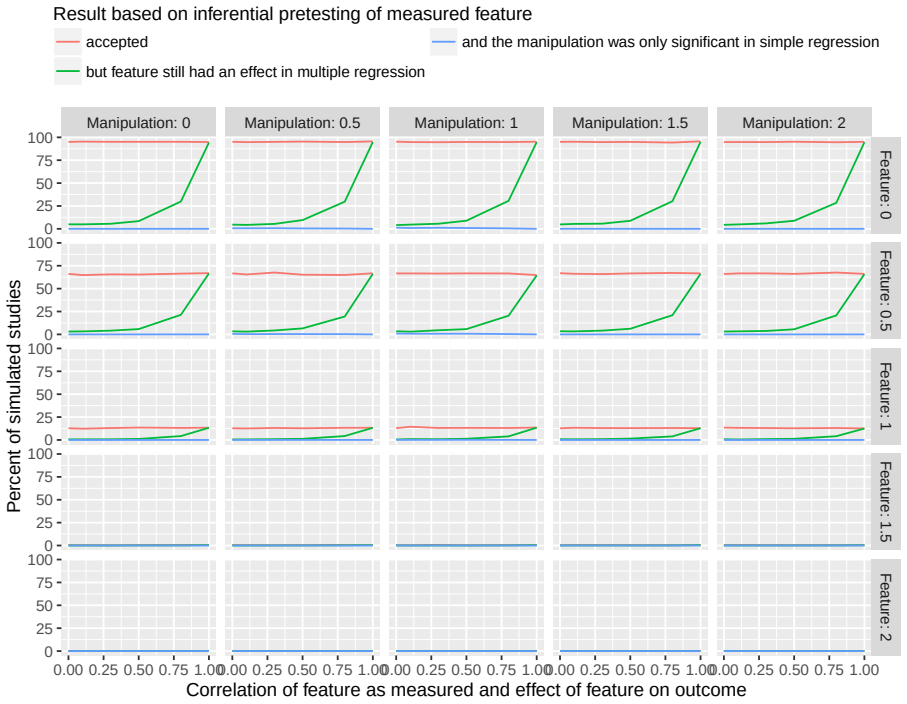

Figure 2: Simulation results for acceptances. The difference in feature size and the effect size of the manipulation are given in Cohen's d. The feature size is not the size of its effect, but rather the measured difference in the feature itself. The impact of the feature is determined by the correlation between its measured size and effect size. Acceptances where the feature still had an effect in multiple rejection are failed detections of a real confounding effect. Acceptances where the manipulation was only significant in simple regression are failed detections of a real confounding effect that completely subsumes the manipulation effect.

Researchers should still use descriptive statistics to demonstrate the success of balancing. That is, quantifying e.g. differences between stimuli via variances, raw means and standardized means (Cohen's d), and correlation coefficients, which many researchers already often do, can be highly informative, and should be routinely done. For more complex designs, cross-correlation matrices can visualize the degree of confounding. In contrast, 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 can include confounds in their statistical model for the data. For traditional t-tests, ANOVAs and regression models, this corresponds to using multiple regression with the confounds as additional nuisance factors (including continuous factors). In multiple regression, all parameters are jointly estimated, and assuming the model is correct (includes all relevant variables) and the included variables are reliably measured (Westfall and Yarkoni 2016), these estimates are unbiased.[1] Thus, a manipulation effect estimated by a model also containing nuisance factors corresponds to the effect of manipulation while accounting for nuisance influence. Importantly, to prevent p-value "fishing", the choice of selecting covariates to include must be made on principled grounds, 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 (computed with e.g *lme4* (Bates et al. 2015)) have the additional advantage of accounting for the combined variance of subjects and items in one model (Clark 1973; Baayen, Davidson, and Bates 2008; Judd, Westfall, and Kenny 2012) and automatically provide a summary of correlation between effects.

One problem in this context is that these stimulus confounds are often correlated with one another, the dependent variables, and the independent variables of interest (e.g., word frequency and word length correlate). Under multicollinearity, standard errors may be inflated. The main technique for dealing with collinearity is one that researchers traditionally already employ: attempting to balance stimulus/subject selection so that differences in confounds are minimised, e.g. via matching or blocking. That is, matching should generally still be performed in addition to multivariate estimation.

Finally, effective parameter estimation in complex regression models requires 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 nuisance control are:

[1] More precisely, the ordinary least-squares procedure is, under the usual assumption of homoskedastic errors with mean zero, the best linear unbiased estimator (BLUE) of the model coefficients.

- attempt to match nuisance factors to a reasonable degree
- use descriptive, but not inferential statistics to guide stimulus selection
- add potentially confounding factors as covariates into the final data analysis process
- use larger samples to provide adequate power

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; Katerina Starikova, Miriam Burk and Antonia Götz are to be thanked for collecting the survey data. This work was supported in part by the German Research Foundation (BO 2471/3-2) and by the ERC grant 617891.

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: 335–59.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112 (1): 55–159.

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." *J Pers Soc Psychol* 103 (1): 54–69.

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: 1715–26.

Westfall, Jacob, and Tal Yarkoni. 2016. "Statistically Controlling for Confounding Constructs Is Harder Than You Think." *PLoS ONE* 3 (11): 1–100.