CDDRL

WORKING PAPER

August 2015

A Method to Reduce Type 1 Error while Maintaining Power Using Split Samples

Marcel Fafchamps
Stanford University

Julien Labonne
University of Oxford

Stanford | CDDRL Center on Democracy, Development, and the Rule of Law

FSI Freeman Spogli Institute for International Studies

CDDRL

ABOUT CDDRL

Since 2002, the Center on Democracy, Development and the Rule of Law (CDDRL) at Stanford University has collaborated widely with academics, policymakers and practitioners around the world to advance knowledge about the conditions for and interactions among democracy, broad-based economic development, human rights, and the rule of law.

The mission of CDDRL is to understand how countries can overcome poverty, instability, and abusive rule to become prosperous, just, democratic, and well-governed states. This concern for the overall trajectory of national development—and for the intricate links among the economic, political, legal, social, and health dimensions of development—sets CDDRL apart from other research centers.

Center on Democracy, Development and the Rule of Law
Freeman Spogli Institute for International Studies
Stanford University
Encina Hall
616 Serra St.
Stanford, CA 94305-6055

Voice: 650-723-4610
Fax: 650-724-2996

Website: http:/cddrl.fsi.stanford.edu

A Method to Reduce Type I Error while Maintaining Power Using Split Samples

Marcel Fafchamps and Julien Labonne[*]

Draft
August 2015

Abstract

We discuss a method aimed at reducing the risk that spurious results are published. Researchers send their datasets to an independent third party who randomly generates training and testing samples. Researchers perform their analysis on the former and once the paper is accepted for publication the method is applied to the latter and it is those results that are published. Simulations indicate that, under empirically relevant settings, the proposed method significantly reduces type I error and delivers adequate power. Unlike alternative approaches such as the registration of a pre-analysis plan, this method allows researchers to learn from the data.

[*] Fafchamps: Stanford University, Freeman Spogli Institute for International Studies, Encina Hall E105, Stanford CA 94305 USA (fafchamp@stanford.edu). Labonne: Yale-NUS College, Singapore (julien.labonne@yale-nus.edu.sg). We thank Rob Garlick for fruitful discussions while working on this paper. All remaining errors are ours.

1 Introduction

The gap between econometric theory and practice makes it challenging to assess the reliability of empirical findings in economics and political science (Leamer 1974 Leamer 1978 Leamer 1983 Lovell 1983 Glaeser 2006). This is due to a combination of researchers' degree of freedom and publication bias. As a result, the probability of Type I error in published research is believed to be larger than the commonly accepted five percent. For example, Gerber and Malhotra (2008) and Brodeur, Le, Sangnier and Zylberberg (forthcoming) report that there is a bunching of p-values just below the 0.05 threshold in top economics and political science journals. This is consistent with researchers and editor unconsciously or consciously selecting outcome variables, regression methods, estimation samples, and control variables to deliver significant results.[1]

A number of reforms of the reviewing process have been proposed to decrease the risk that spurious findings are published and cited (Green, Humphreys and Smith 2013). The common objective is to encourage researchers to transparently select which statistical tests to implement *before* accessing the data on which they will be run. A prominent example is the introduction of pre-analysis plans (PAPs).[2] Such plans are written – and possibly shared with the research community – before any analysis is carried out. This reduces the risk that researchers select hypotheses that can be rejected with the available data (Casey, Glennerster and Miguel 2012 Olken 2015). Some researchers are more skeptical and argue that the profession should encourage replications instead (Coffman and Niederle 2015). The profession has also insisted on the need to correct for multiple comparisons.

While filing a PAP may offer some protection in lab and field experiments, it offers no protection in the case of analysis of observational data. In this paper we discuss an alternative – and possibly complementary – method that can be applied to both new and existing datasets. The process involves sending the data to an independent third party who randomly generates two non-overlapping subsets of the data. Researchers only have access to one subset – called the training dataset – while the third party keeps the second one – called the testing dataset.

[1] This behavior isn't confined to economics and political science, however. As reported by Shea (2013), Dirk Smeesters, a social psychologist at Erasmus University Rotterdam, whose own university publicly announced that it had "*no confidence in the scientific integrity*" of three of his articles, stated that "*many authors knowingly omit data to achieve significance, without stating this*".

[2] Building on J-PAL Hypothesis Registry, The American Economic Association has recently set-up a RCT registry. To-date the American Political Science Association hasn't followed suit but the E-GAP (Experiments in Governance and Politics) network allows researchers to register both experimental and observational studies.

Researchers are free to analyze the training dataset and can interact with seminar audiences, editors and referees based on results obtained from it. Once the paper has been accepted for publication, the analysis is implemented, unchanged, on the testing dataset, and these results are the ones that are published.[3] An important feature of the method is that only hypotheses that are rejected on the training sample are carried over the testing sample. As we correct for multiple testing, this compensates for the loss of power due to smaller sample sizes.

The proposed method offers three methodological benefits: reduced type I errors, as in PAPs; increased ability to learn from the data and to test hypotheses that researches did not think about when they started their project – an issue that is not addressed by PAPs; and reduced the risk of publication bias. The main potential cost of split samples is loss of power. First, it reduces the risk of Type I error because researchers fully specify the regressions they want to estimate before having access to the dataset on which hypotheses will be tested. This reduces the risk of focusing on specifications where a spurious null happens to be rejected. Second, the method allows researchers to test hypotheses that they did not think about before starting their analysis: researchers can refine their research plans based on initial findings, interactions with seminar audiences, and requests from referees and editors.[4] Third, the method reduces the risk of publication bias because journal editors decide whether to publish a paper before seeing the final results.[5] This last benefit could also be achieved by PAPs if journal editors were willing to accept a paper on the basis of the quality of PAP design alone.

To capture these features in a simple way, we imagine a situation in which the researcher wishes to test multiple hypotheses, without strong a priori information on which hypotheses are most relevant. In such a situation, it is common for researchers to adjust for multiple testing. We present results from simulations quantifying the trade-off between reduced type I error and

[3] In the recent field of genoeconomics, researchers often attempt to replicate their findings by testing whether the identified genes are correlated with the outcomes of interest on another sample (Benjamin, Cesarini, Chabris, Glaeser, Laibson, Guonason, Harris, Launer, Purcell, Smith, Johannesson, Magnusson, Beauchamp, Christakis, Atwood, Hebert, Freese, Hauser, Hauser, Grankvist, Hultman and Lichtenstein 2011). Other researchers have used a related method and construct a genetic score on a sub-sample and check its predictive accuracy on the remaining sample (Benjamin, Cesarini, van der Loos, Dawes, Koellinger, Magnusson, Chabris, Conley, Laibson, Johannesson and Visscher 2012). In both cases, no third-party is involved and researchers have control over the choice of the other sample and over the split between the training and the testing sample.

[4] One could argue that additional hypotheses can be addressed in future research. But given the cost of collecting additional data and the long publication lag in economics, this would unnecessarily delay the availability of evidence.

[5] Franco, Malhotra and Simonovits (2014) take advantage of an NSF-sponsored program to quantify publication bias. They show that strong results are much more likely to be published. This effect is partially explained by the fact that researchers do not write up null findings (Franco et al. 2014), and partly by the fact that editors and referees are reluctant to publish null results.

loss of power, compared to a situation where the researcher test hypotheses on the full sample. In both cases we adjust for multiple testing. We show that the loss of power from using split samples instead of the full sample is lowest when the total number of tests is large – that is, when the researcher most wishes to learn from the data. In this case, multiple comparison adjustments can induce a large reduction in power when using the full sample. The split sample approach allows the researcher to curtail the number of tests carried on the testing sample, and this compensates for the loss of power due to smaller sample size. We also provide guidance on the optimal way of splitting the full sample into training and testing subsamples.

Results presented in the paper indicate that in a large number of relevant empirically settings, the loss of power associated with the split sample is manageable. Economically significant effect size (above 0.2 standard deviation) can be detected with power comfortably above 80 percent as soon as sample size is above 3,000. For a smaller effect size (*e.g.,* 0.1 standard deviation), a sample of 10,000 observations or more is required. In addition, we provide evidence that the split sample approach is more likely than a PAP approach to identify a null hypothesis that should be rejected. When using a PAP, researchers often keep the number of tested hypotheses small to counteract the loss of power due to multiple testing. Our proposed method allows researchers to test a large number of null hypotheses with only a small loss in power. As a results, they are better able to learn from the data.

It is important to note that there are other ways through which spurious results can be published, but dealing with them is beyond the scope of this paper. The method would still deliver biased estimates if researchers use unreliable data, or faulty code and software. For example, Bell and Miller (forthcoming) could replicate Rauchhaus (2009)'s findings in STATA but not in R, which they attribute to a problem in STATA. More perniciously, some researchers have been caught fabricating data. In line with current practice, we argue that the best way to deal with those issues is to ask researchers to make their code and data publicly available after publication. This would increase the likelihood that potential mistakes are quickly identified.

The remainder of the paper is organised as follows. In Section 2, we present a canonical setup often encountered in empirical work. The proposed method is described in Section 3 and results are discussed in Section 4. Section 5 concludes.

2 The Problem

In this section, we discuss current empirical practices and why they might lead to the publication of spurious findings. We also describe how researchers currently attempt to deal with those issues.

2.1 Canonical set-up

We consider the following canonical setup. Researcher A is interested in estimating the effect of an exogenous treatment T (with $T = 1$ for half of the observations). She has access to a sample S of size N that includes a set of m potential outcome variables $(y^k)_{k=1,...,m}$. The m outcome variables can either capture different concepts, related concepts, or different ways of measuring the same concept. For example, the researcher may have access to firm data on firms' hiring practices, number of employees, value-added, profits, etc. Unsure of which aspects of firm performance is affected by treatment, the researcher runs regressions of the form:

$$y^k = a + b_k T + u \tag{1}$$

The researcher then runs a series of tests $H_0^k : b_k = 0$. Some of these null hypotheses are true, some are non-true. The researcher faces a multiple comparison problem: without adequate adjustment, the probability that a true null hypothesis is rejected is higher than the level α at which each individual test is carried out. The set-up, adapted from Benjamini and Hochberg (1995), is summarised in Table 1.

Table 1: Set-up

	Declared Non-significant	Declared Significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
Total	m-R	R	m

The researcher is concerned about Type I errors and wants to find ways to control the Family Wise Error Rate (FWER).

Definition 1 *The Family Wise Error Rate is the probability of rejecting at least one true null hypothesis. In the notation of Table 1, it is equal to $Pr(V > 0)$.*

The most basic way to keep the FWER in check is to make Bonferroni adjustments: instead of rejecting H_0 if the p-value is smaller than α, reject if it is smaller than α/m. Let R_k be a variable indicating whether hypothesis k was rejected. It is straightforward to show that the adjustment controls the FWER:

$$P(V>0) \leq P(R>0) = P(\bigcup_{k=1}^{m} R_k) \leq \sum_{k=1}^{m} P(R_k) = m * \frac{\alpha}{m} = \alpha$$

The adjustment is only valid if all null hypotheses are true $(m = m_0)$ and all tests are independent. It is well known that this correction tends to be very conservative and can lead to serious loss of power. In addition, the method is only valid if the researcher can keep track of all tests she performed. If for example, the researcher ran $m^{'}$ tests and attempt to control the FWER as if only m tests had been carried out (with $m < m^{'}$), the reported FWER will understimate the actual FWER.

Let α be the significance level used to test H_0^k and let δ_k be the standardized effect size for the $m - m_0$ non-true null hypotheses. In this convenient set-up we can use standard power calculations formula (see McConnell and Vera-Hernández (2015)). Power, denoted as $1-\beta$, is the probability of rejecting the null hypothesis when the alternative is correct. Under our assumptions, it is given by:

$$1 - \beta_k = \Phi(\delta_k \sqrt{\frac{N}{4}} - Z_{1-\frac{\alpha}{2}}) \tag{2}$$

where Φ is the cumulative distribution function for the standard normal distribution. The detailed calculations are available in the Appendix. If the researcher carries out Bonferonni corrections, power becomes:

$$1 - \beta_k^{Bonf} = \Phi(\delta_k \sqrt{\frac{N}{4}} - Z_{1-\frac{\alpha}{2m}}) \tag{3}$$

Comparing the two formulas directly shows that Bonferonni corrections lead to a loss of statistical power. This loss is increasing in m, the number of tests that are carried out.

Since the probability of rejecting each true null hypotheses is α, the probability of rejecting at least one is given by:

$$FWER = 1 - (1-\alpha)^{m_0} \tag{4}$$

where m_0 is the (unknown) number of true null hypotheses. It is important to note that the FWER is not a function of sample size or effect size. If the researcher carries out Bonferonni corrections, the FWER becomes:

$$FWER^{Bonf} = 1 - (1 - \frac{\alpha}{m})^{m_0} \tag{5}$$

2.2 Pre-Analysis Plan

Before having access to the data, the researcher can prepare and register a pre-analysis plan (Coffman and Niederle 2015 Olken 2015). Such a plan lists the hypotheses to be tested and describes how they will be tested, including which variables to include, how they will be included, and how researchers intend to deal with the multiple comparison problems.

This approach is appealing but it has some drawbacks. First, unless pre-analysis plans fully specify the regressions to be estimated, it still leaves some room for data mining. As a result, Humphreys, Sanchez de la Sierra and van der Windt (2013) argue that researchers should write a mock report with fake data. This forces researchers to make all decisions regarding the analysis (including micro-decisions such as the precise way of defining all variables) before having access to the dataset on which the regressions will be estimated. The methodology is then applied to the real data.

Second, following a PAP to the letter does not allow researchers to learn from the data, and this can slow down the pace of new discoveries. Indeed, PAPs can only cover hypotheses that the researcher could think of before carrying out their experiment. There often are other testable hypotheses that the researcher did not think of beforehand. A number of social scientists have recently argued that some of their most important findings were the direct result of time spent with the data (Laitin 2013 Gelman 2014). For example, Simonsohn (cited by Laitin (2013)) argues that: "*I also think of science as a process of discovery . . . Every paper I have [written] has some really interesting robustness, extensions, follow-ups that I would have never thought about at the beginning.*" Similarly, Gelman (2014) states that "*Many of my most important applied results were interactions that my colleagues and I noticed only after spending a lot of time with our data.*"

Third, it is difficult to credibly implement PAPs in observational studies because it is difficult to guarantee that the researcher has not run the regressions before registering the PAP. This concern is especially acute in situations where the data have already been used by other researchers. PAPs are better suited for analysis of experimental data.

Fourth, unless editors are willing to unconditionally accept papers based on a detailed pre-analysis plan, there is always room for what Pepinsky (2013) refers to as *referee degree of freedom*, i.e., the referees (and editor) may require the researcher to conduct analysis that was not in the PAP.

Fifth, PAP forces researchers to divulge their research design with other, possibly competing researchers at an early stage of the research process. Given the long publication lags in economics, this opens the door to abuse.

Finally, as long as the decision to publish results is based on whether or not some null hypothesis is rejected, there remains a risk that, even if all research follows a PAP, many published findings are spurious. To illustrate, imagine m researchers, each with access to data on treatment T and one of the outcome variables $(y^k)_{k=1,...,m}$. Each of these m researchers registers a PAP to estimate the effect of T on a single y^k. All tests for which the null is rejected are then published. Ioannidis (2005) argues that since there are many more true null hypotheses than false ones, as long as m is sufficiently large there will be more cases of Type I error than of cases where the null is correctly rejected.

3 The Method

We now describe the split sample approach in details.[6] As above, we assume that researcher A is interested in estimating the effect of T on a list of possible outcomes $(y^k)_{k=1,...,m}$. There is some uncertainty regarding which particular hypotheses to test and how to best test them. The research project proceeds as follows:

- Step 1: Guided by theory and existing evidence, researcher A puts together a sample S including a number of variables that broadly captures the general set of hypotheses that she wants to test. The researcher also includes variables used to test for heterogeneous effects.

- Step 2: The data is then sent to a third-party B who randomly generates two non-overlapping subsets. If the researcher is interested in studying particular subgroups the sample should be stratified accordingly. The first sub-sample (*training sample*) is sent back to A. The third-party keeps the second one (*testing sample*). All relevant IDs are scrambled during the process so that A is unable to 'reverse engineer' the randomization.

[6] Our method differ from earlier efforts to use split-sample in applied econometrics. Researchers focused on pretesting bias; more specifically of how the potential bias arising from dropping regressors based on the associated t-statistics in both OLS and IV estimation (e.g. Angrist and Krueger (1995)). Researchers were concerned about the determinants of a single outcome variables. We are concerned about how one treatment variable affects a large number of potential outcome variables.

- Step 3: A runs regressions, presents the results at seminars and conferences, and refines the methodology based on feedback received.

- Step 4: The paper is submitted to a journal, referees make their comments and A amends her analysis in response, possibly several times.

  The discovery process described by steps 3 and 4 identifies a final subset J of the m outcome variables such that each of these outcome variables is significant at the α level in the training set, conditional on a choice of estimator, control variables, and standard error correction. We call this the final methodology for analysis. In most contexts $J \ll m$ which compensates for the loss of power due to lower sample size when correcting for multiple testing.

- Step 5: The editor accepts the paper conditional on the agreed upon final methodology for analysis. A then secures the testing sample from B and applies the agreed upon methodology to it. The published version of the paper only includes the results obtained from the testing sample.

We think of Steps 3 and 4 as a way for researchers to refine their research plan. The methodology that is accepted in Step 5 is akin to a detailed pre-analysis plans that fully specifies the regressions to be estimated on the testing sample. As researchers have the opportunity to interact with the data, seminar audiences, and referees, there is room in the analysis plan to incorporate interesting hypotheses that A would not have tested otherwise.

More formally, the process looks as follows. Researcher A puts together a sample S that includes N observations and a set of $m+1$ variables: T_i and $(y^k)_{k=1,\ldots,m}$. A third party B then randomly splits the dataset into two sub samples S_1 and S_2 such that: $S = S_1 \cup S_2$ and $S_1 \cap S_2 = \emptyset$. At first, the researcher does not have access to S_2. The researcher starts with a set of specific hypotheses to test. Feedback from other researchers is then used to help A finalize a list of hypotheses to test. This list can be represented most generally as a series of J triplets consisting of: (1) a set of outcome variables $(z^j)_{j=1,\ldots,J}$ which we allow to be transformations of the original data $(y^k)_{k=1,\ldots,m}$ such that $\forall j$ we have $z_i^j = f(y_i^1, \ldots, y_i^m)$;[7] (2) an estimation method (e.g., estimator, control variables); and (3) a set of rules that define the estimation sample (e.g., excluded outliers). Once the J triplets are agreed with an editor, the associated regressions are

[7] In the simplest case, the $(z^j)_{j=1,\ldots,J}$ are simply the subset of the $(y^k)_{k=1,\ldots,m}$ variables that are significant at the α level.

estimated on S_2 and this is the set of results that are published. The method's key feature is that, given that the training and testing samples are independent, the probability of type I error in the two samples are independent.

4 Computations

4.1 Set-up

We now illustrate the method for the canonical setup described above. We compute power and FWER under the full sample approach and the proposed split sample approach. In both cases, we present results both with and without Bonferonni adjustments. We show the sensitivity of power and FWER to variation in the following parameters: the sample size (N); the standardized effect size (δ); the number of tested hypotheses (m); the number of tested null hypotheses that are true (m_0); and the share of the total sample that is allocated to the training set (s).

Throughout we assume that the researcher starts with m possible null hypotheses. Of these, a subset J are found to be significant at the α level in the training set. This subset determines the list of tests estimated on the testing set. To illustrate, let $m = 20$ and imagine that, in the training sample, treatment is significant at the $\alpha = 5\%$ level for seven of these 20 outcome variables. Then we only regress treatment on these seven outcome variables in the testing sample. It is this shrinking of the set of hypotheses that delivers power while keeping FWER low, as we now demonstrate.

Split sample without Bonferonni correction We start by showing how the formula introduced in Section 2 can be adjusted to compute power and the FWER with the split sample methodology. For a null hypothesis to be considered to be rejected, it is necessary that it be rejected first on the training sample, and then again on the testing sample. As a result, power is given by:

$$1 - \beta_k^{Split} = \Phi\left(\delta_k \sqrt{\frac{sN}{4}} - Z_{1-\frac{\alpha}{2}}\right) \Phi\left(\delta_k \sqrt{\frac{(1-s)N}{4}} - Z_{1-\frac{\alpha}{2}}\right) \tag{6}$$

Split sample with Bonferonni correction With Bonferonni correction, the calculations are as above except that we need to account for the number of tests carried out on the testing sample. Power is the expected value of the following random variable:

$$1-\beta_k^{Split/Bonf}=\Phi\left(\delta_k\sqrt{\frac{sN}{4}}-Z_{1-\frac{\alpha}{2}}\right)\Phi\left(\delta_k\sqrt{\frac{(1-s)N}{4}}-Z_{1-\frac{\alpha}{2B}}\right) \tag{7}$$

where B is the number of tests carried out on the testing sample. It distributed according to:

$$B(m_0,\alpha)+B\left(m-m_0,\Phi\left(\delta_k\sqrt{\frac{sN}{4}}-Z_{1-\frac{\alpha}{2}}\right)\right) \tag{8}$$

where $B(n,p)$ is a binomial distribution with n trials and p probability of success in each trial. The number of tests conducted on the testing sample is the sum of two terms: the number of true null hypotheses that are incorrectly rejected on the training sample; and the number of non-true null hypotheses that are correctly rejected on the training sample. To obtain an approximation of expected power, we take 10,000 draws of the distribution B using (8), compute power (7) for each iteration, and then take the average over all 10,000 iterations.

4.2 Results

We now present the results from applying the above formulas and simulation method to various parameter values. To capture the idea that there are many more true null hypotheses than false ones, we organize the simulations around the assumption that, out of 100 possible null hypotheses, only one is non-true, i.e., should be rejected. Hence, unless stated otherwise, the results presented below are based on $m=100$ and $m_0=99$. Given these parameter values, the majority of the results found significant are spurious. For instance, if $\alpha=5\%$, there will on average be five false rejections and, provided that power is high enough, one true rejection in the training sample. For now we use a 50-50 split between the training and testing samples, i.e., we set $s=0.5$.

We organize our simulations around four stylized testing scenarios: (1) testing all 100 null hypotheses on the full sample without correction; (2) testing all 100 null hypotheses on the full sample with Bonferonni corrections; (3) testing all 100 null hypotheses on the training sample, and only testing on the training sample those null hypotheses that were significant in the training sample; and (4) proceeding as in (3) but adding Bonferroni corrections to the testing sample results.

We start by investigating the effect of sample size on the power to detect a true effect of size 0.2. In other words, we compute the likelihood of rejecting the null hypothesis when this hypothesis is false and the true effect is 0.2. Figure 1 plots power under the four scenarios for sample sizes varying between 500 to 10,000. Even with Bonferonni corrections, power under the split sample approach is well above 0.8 for the kind of sample sizes of 3,000 or more that are commonly encountered in empirical work. As expected, the Bonferonni corrections lead to a loss in power. But this loss of lower is less with the split sample approach than with the full sample approach. This makes sense because the split sample approach reduces the number of tests that are carried out on the testing sample.

Figures 2 and 3 plot similar results for different effect sizes of 0.1 and 0.3. Larger, but still relatively common, sample sizes are required to have power above 0.8 with smaller expected effect sizes (Figure 2). For example, with a small expected effect size of 0.1, raising power above 0.8 under the split sample approach requires sample sizes of 10,000 or more. When the expected effect size is 0.3, power under the split sample approach reaches 0.8 as soon as sample size is above 1,500.

So far we have set $m = 100$ and $m_0 = 99$. Next, we simulate what happens to power when we vary the total number of hypotheses that are being tested (m) and the number of non-true hypotheses (m_0). The effect size that we are trying to detect is 0.2, as in Figure 1. Figure 4 shows our simulation results for scenario (4) – the split sample approach with Bonferroni correction applied to the testing sample results. Results show that power is a decreasing function of m and m_0. This is because the Bonferroni correction becomes more stringent with m or m_0 increase.

Having shown that the split sample approach need not have a prohibitive cost in terms of loss of power, we now turn to its advantages in terms of minimizing the risk of false rejection. In Table 2 we compare the FWER under our four scenarios. Recall that the FWER is the probability of rejecting at least one true null hypothesis. We start by observing that, for $m = 100$ and $m_0 = 99$, the FWER is close to 1 in column, that is, when we test all 100 null hypotheses on the full sample without Bonferroni corrections. Even without Bonferroni correction, moving from the full sample approach to the split sample approach result in a massive reduction in the FWER from 0.994 to 0.219. This improvement is due solely to the reduction in the number of hypothesis tests that are carried out on the testing sample. If we add Bonferroni corrections, the FWER falls below 5% with or without split sample. The formal similarity between the two approaches is misleading, however. For the FWER to be truly below 5% in the full sample approach, the researcher must credibly track and report all the tests they run. We argue that this is unlikely to be the case in most empirical applications (Gelman 2013). In contrast, the split sample approach does not suffer from this type of under-reporting bias.

We also investigate whether it is optimal to split the sample 50-50 between training and testing sets. We continue to focus on scenario (4) – sample split with Bonferroni corrections – and we simulate power under alternative sample splitting rules, i.e., 30/70 and 70/30. The results, displayed in Figure 6, indicate that, across all considered sample sizes, a 50/50 split delivers the best power.

Next, we investigate whether power in the split sample approach with Bonferroni correction depends on the threshold level of significance used to select hypotheses in the training sample. So far we have assumed that this threshold is the same in the training and testing samples, i.e., $\alpha = 0.05$. We now compare this situation to using a threshold of 0.1 when selecting hypotheses on the training sample. Three effect sizes $0.1, 0.2$ and 0.3 are considered. We find that, for all three effect sizes, power appears to be marginally larger with a 0.05 threshold than a 0.1 threshold. This is because applying a less restrictive threshold to the training sample increases the number of true null hypotheses that are rejected, and thus the number of hypotheses that are tested on the testing sample. A larger number of hypotheses means that a stronger Bonferroni correction is required on the testing sample, and this is what drives the loss of power.

The split sample approach has one important additional benefit: it allows the researcher to test a large number of hypotheses with little loss in power. When using a PAP with full sample analysis, the researcher is often induced to select a short list of tested hypotheses in order avoid the loss of power due to Bonferroni correction. This short list typically includes hypotheses that the researcher a priori believes are most likely to be rejected. This means that many hypotheses (e.g., outcome variables) are excluded from the PAP, thereby preventing the researcher from learning from observations made during data collection and field experimentation. Because our method reduces the loss of power due to multiple testing, it allows researchers to learn from the data and to test hypotheses that they did not think about when they started the project.

To illustrate, let's imagine that the researcher has a dataset with 100 potential outcome variables but decided to only include 10 of them in the PAP. We keep other assumptions unchanged. In particular, we continue to assume that the null hypothesis should only be rejected for one of the 100 potential outcome variables. The question is whether this hypothesis is included in the shortlist or not. If it is, the shortlist approach yields correct inference. But if it is left out, the researcher might wrongly declare that the treatment has no effect. We now show that under a variety of settings the split sample approach reduces that risk.

Let ψ be the likelihood that the one hypothesis to be rejected is included in the shortlist of 10 tests. Once Bonferonni adjustments are taken into account, power under the full sample is given by:

$$Power_{PAP} = \psi * \Phi(\delta_k \sqrt{\frac{N}{4}} - Z_{1-\frac{\alpha}{2*10}}) \tag{9}$$

Power under the split sample approach with Bonferonni corrections is given as before by equation (7). Using these two formulas, we can compute the value ψ^* at which the two methods yield similar power. In Figure 7, we plot the value of ψ^* for various effect sizes (.1, .2 and .3). For all values of ψ below the curve, the split sample approach delivers more power. In a large number of cases, ψ needs to be close to one for the full sample approach with a PAP to be superior (or equivalent) to the split sample approach. For example, for effect sizes of .3 as soon as sample size is above 2,000, ψ needs to be one for the two approaches to yield similar results. Even with an effect size of .1 and a sample size of 7,000, ψ needs to be above .6 for the full sample approach with a PAP to dominate. This set of results thus confirms that the split sample approach increases researchers' ability to learn from the data.

5 Conclusion

In this paper we contribute to the nascent literature on ways to increase the likelihood that published findings are true. We investigate the effectiveness of a method that can be applied to both new and existing datasets. The method relies on a third-party randomly splitting the data in two non-overlapping subsets. Researchers use the first half to refine their research plan, present their findings during seminars and conference and submit them to journals. Once the paper is accepted, the precise research plan is then implemented on the second half and this is the set of results that are published.

We find that for a large number of empirically-relevant settings, the loss in statistical power associated with the split sample approach is manageable and we strongly encourage researchers to adopt the approach. This is especially true for quasi-experiments and observational relying on large datasets. For experiments, If researchers have strong prior that some hypotheses are true, they could set up a PAP for this subset of hypotheses. They could then use the split sample approach to test other, more exploratory hypotheses.

We believe that either journals or a professional association should set up and maintain an online platform where researchers can upload their dataset and have someone carry out the split sample. Importantly, the method can still be implemented by researchers working with proprietary data, e.g., researchers can send their anonymized dataset with garbled variable names to the third party.

References

Angrist, Joshua D. and Alan B. Krueger, "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 1995, *13* (2), 225–235.

Bell, Mark and Nicholas Miller, "Questioning the Effect of Nuclear Weapons on Conflict," *Journal of Conflct Resolution*, forthcoming.

Benjamin, Daniel J., David Cesarini, Christopher F. Chabris, Edward L. Glaeser, David I. Laibson, Vilmundur Guonason, Tamara B. Harris, Lenore J. Launer, Shaun Purcell, Albert Vernon Smith, Magnus Johannesson, Patrik K. E. Magnusson, Jonathan P. Beauchamp, Nicholas A. Christakis, Craig S. Atwood, Benjamin Hebert, Jeremy Freese, Robert M. Hauser, Taissa S. Hauser, Alexander Grankvist, Christina M. Hultman, and Paul Lichtenstein, "The Promises and Pitfalls of Genoeconomics," *Annual Review of Economics*, 2015/08/23 2011, *4* (1), 627–662.

___ , ___ , Matthijs J. H. M. van der Loos, Christopher T. Dawes, Philipp D. Koellinger, Patrik K. E. Magnusson, Christopher F. Chabris, Dalton Conley, David Laibson, Magnus Johannesson, and Peter M. Visscher, "The genetic architecture of economic and political preferences," *Proceedings of the National Academy of Sciences*, 2012, *109* (21), 8026–8031.

Benjamini, Yoav and Yosef Hochberg, "Controlling the False Discovery Rate: A Pactrical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1995, *57* (1), 289–300.

Brodeur, Abel, Mathias Le, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, forthcoming.

Casey, Katherine, Rachel Glennerster, and Edward Miguel, "Reshaping Institutions: Evidence on Aid Impacts Using a Pre-Analysis Plan," *Quarterly Journal of Economics*, 2012, *127* (4), 1755–1812.

Coffman, Lucas C. and Muriel Niederle, "Pre-Analysis Plans are not the Solution Replications Might Be," *Journal of Economic Perspectives*, 2015, *29* (3), 81–98.

Franco, Annie, Neil Malhotra, and Gabor Simonovits, "Publication bias in the social sciences: Unlocking the file drawer," *Science*, 2014.

Gelman, Andrew, "False memories and statistical analysis," 2013.

___ , "Preregistration: what's in it for you?," 2014.

Gerber, Alan and Neil Malhotra, "Do statistical reporting standards affect what is published? Publication bias in two leading political science journals," *Quaterly Journal of Political Science*, 2008, *3*, 313–326.

Glaeser, E., "Researcher Incentives and Empirical Methods," *Harvard Institute of Economic Research, Discussion Paper Number 2122*, 2006.

Green, Don, Macartan Humphreys, and Jenny Smith, "Read it, understand it, believe it, use it: Principles and proposals for a more credible research publication," *Columbia University, mimeo*, 2013.

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt, "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration," *Political Analysis*, 2013, *21* (1), 1–20.

Ioannidis, John, "Why Most Published Research Findings Are False," *PLOS Medicine*, 2005, *2* (8).

Laitin, David D., "Fisheries Management," *Political Analysis*, 2013, *21*, 42–47.

Leamer, Edward, "False Models and Post-Data Model Construction," *Journal of the American Statistical Association*, 1974, *69* (345), pp. 122–131.

——, *Specification Searches. Ad Hoc Inference with Nonexperimental Data*, New York, NY: Wiley, 1978.

——, "Let's Take the Con out of Econometrics," *American Economic Review*, 1983, *73* (1), 31–43.

Lovell, M., "Data Mining," *Review of Economic and Statistics*, 1983, *65* (1), 1–12.

McConnell, Brendon and Marcos Vera-Hernández, "Going beyond simple sample size calculations: a practitioner's guide," *IFS Working Paper W15/17*, 2015.

Olken, Benjamin, "Pre-Analysis Plans in Economics," *Journal of Economic Perspectives*, 2015, *29* (3), 61–80.

Pepinsky, Tom, "The Perilous Peer Review Process," 2013.

Rauchhaus, Robert, "Evaluating the Nuclear Peace Hypothesis A Quantitative Approach," *Journal of Conflict Resolution*, 2009, *53* (2), 258–277.

Shea, Christopher, "The Data Vigilante," *The Atlantic*, November 28 2013.

Wittes, Janet, "Sample Size Calculations for Randomized Controlled Trials," *Epidemiologic Reviews*, 2002, *24* (1), 39–53.

Table 2: Family Wise Error Rate

		Full Sample		Split Sample	
m	m_0	Bonferonni Corrections:			
		No	Yes	No	Yes
10	9	0.370	0.044	0.022	0.018
100	90	0.990	0.044	0.202	0.016
100	99	0.994	0.048	0.219	0.048
1,000	900	1	0.044	0.895	0.016
1,000	990	1	0.048	0.916	0.041

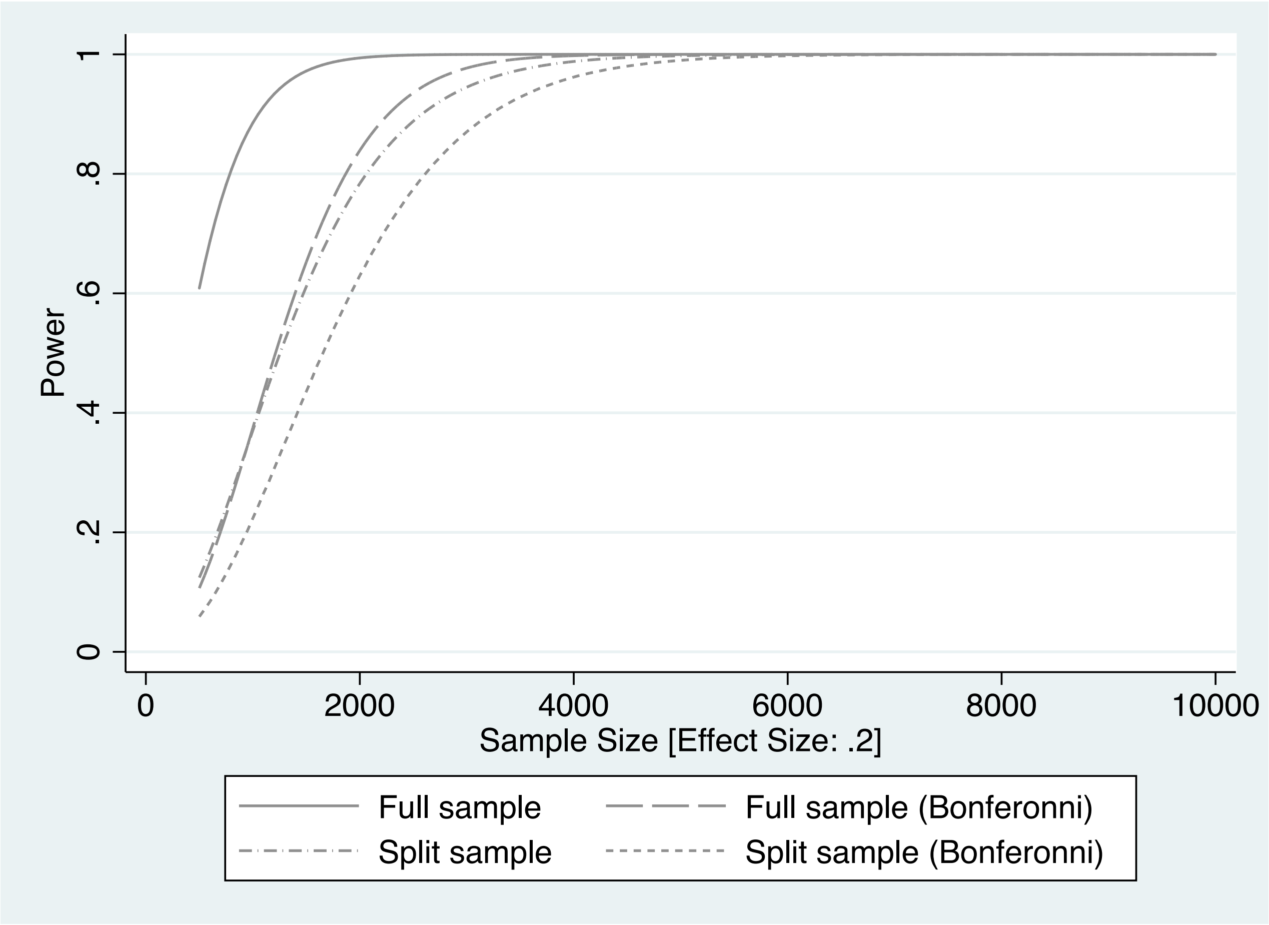


Figure 1: Comparing Power : Full Sample vs. Split Sample [Effect size = .2]

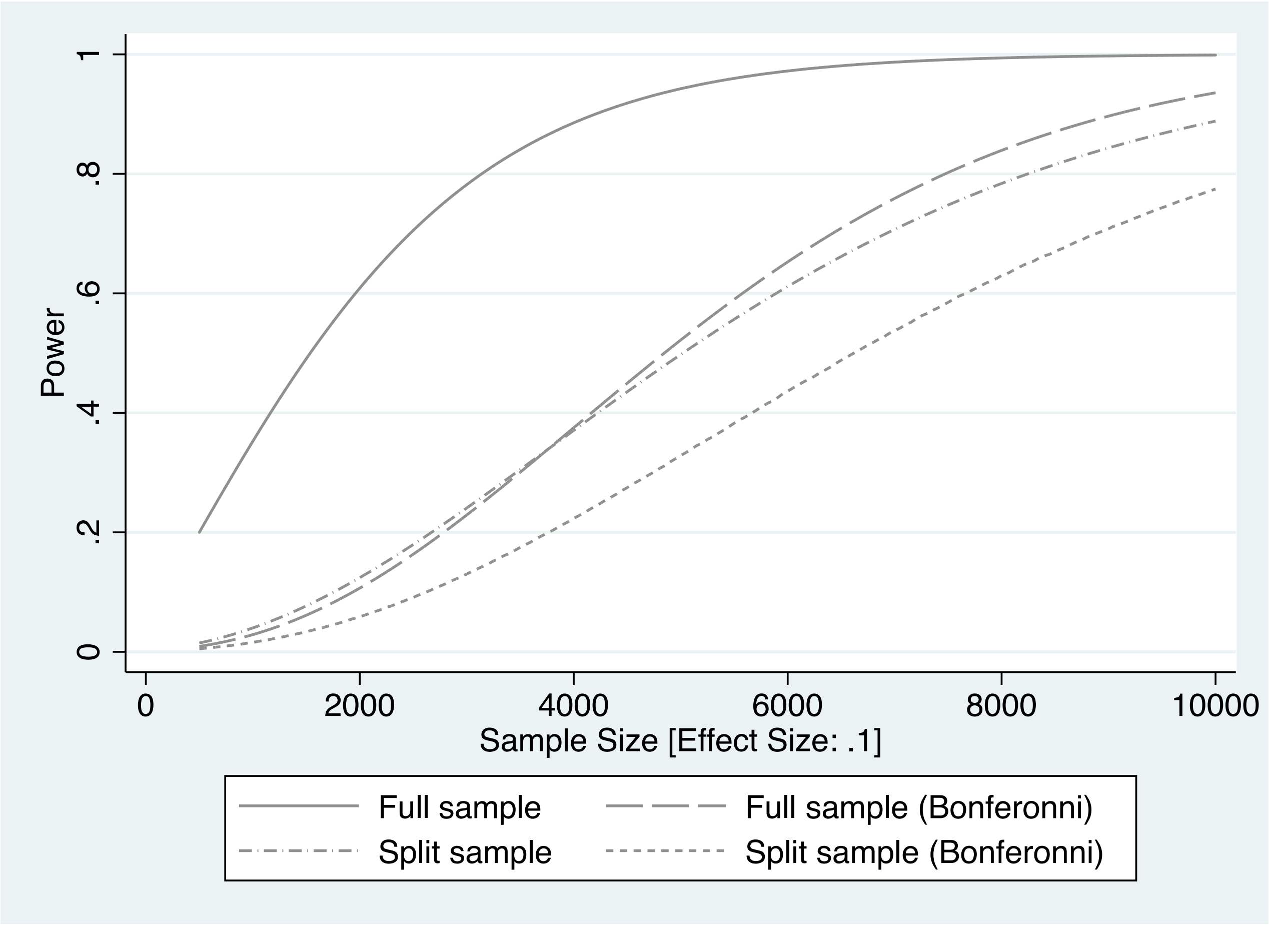


Figure 2: Comparing Power : Full Sample vs. Split Sample [Effect size = .1]

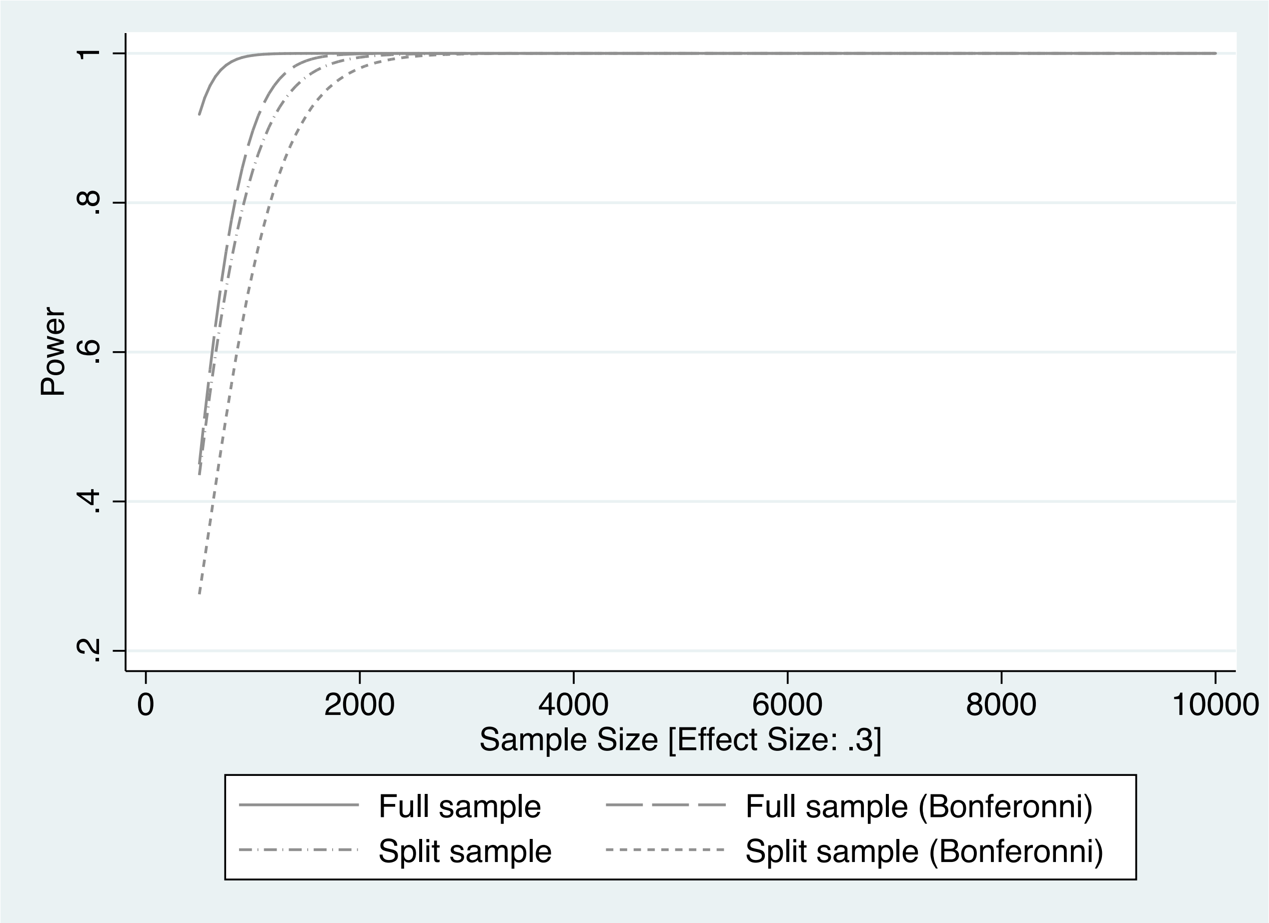


Figure 3: Comparing Power : Full Sample vs. Split Sample [Effect size = .3]

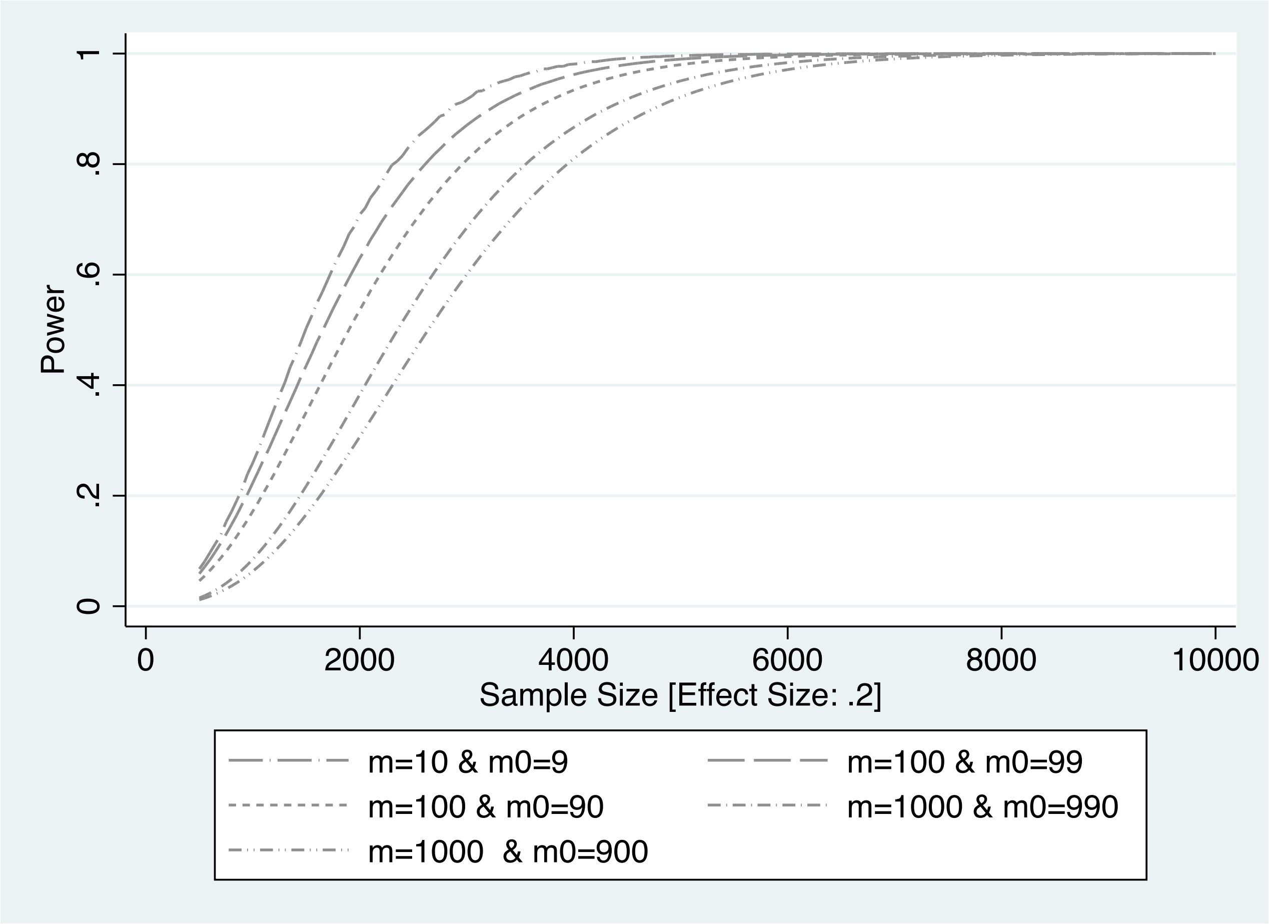


Figure 4: Power Under the Sample Split Approach with Bonferonni Corrections: Number of variables

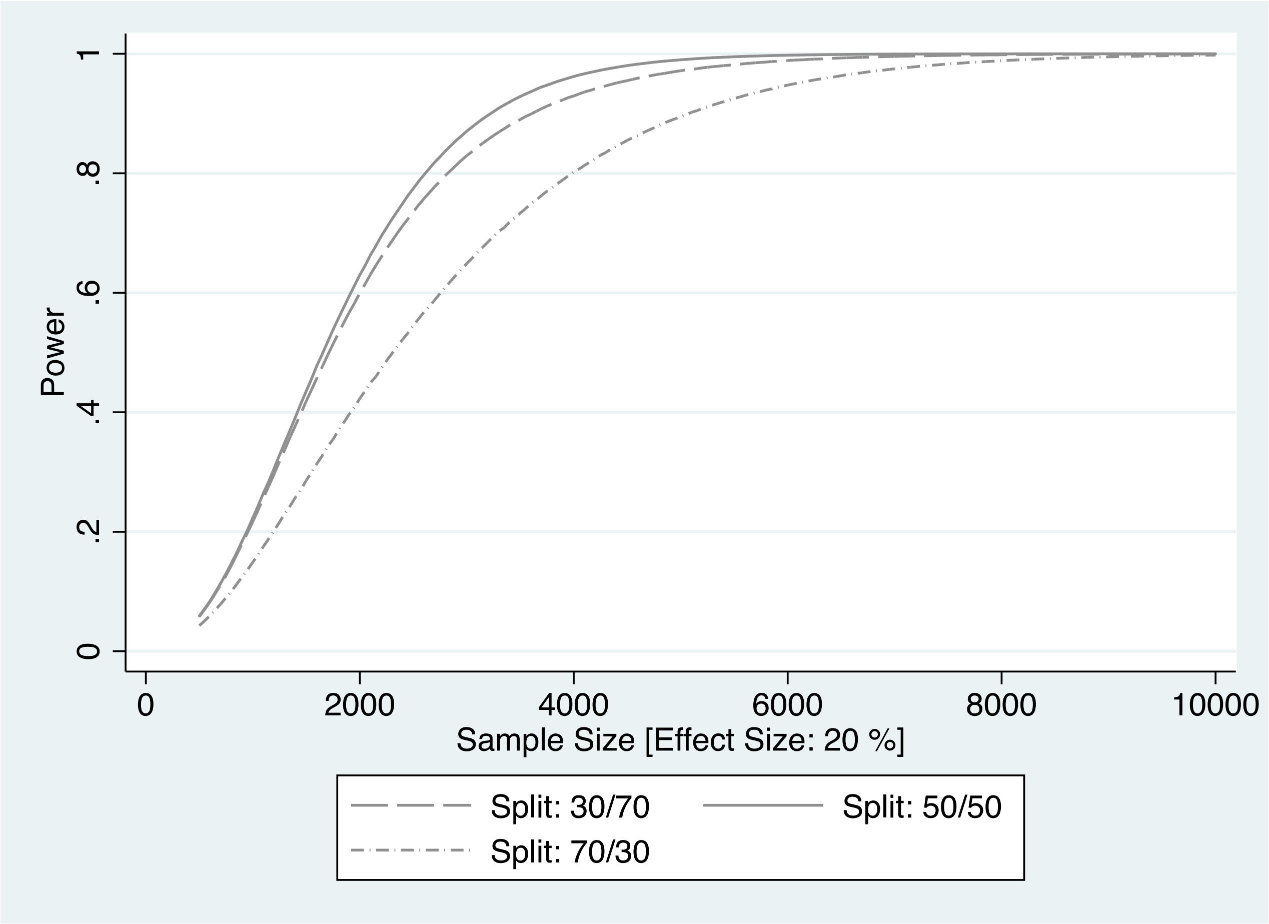


Figure 5: Power Under the Sample Split Approach with Bonferonni Corrections: Share in the Training Sample

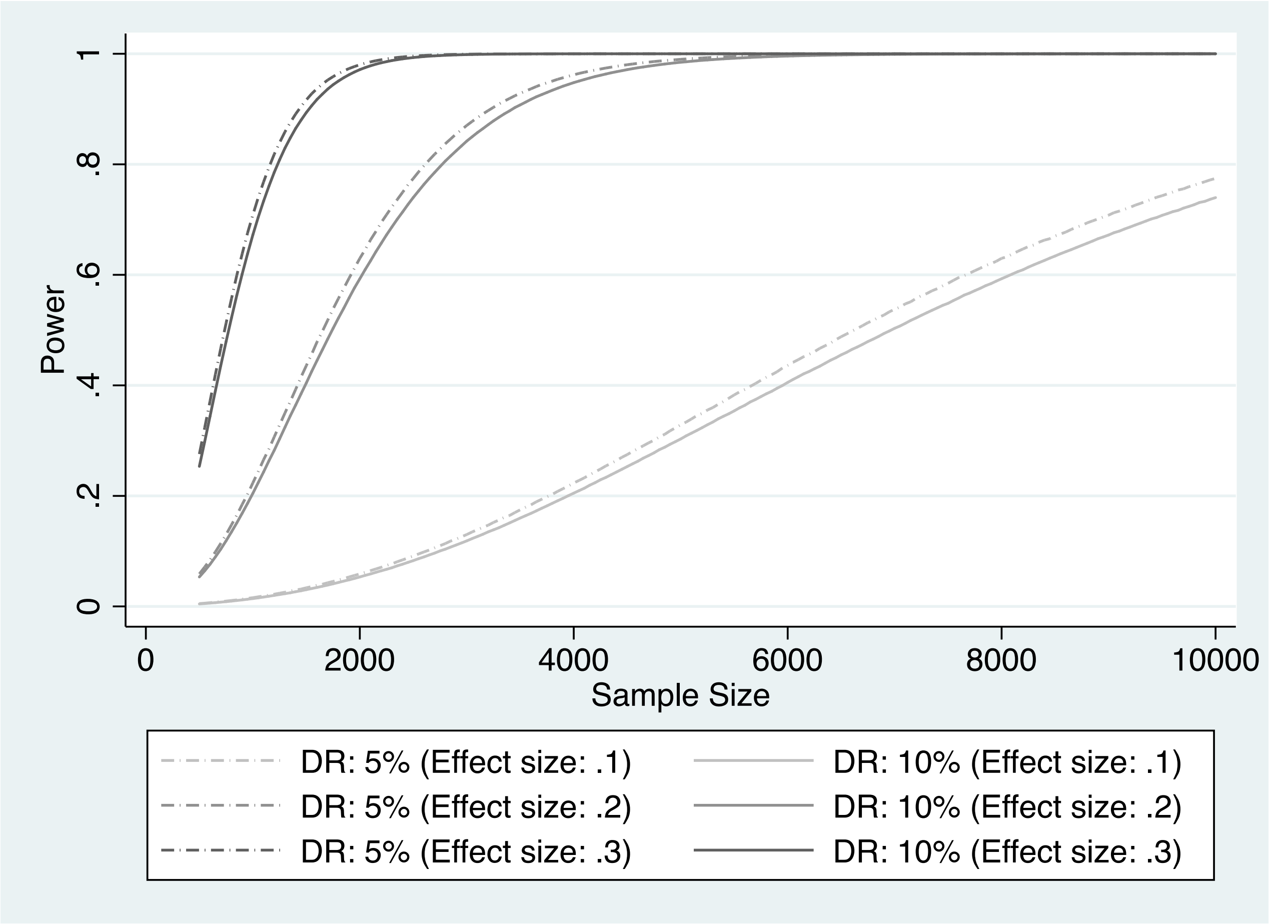
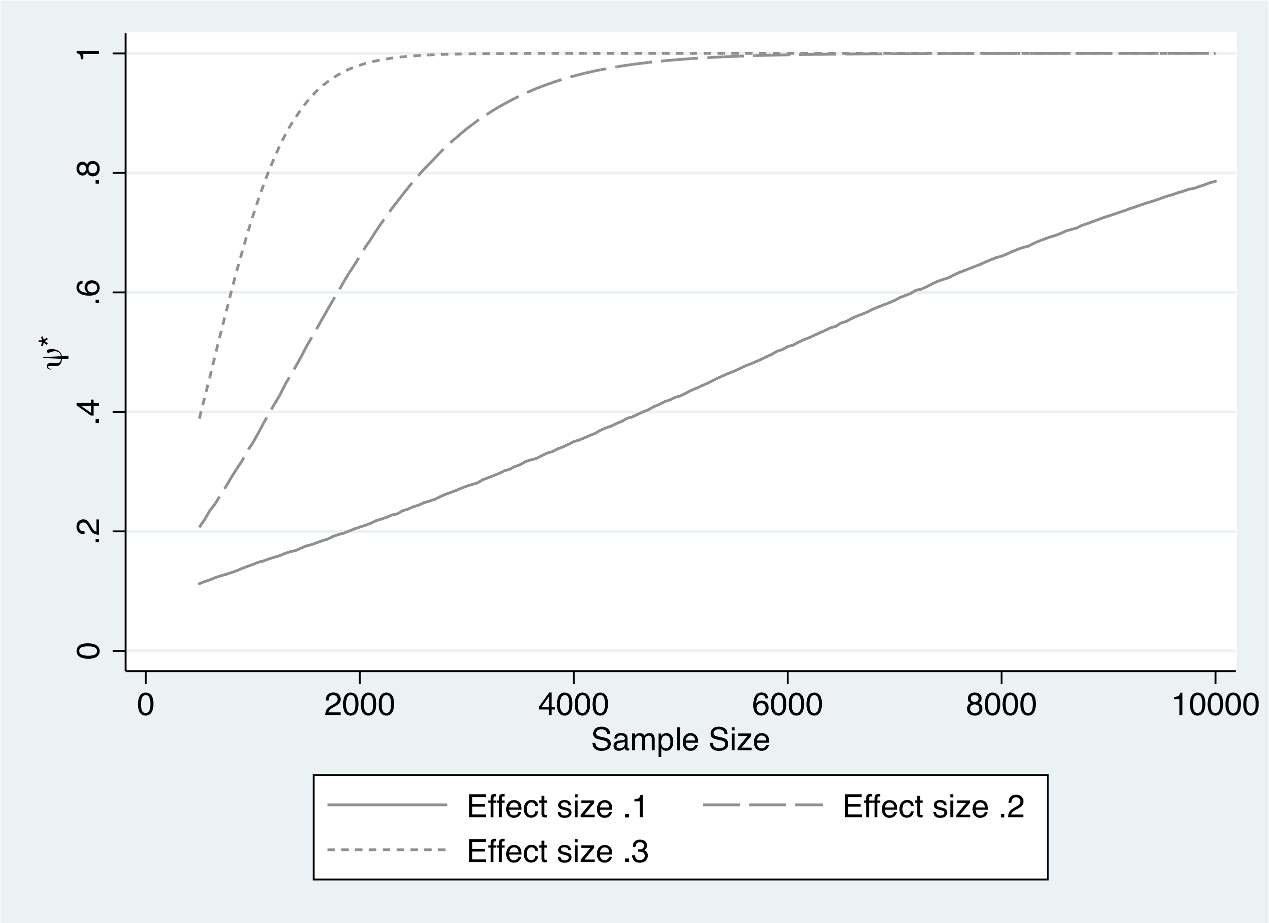


Figure 6: Power Under the Sample Split Approach with Bonferonni Corrections: Decision Rule on Training Sample

Figure 7: Value of ψ at which the full sample approach with a PAP and the split sample approach yields the same power.



Note: ψ is the likelihood that the non-true hypothesis is in the set of tests included in the PAP.

Appendix: Deriving the Power Calculations Formula

This appendix relies on Wittes (2002) and McConnell and Vera-Hernández (2015). Researcher A is interested in estimating the effect of T (the treatment randomly assigned to a subset of the sample) and she has access to sample S that includes a set of m potential outcome variables $(y^k)_{k=1,\ldots,m}$. The researcher decides to run a series of regressions:

$$y^k = a + b_k T + u \tag{10}$$

and carries out a series of tests: $H_0^k : b_k = 0$. The z-statistic associated with each test is given by:

$$Z^k = \frac{\bar{Y}_1^k - \bar{Y}_0^k}{\sigma_k \sqrt{1/n_0 + 1/n_1}} \tag{11}$$

Where $\bar{Y}_1^k$ ($\bar{Y}_0^k$) is the sample average of Y^k for observations with $T = 1$ ($T = 0$) and n_0 (n_1) is the number of observations with $T = 1$ ($T = 0$). Under H_0^k, $\bar{Y}_1^k = \bar{Y}_0^k$ and Z^k follows a normal distribution with mean zero and variance one.

The choice of α and β lead to the following set of equations:

$$Pr(|z| > Z_{1-\alpha/2}|H_0) < \alpha \tag{12}$$

$$Pr(|z| > Z_{1-\alpha/2}|H_A) > 1 - \beta \tag{13}$$

Assuming that, for non-true null hypotheses, the effect is δ_k and that $n_0 = n_1 = N/2$ leads to

$$Pr(\frac{\sqrt{N}|\bar{Y}_1^k - \bar{Y}_0^k|}{\sigma_k \sqrt{4}} > Z_{1-\alpha/2}|H_A) > 1 - \beta \tag{14}$$

Subtracting both sides by δ_k and dividing both sides by $\sigma_k \sqrt{4/N}$ leads to

$$Pr(\frac{\sqrt{N}(|\bar{Y}_1^k - \bar{Y}_0^k| - \delta_k)}{\sigma_k \sqrt{4}} > Z_{1-\alpha/2} - \frac{\sqrt{N}\delta_k}{\sigma_k \sqrt{4}}|H_A) > 1 - \beta \tag{15}$$

Given that under H_A, the expectation of $(\bar{Y}_1^k - \bar{Y}_0^k)$ is δ_k, $\frac{\sqrt{N}(|\bar{Y}_1^k - \bar{Y}_0^k| - \delta_k)}{\sigma_k \sqrt{4}}$ is normally distributed. It follows that

$$Z_{1-\alpha/2} - \frac{\sqrt{N}\delta_k}{\sigma_k \sqrt{4}} = Z_\beta = -Z_{1-\beta} \tag{16}$$

Rearranging the equation leads to:

$$Z_{1-\beta} = \delta_k \sqrt{\frac{N}{4\sigma_k^2}} - Z_{1-\alpha/2} \tag{17}$$

and so:

$$1 - \beta = \Phi(\delta_k \sqrt{\frac{N}{4\sigma_k^2}} - Z_{1-\alpha/2}) \tag{18}$$

If the researcher has access to m variables and plans to use Bonferonni corrections, power is:

$$1 - \beta^{Bonf} = \Phi(\delta_k \sqrt{\frac{N}{4\sigma_k^2}} - Z_{1-\alpha/(2m)}) \tag{19}$$