Queen's Economics Department Working Paper No. 1355

Randomization Inference for Difference-in-Differences with Few Treated Clusters

James G. MacKinnon
Queen's University

Matthew D. Webb
Carleton University

Department of Economics
Queen's University
94 University Avenue
Kingston, Ontario, Canada
K7L 3N6

6-2016

Randomization Inference for Difference-in-Differences with Few Treated Clusters *

James G. MacKinnon
Queen's University
jgm@econ.queensu.ca

Matthew D. Webb
Carleton University
matt.webb@carleton.ca

June 17, 2016

Abstract

Inference using difference-in-differences with clustered data requires care. Previous research has shown that, when there are few treated clusters, t tests based on a cluster-robust variance estimator (CRVE) severely over-reject, different variants of the wild cluster bootstrap can over-reject or under-reject dramatically, and procedures based on randomization inference show promise. We demonstrate that randomization inference (RI) procedures based on estimated coefficients, such as the one proposed by Conley and Taber (2011), fail whenever the treated clusters are atypical. We propose an RI procedure based on t statistics which fails only when the treated clusters are atypical and few in number. We also propose a bootstrap-based alternative to randomization inference, which mitigates the discrete nature of RI P values when the number of clusters is small. Two empirical examples demonstrate that alternative procedures can yield dramatically different inferences.

Keywords: CRVE, grouped data, clustered data, panel data, wild cluster bootstrap, difference-in-differences, DiD, randomization inference

*This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Taylor Jaworski for directing our attention to what became the empirical example. We would also like to thank Chris Taber and participants at the University of Calgary Empirical Microeconomics Conference (2015), the Canadian Econometric Study Group (2015), McMaster University, New York Camp Econometrics (2016), NYU-Shanghai, the CIREQ Econometrics Conference in Honor of Jean-Marie Dufour, University of Copenhagen, the Society of Labor Economists (2016), the Canadian Economics Association Meetings (2016), and Dalhousie University for helpful comments on preliminary versions.

1 Introduction

Inference for estimators that use clustered data, which in practice are very often difference-in-differences estimators, has received considerable attention in the past decade. Cameron and Miller (2015) provides a recent and comprehensive survey. While much progress has been made, there are still situations in which reliable inference is a challenge. It is particularly challenging when there are very few treated clusters. Past research, including Conley and Taber (2011), has shown that inference based on cluster-robust test statistics greatly over-rejects in this case. MacKinnon and Webb (2016) explains why this happens and why the wild cluster bootstrap of Cameron, Gelbach and Miller (2008) does not solve the problem. In fact, the wild cluster bootstrap either greatly under-rejects or greatly over-rejects, depending on whether or not the null hypothesis is imposed on the bootstrap DGP.

Several authors have considered randomization inference (RI) as a way to obtain tests with accurate size when there are few treated groups (Conley and Taber, 2011; Canay, Romano and Shaikh, 2014; Ferman and Pinto, 2015). RI procedures necessarily rely on strong assumptions about the comparability of the control groups to the treated groups. We show that, for the Conley-Taber procedure, these assumptions almost always fail to hold when the treated groups have either more or fewer observations than the control groups. As a consequence, the procedure can severely over-reject or under-reject if the treated groups are substantially smaller or larger than the controls.

We are motivated by the many studies that use individual data, in which there is variation in treatment across both groups and time periods. Such models are often expressed as follows. If i indexes individuals, g indexes groups, and t indexes time periods, then a classic "difference-in-differences" (or "DiD") regression can be written as

$$\begin{aligned} y_{igt} &= \beta_1 + \beta_2 \mathrm{GT}_{igt} + \beta_3 \mathrm{PT}_{igt} + \beta_4 \mathrm{GT}_{igt}\mathrm{PT}_{igt} + \epsilon_{igt}, \\ & i = 1, \ldots, N_g, \quad g = 1, \ldots, G, \quad t = 1, \ldots, T. \end{aligned} \tag{1}$$

Here GT_{igt} is a "group treated" dummy that equals 1 if group g is treated in any time period, and PT_{igt} is a "period treated" dummy that equals 1 if any group is treated in time period t. The coefficient of most interest is β_4, which shows the effect on treated groups in periods when there is treatment.[1] Following the literature, we divide the G groups into G_0 control groups, for which $\mathrm{GT}_{igt} = 0$, and G_1 treated groups, for which $\mathrm{GT}_{igt} = 1$, so that $G = G_0 + G_1$. We are concerned with the case in which G_1 is small.

Section 2 deals with a variety of procedures for inference with clustered errors. Subsection 2.1 discusses cluster-robust variance estimation and shows why it fails when there are few treated clusters. Subsection 2.2 briefly discusses some alternative procedures that we do not study. Subsection 2.3 explains how the wild cluster bootstrap works. Subsection 2.4 introduces randomization inference, Subsection 2.5 describes the Conley-Taber approach to RI, and Subsection 2.6 suggests an alternative RI procedure based on t statistics instead of coefficients. Subsection 2.7 examines some of the theoretical properties of these two RI procedures and shows that neither of them can be expected to perform well in certain cases.

[1] In many cases, of course, regression (1) would contain additional regressors, often including group and/or time dummies, which might make it necessary to drop GT_{igt}, PT_{igt}, or both.

In Section 3, we use Monte Carlo simulation experiments to compare several procedures. The model and DGP used in the experiments are described in Subsection 3.1, and the results are presented in Subsections 3.2, 3.3, and 3.4. We find, as the theory of Subsection 2.7 suggests, that none of the existing procedures yields reliable inferences when groups are heterogeneous and only one group is treated. However, the new RI procedure always outperforms the Conley-Taber procedure when two or more groups are treated.

Section 4 discusses a practical problem with all random inference procedures when the number of control groups is small. Subsection 4.1 then introduces a bootstrap-based modified RI procedure that solves this problem. In Subsection 4.2, we show that this procedure substantially improves inference in cases where the only problem is an insufficient number of control groups.

Section 5 presents results for two empirical examples, one based on Bailey (2010) and one based on Conley and Taber (2011). Section 6 concludes.

2 Inference with Few Treated Clusters

A linear regression model with clustered errors may be written as

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_G \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_G \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_G \end{bmatrix}, \tag{2}$$

where each of the G clusters, indexed by g, has N_g observations. The matrix $\boldsymbol{X}$ and the vectors $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ have $N = \sum_{g=1}^{G} N_g$ rows, $\boldsymbol{X}$ has k columns, and the parameter vector $\boldsymbol{\beta}$ has k rows. OLS estimation of equation (2) yields estimates $\hat{\boldsymbol{\beta}}$ and residuals $\hat{\boldsymbol{\epsilon}}$.

Because the elements of the $\boldsymbol{\epsilon}_g$ are in general neither independent nor identically distributed, both classical OLS and heteroskedasticity-robust standard errors for $\hat{\boldsymbol{\beta}}$ are invalid. As a result, conventional inference can be severely unreliable. It is therefore customary to use a cluster-robust variance estimator, or CRVE. There are several of these, of which the earliest may be the one proposed in Liang and Zeger (1986). The CRVE we investigate is defined as:

$$\frac{G(N-1)}{(G-1)(N-k)} (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \hat{\boldsymbol{\epsilon}}_g \hat{\boldsymbol{\epsilon}}_g' \boldsymbol{X}_g \right) (\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{3}$$

where $\hat{\boldsymbol{\epsilon}}_g$ is the subvector of $\hat{\boldsymbol{\epsilon}}$ that corresponds to cluster g. This is the estimator that is used when the `cluster` command is invoked in Stata. It yields reliable inferences when the number of clusters is large (Cameron, Gelbach and Miller, 2008) and the number of observations per cluster does not vary too much (Carter, Schnepel and Steigerwald, 2015; MacKinnon and Webb, 2016). However, Conley and Taber (2011) and MacKinnon and Webb (2016) show that t statistics based on (3) over-reject severely when the parameter of interest is the coefficient on a treatment dummy and there are very few treated clusters. Rejection frequencies can be over 80% when only one cluster is treated, even when the t statistics are assumed to follow a $t(G-1)$ distribution, as is now commonly done based on the results of Donald and Lang (2007) and Bester, Conley and Hansen (2011).

2.1 Cluster-Robust Variance Estimation

It is of interest to see precisely why inference based on the CRVE (3) fails so dramatically when there is just one treated cluster.[2] Consider, for simplicity, the dummy variable regression model

$$y_{ig} = \beta_1 + \beta_2 d_{ig} + \epsilon_{ig}, \tag{4}$$

where the treatment dummy d_{ig} equals 1 for the first G_1 clusters and 0 for the remaining G_0 clusters. Making equation (4) more complicated by adding additional regressors, or by allowing only some observations within the treated clusters to be treated, as in the DiD regression (1), would not change anything important. The fundamental problem for all these models, as we will see shortly, is that the residuals sum to zero over all treated observations.

Equation (4) may be rewritten in vector notation as $\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \beta_2 \boldsymbol{d} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$, $\boldsymbol{\iota}$, $\boldsymbol{d}$, and $\boldsymbol{\epsilon}$ are N-vectors with typical elements y_{ig}, 1, d_{ig}, and ϵ_{ig}, respectively, and i is assumed to vary more rapidly than g. Then the OLS estimate of β_2 is

$$\hat{\beta}_2 = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{y}}{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})} = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\epsilon}}{N(\bar{d} - \bar{d}^2)}, \tag{5}$$

where the second equality holds under the null hypothesis that $\beta_2 = 0$, and $\bar{d} = \left(\sum_{g=1}^{G_1} N_g\right)/N$ is the sample mean of the d_{ig}, that is, the proportion of treated observations. The variance of $\hat{\beta}_2$ is evidently

$$\mathrm{Var}(\hat{\beta}_2) = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\Omega}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})}{\left((\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})\right)^2} = \frac{(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})'\boldsymbol{\Omega}(\boldsymbol{d} - \bar{d}\boldsymbol{\iota})}{N^2\bar{d}^2(1-\bar{d})^2}, \tag{6}$$

where $\boldsymbol{\Omega}$ is an $N \times N$ block diagonal matrix with G covariance matrices $\boldsymbol{\Omega}_g$ of dimensions $N_g \times N_g$ forming the diagonal blocks.

From expression (3), it is easy to see that the CRVE for $\hat{\beta}_2$ is proportional to

$$\frac{1}{N^2\bar{d}^2(1-\bar{d})^2} \sum_{g=1}^{G} (\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g)'\hat{\boldsymbol{\epsilon}}_g\hat{\boldsymbol{\epsilon}}_g'(\boldsymbol{d}_g - \bar{d}\boldsymbol{\iota}_g), \tag{7}$$

where $\boldsymbol{d}_g$ is the subvector of $\boldsymbol{d}$ that corresponds to cluster g, and $\boldsymbol{\iota}_g$ is an N_g-vector of 1s. Thus expression (7) should provide a good estimate of $\mathrm{Var}(\hat{\beta}_2)$ if the summation provides a good estimate of the quadratic form in (6). Unfortunately, this is not the case when the number of treated clusters is small.

It is not difficult to show that the summation in expression (7) is equal to

$$(1-\bar{d})^2 \sum_{g=1}^{G_1} (\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 + \bar{d}^2 \sum_{g=G_1+1}^{G} (\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2. \tag{8}$$

This expression is supposed to estimate the quadratic form in expression (6), which can be written as

$$(1-\bar{d})^2 \sum_{g=1}^{G_1} \boldsymbol{\iota}_g'\boldsymbol{\Omega}_g\boldsymbol{\iota}_g + \bar{d}^2 \sum_{g=G_1+1}^{G} \boldsymbol{\iota}_g'\boldsymbol{\Omega}_g\boldsymbol{\iota}_g. \tag{9}$$

[2] This subsection is based on parts of Section 6 of MacKinnon and Webb (2016).

Unfortunately, expression (8) estimates expression (9) very badly when G_1 is small. Consider the extreme case in which $G_1 = 1$. Since $d_{i1} = 1$ and $d_{ig} = 0$ for $g > 1$, the residuals for cluster 1 must sum to zero. This implies that the first term in expression (8) equals zero. In contrast, the first term in expression (9) is not zero, and it will generally be quite large relative to the second term, because $\bar{d}$ will normally be small if just one cluster is treated. In fact, if we let $G \to \infty$ while keeping $G_1 = 1$, it must be the case that $\bar{d} \to 0$. Evidently (8) will severely underestimate (9) when $G_1 = 1$.

When two or more clusters are treated, the residuals for those clusters will not sum to zero for each cluster, but they must sum to zero over all the treated clusters. In consequence, the expectation of the squared summation for the first treated cluster must underestimate the corresponding true variance. When the errors are homoskedastic and independent, it does so by a factor of $(M_1 - N_1)/M_1$, where N_1 is the number of treated observations in cluster 1, and M_1 is the number of treated observations in all treated clusters.[3] Although this result is somewhat special, it strongly suggests that, for G_1 small, the sum of squared summations in the first term of (8) will severely underestimate the corresponding double summation in (9). The problem evidently goes away as G_1 increases, provided the sizes of the treated clusters are not too variable, and simulation results in MacKinnon and Webb (2016) suggest that it does so quite quickly.

2.2 Other Procedures

Building off results in Donald and Lang (2007), Ibragimov and Müller (2016) studies the Behrens-Fisher problem of comparing means of two groups with different variances. The paper focuses on differences in means for treated and control groups and proves that t tests for these differences in means follow asymptotic distributions with degrees of freedom equal to $\min(G_0, G_1) - 1$. When $G_1 = 1$, this number is 0, which implies that the Ibragimov-Müller procedure is not appropriate when there is only one treated group.

A very different procedure is proposed by Abadie, Diamond and Hainmueller (2010). It also bases inference on an empirical distribution generated by perturbing the assignment of treatment. However, the procedure differs substantially from the ones considered in this paper, because it constructs a "synthetic control" as a weighted average of potential control groups, based on the characteristics of the explanatory variables. This results in both a different estimate of the "treatment effect" and a different P value. For this reason, we do not study the synthetic controls approach in this paper.

Recent work by Carter, Schnepel and Steigerwald (2015) develops the asymptotic properties of the CRVE when the number of observations per cluster is not constant. The authors show that, when clusters are unbalanced, the dataset has an effective number of clusters, G^*, which is less than G (sometimes very much less). Simulations in MacKinnon and Webb (2016) show that using critical values based on G^* can work fairly well when intermediate numbers of clusters are treated. However, when few clusters are treated (or untreated), it can either over-reject or under-reject severely. When G^* is extremely small, which happens whenever G_1 is small, sharply different results can be obtained depending on whether the test statistic is assumed to follow a $t(G^*)$ or a $t(G^* - 1)$ distribution. For this reason, we do not consider these procedures in our simulation experiments. Alternative degrees-of-freedom

[3] This result is equation (A.2) of the online appendix of MacKinnon and Webb (2016).

corrections, in some cases based on alternative CRVEs, have also been proposed by Imbens and Kolesar (2016) and Young (2015b). All of these procedures are computationally challenging, and in some cases infeasible, for datasets with large clusters. We therefore do not study them in our experiments.

2.3 The Wild Cluster Bootstrap

The wild cluster bootstrap was proposed in Cameron, Gelbach and Miller (2008) as a method for reliable inference in cases with a small number of clusters.[4] It was studied extensively in MacKinnon and Webb (2016) for the cases of unbalanced clusters and/or few treated clusters. Because we will be proposing a new procedure that is closely related to the wild cluster bootstrap in Subsection 4.1, we review how the latter works.

Suppose we wish to test the hypothesis that a single coefficient in equation (2) is zero. Without loss of generality, we let this be β_k, the last coefficient of $\boldsymbol{\beta}$. The restricted wild cluster bootstrap works as follows:

1. Estimate equation (2) by OLS.

2. Calculate $\hat{t}_k$, the t statistic for $\beta_k = 0$, using the square root of the k^{th} diagonal element of (3) as a cluster-robust standard error.

3. Re-estimate the model (2) subject to the restriction that $\beta_k = 0$, so as to obtain the restricted residuals $\tilde{\boldsymbol{\epsilon}}$ and the restricted estimates $\tilde{\boldsymbol{\beta}}$.

4. For each of B bootstrap replications, indexed by b, generate a new set of bootstrap dependent variables y_{ig}^{*b} using the bootstrap DGP

$$y_{ig}^{*b} = \boldsymbol{X}_{ig}\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}_{ig} v_g^{*b}, \tag{10}$$

   where y_{ig}^{*b} is an element of the vector $\boldsymbol{y}^{*b}$ of observations on the bootstrap dependent variable, $\boldsymbol{X}_{ig}$ is the corresponding row of $\boldsymbol{X}$, and so on. Here v_g^{*b} is a random variable that follows the Rademacher distribution; see Davidson and Flachaire (2008). It takes the values 1 and -1 with equal probability. Observe that v_g^{*b} takes the same value for all observations within each group. Because of that, we would not want to use the Rademacher distribution if G were smaller than about 12; see Webb (2014), which proposes an alternative for such cases.

5. For each bootstrap replication, estimate regression (2) using $\boldsymbol{y}^{*b}$ as the regressand, and calculate t_k^{*b}, the bootstrap t statistic for $\beta_k = 0$, using the square root of the k^{th} diagonal element of (3), with bootstrap residuals replacing the OLS residuals, as the standard error.

6. Calculate the bootstrap P value as

$$\hat{p}_{\text{s}}^{*} = \frac{1}{B}\sum_{b=1}^{B} \text{I}\big(|t_k^{*b}| > |\hat{t}_k|\big), \tag{11}$$

[4] A different, but much less effective, bootstrap procedure for cluster-robust inference was previously suggested by Bertrand, Duflo and Mullainathan (2004).

where $I(\cdot)$ denotes the indicator function. Equation (11) assumes that the distribution of t_k is symmetric. Alternatively, one can use a slightly more complicated formula to calculate an equal-tail bootstrap P value.

The procedure just described is known as the restricted wild cluster, or WCR, bootstrap, because the bootstrap DGP (10) uses restricted parameter estimates and restricted residuals. The unrestricted wild cluster, or WCU, bootstrap is a closely related procedure. It uses unrestricted estimates and unrestricted residuals in step 4, and the bootstrap t statistics in step 5 now test the hypothesis that $\beta_k = \hat{\beta}_k$.

MacKinnon and Webb (2016) explains why the wild cluster bootstrap fails when the number of treated clusters is small. The WCR bootstrap, which imposes the null hypothesis, leads to severe under-rejection. In contrast, the WCU bootstrap, which does not impose the null hypothesis, leads to severe over-rejection. When just one cluster is treated, it over-rejects at almost the same rate as using CRVE t statistics with the $t(G-1)$ distribution. In many of the Monte Carlo experiments discussed below, we obtained results for the WCR and WCU bootstraps, but we do not report most of them because those procedures are treated in detail in MacKinnon and Webb (2016).

2.4 Randomization Inference

Randomization inference was first proposed by Fisher (1935) as a procedure for performing exact tests in the context of experiments. Rosenbaum (1996) mentions the possibility of using randomization inference for group level interventions.[5] The idea is to compare an observed test statistic $\hat{T}$ with an empirical distribution of test statistics T_j^* for $j = 1, \ldots, S$ generated by re-randomizing the assignment of treatment across experimental units. To compute each of the T_j^*, we use the actual outcomes while pretending that certain non-treated experimental units were treated. If $\hat{T}$ is in the tails of the empirical distribution of the T_j^*, then this is evidence against the null hypothesis of no treatment effect.

The general practice is to incorporate all available information about treatment assignment in conducting the re-randomization (Yates, 1984). This is done because randomization tests are valid only when the distribution of the test statistic is invariant to the realization of the re-randomizations across permutations of assigned treatments (Lehmann and Romano, 2008). In practice, this means making use of any information that determines treatment assignment in the original data.

When treatment is randomly assigned at the individual level, the invariance of the distribution of the test statistic to re-randomization will follow naturally. However, if treatment assignment is instead at the group level, then the extent of unbalancedness can determine how close the distribution is to being invariant. When clusters are balanced, the value of $\bar{d}$ in equation (8) will be constant across re-randomizations. However, when clusters are unbalanced, $\bar{d}$ may vary considerably across re-randomizations. The implications of this are discussed below in Subsection 2.7.

2.5 Randomization Inference – Coefficients

Conley and Taber (2011) suggests two procedures for inference with few treated groups. Both of these procedures involve constructing an empirical distribution by randomizing the

[5] Monte Carlo tests are closely related to randomization inference; see Dufour (2006).

assignment of groups to "treatment" and "control" and using this empirical distribution to conduct inference. These procedures are quite involved, and the details can be found in their paper. Of the two procedures, we restrict attention to the one based on randomization inference, because it can be used whether or not $G_0 > G_1$ and because it often has better size properties in their Monte Carlo experiments.

For the RI procedure of Conley and Taber (2011), a coefficient equivalent to β_4 in equation (1) is estimated, and the estimate, say $\hat{\beta}$, is compared to an empirical distribution of estimated β_j^*, where j indexes repetitions. The β_j^* are obtained by pretending that various sets of G_1 groups are actually treated. When $G_1 = 1$, the number of β_j^* is just G_0. For $G_1 > 1$, the number is generally much larger, as we discuss in the next subsection.

This procedure evidently depends on the strong assumption that $\hat{\beta}$ and the β_j^* follow the same distribution. But that cannot be the case if the coefficients for some clusters are estimated more efficiently than for others, perhaps because some clusters have more observations. As we will demonstrate in Subsection 3.2, when clusters are of different sizes, and unusually large or small clusters are treated, this type of RI procedure can have very poor size properties. A similar problem arises whenever the variance of the error terms for the treated clusters differs from the variance of the error terms for the controls; see Subsection 3.4.

2.6 Randomization Inference – t statistics

As an alternative to the Conley-Taber procedure, we consider an RI procedure based on cluster-robust t statistics. Instead of comparing $\hat{\beta}$ to the empirical distribution of the β_j^*, we compare the actual t statistic $\hat{t}$ to an empirical distribution of the t_j^* that correspond to the β_j^*. This is similar to one of the procedures studied in Young (2015a).

When there is just one treated group, it is natural to compare $\hat{t}$ to the empirical distribution of G_0 different t_j^* statistics. However, when there are two or more treated groups and G_0 is not quite small, the number of potential t_j^* to compare with can be very large. In such cases, we may pick B of them at random. Note that, to avoid ties, we never include the actual $\hat{t}$ among the t_j^*.

Our randomization inference procedure works as follows:

1. Estimate the regression model and calculate $\hat{t}$, the cluster-robust t statistic for the coefficient of interest. For the DiD model (1), this is the t statistic for $\beta_4 = 0$.

2. Generate a number of t_j^* statistics to compare $\hat{t}$ with.

   - When $G_1 = 1$, assign a group from the G_0 control groups as the "treated" group g^* for each repetition, re-estimate the model using the observations from all G groups, and calculate a new t statistic, t_j^*, indicating randomized treatment. Repeat this process for all G_0 control groups. Thus the empirical distribution of the t_j^* will have G_0 elements.
   - When $G_1 > 1$, sequentially treat every set of G_1 groups except the set actually treated, re-estimate equation (1), and calculate a new t_j^*. There are potentially ${}_G\mathrm{C}_{G_1} - 1$ sets of groups to compare with, where ${}_n\mathrm{C}_k$ denotes "n choose k." When this number is not too large, obtain all of the t_j^* by enumeration. When it exceeds

B (picked on the basis of computational cost), choose the comparators randomly, without replacement, from the set of potential comparators. Thus the empirical distribution will have $\min({}_G\mathrm{C}_{G_1} - 1, B)$ elements.

3. Sort the vector of t_j^* statistics.

4. Determine the location of $\hat{t}$ within the sorted vector of the t_j^*, and compute a P value. This may be done in more than one way; see Subsection 4.

We will refer to the procedure just described as "t statistic randomization inference," or RI-t for short, and to the Conley-Taber procedure described in the preceding subsection as "coefficient randomization inference," or RI-β for short.

2.7 Properties of Randomization Inference Procedures

It seems plausible that randomization inference should perform better when it is based on t statistics than when it is based on coefficients, because t statistics are asymptotically pivotal (that is, invariant to any unknown parameters) and coefficients are not. Thus it is less unrealistic to assume that $\hat{t}$ and the t_j^* follow the same distribution than it is to assume that $\hat{\beta}$ and the β_j^* do so. Unfortunately, as we now demonstrate, the two RI procedures do not differ much when G_1 is very small. As a result, there are many circumstances in which they are both likely to yield similar inferences, which can be grossly invalid.

Consider again the pure treatment model (4). Under the null hypothesis, the parameter estimate $\hat{\beta}_2$ is given in equation (5) and can be rewritten as

$$\hat{\beta}_2 = \frac{1}{N\bar{d}(1-\bar{d})}\left((1-\bar{d})\sum_{g=1}^{G_1}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g - \bar{d}\sum_{g=G_1+1}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g\right). \tag{12}$$

Combining this result with (7) and (8), we find that the t statistic is

$$\hat{t}_2 = \frac{(1-\bar{d})\sum_{g=1}^{G_1}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g - \bar{d}\sum_{g=G_1+1}^{G}\boldsymbol{\iota}_g'\boldsymbol{\epsilon}_g}{\left((1-\bar{d})^2\sum_{g=1}^{G_1}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2 + \bar{d}^2\sum_{g=G_1+1}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2\right)^{1/2}}. \tag{13}$$

Note that the numerator depends on the error terms $\boldsymbol{\epsilon}_g$, and the denominator depends on the residuals $\hat{\boldsymbol{\epsilon}}_g$.

Now suppose that $G_1 = 1$. In that case, as we saw in Subsection 2.1, the first term in the denominator of (13) equals zero. Consider the two terms in parentheses in equation (12), which also form the numerator of $\hat{t}_2$. If we further assume that G_0 is large relative to G_1 and that N_1 is not unusually large, the first of these two terms must be much larger than the second. Therefore, it must be approximately the case that

$$\hat{\beta}_2 \approx \frac{1}{N\bar{d}}\boldsymbol{\iota}_1'\boldsymbol{\epsilon}_1 = O_p(N_1^{-1/2}), \tag{14}$$

where the order relation uses the facts that $\bar{d} = O(N_1)/O(N)$ and $\boldsymbol{\iota}_1'\boldsymbol{\epsilon}_1 = \sum_{i=1}^{N_1}\epsilon_{i1} = O_p(N_1^{1/2})$. The result (14) is intuitive: The larger the number of treated observations, the less dispersed will be the estimate of β.

Similarly,

$$\hat{t}_2 \approx \frac{(1-\bar{d})\boldsymbol{\iota}_1'\boldsymbol{\epsilon}_1}{\bar{d}\left(\sum_{g=2}^{G}(\boldsymbol{\iota}_g'\hat{\boldsymbol{\epsilon}}_g)^2\right)^{1/2}} = O_p(N_1^{-1/2})O_p(N^{1/2}), \tag{15}$$

where the order relation uses the same facts. For simplicity, we have also assumed that $G = O(N)$, so that the square root in the denominator is $O_p(N^{1/2})$. Since the rightmost factor on the right-hand side of equation (15) does not depend on which cluster is being treated, this assumption does not matter for RI inference, and that factor can be ignored. The important thing is that the absolute values of both $\hat{t}_2$ and $\hat{\beta}_2$ are $O_p(N_1^{-1/2})$. They can both therefore be expected to shrink as the number of treated observations increases.

Equations (14) and (15) make it evident that both RI procedures will tend to over-reject when N_1 is small and under-reject when N_1 is large. In the former case, both $\hat{\beta}_2$ and $\hat{t}_2$ will tend to be more variable than the β_j^* and t_j^* with which they are being compared, because $N_1^{-1/2}$ is larger than $N_j^{*-1/2}$ for most of the other clusters. In the latter case, by the same argument in reverse, both $\hat{\beta}_2$ and $\hat{t}_2$ will tend to be less variable than the β_j^* and t_j^* with which they are being compared. Thus neither RI procedure can possibly provide valid inferences when $G_1 = 1$ and the treated cluster is larger or smaller than the controls.

The case of $G_1 = 1$ is the most extreme one. As G_1 increases, we would expect the distribution of $\hat{t}$ eventually to lose any dependence on the sizes of the treated clusters, because the first term in the denominator of (13) will no longer be zero, and $\bar{d}$ will increase with G_1. In contrast, the distribution of $\hat{\beta}_2$ will continue to depend on the sizes of the treated clusters. Thus we would expect the behavior of the two RI procedures to become less and less similar as G_1 increases in cases with unbalanced clusters where neither of them yields valid inferences when $G_1 = 1$.

The failure of both RI-β and RI-t when G_1 is small and cluster sizes vary, and of the former even when G_1 is not small, is a consequence of the fact that $\hat{\beta}_2$ and $\hat{t}_2$ depend on $\bar{d}$, which is not invariant across re-randomizations. As such, it is not surprising that the randomization inference procedures fail with unbalanced clusters, as the simulation results in Subsections 3.2 and 3.3 will demonstrate.

The RI-β procedure was originally suggested for use with aggregate data, or with individual data that have been aggregated into time-cluster cells. It is probably less unreasonable to expect $\hat{\beta}_2$ and the β_j^* to follow the same distribution in those cases than in the case of individual data. Nevertheless, the assumption that $\hat{\beta}$ and the β_j^* follow the same distribution is still a very strong one. Variations across clusters in the number of underlying observations per cell, in the values of other regressors, or in the variances of the error terms may all invalidate this crucial assumption.[6] In contrast, $\hat{t}$ and the t_j^* can be expected to follow approximately the same distribution whenever G_1 and G are not too small.

3 Simulation Experiments

We conduct a number of Monte Carlo experiments to study the performance of various inferential procedures when the number of treated clusters is small and cluster sizes are

[6] Ferman and Pinto (2015) show that aggregation of unbalanced clusters introduces heteroskedasticity in the aggregate data, which causes similar problems for randomization inference when either large or small clusters are treated.

heterogeneous. In the experiments, we vary the total number of clusters, the number of treated clusters, and which clusters are treated. Since sample size does not seem to matter, we either hold it fixed or let it vary with the number of clusters.

3.1 Monte Carlo Design

In all experiments, we assign N total observations unevenly among G clusters using the following formula:

$$N_g = \left[N \frac{\exp(\gamma g/G)}{\sum_{j=1}^{G} \exp(\gamma j/G)} \right], \quad g = 1, \ldots, G-1, \tag{16}$$

where $[x]$ means the integer part of x. The value of N_G is then set to $N - \sum_{g=1}^{G_1} N_g$. The key parameter here is $\gamma \geq 0$, which determines how uneven the cluster sizes are. When $\gamma = 0$ and N/G is an integer, equation (16) implies that $N_g = N/G$ for all g. As γ increases, however, cluster sizes vary more and more.[7]

The data generating process is the DiD model, equation (1), with $\beta_4 = 0$. Each observation is assigned to one of 20 "years", and the starting year of "treatment" is randomly assigned to years between 4 and 14. The error terms are homoskedastic and correlated within each cluster, with correlation coefficient 0.05. The number of Monte Carlo replications in all experiments is 100,000. Rejection frequencies are calculated at the 1%, 5%, and 10% levels, although only the 5% rejection frequencies are discussed below.

3.2 Monte Carlo Results for RI Procedures

In the first set of experiments, $N = 4000$, $G = 40$, and the number of treated clusters varies from $1, 2, ..., 10$. The treated clusters are chosen in three ways: the smallest first, the largest first, or at random. The more observations the treated clusters have (at least up to about half the sample size), the more efficiently β_4 should be estimated.[8] Thus, as discussed in Subsection 2.7, we would expect randomization inference based on coefficient estimates to perform less well than randomization inference based on t statistics when the treated clusters are unusually large or small.

For the two varieties of randomization inference, the number of randomizations is as follows: 39 for $G_1 = 1$; ${}_{40}C_2 - 1 = 779$ for $G_1 = 2$; and 999 for $G_1 \geq 3$. We set $G = 40$ to avoid the problem of interval P values, which is discussed in Section 4.

The most striking result is that the size of some tests depends heavily on which clusters are treated. For instance, with RI-β, there is severe under-rejection when the largest clusters are treated first and fairly severe over-rejection when the smallest clusters are treated first. As the analysis in Subsection 2.7 predicts, RI-t also performs poorly when $G_1 = 1$, with the same pattern of under-rejection and over-rejection as RI-β.

These patterns can be seen clearly in Figure 1, which graphs rejection frequencies against G_1 for both RI-β and RI-t tests. For RI-β, not much changes as G_1 increases. In contrast, for

[7] For the experiments with 4000 observations and $\gamma = 2$, the sizes of the 40 clusters are: 32, 33, 35, 37, 39, 41, 43, 45, 47, 50, 52, 55, 58, 61, 64, 67, 71, 75, 78, 82, 87, 91, 96, 101, 106, 112, 117, 123, 130, 136, 143, 151, 158, 167, 175, 184, 194, 204, 214, and 246.

[8] It is difficult to be precise about this, because efficiency will also depend on intra-cluster correlations, the number of treated years, and the values of other regressors.

RI-t, the rejection frequencies improve rapidly as G_1 increases. When the smallest clusters are treated, the procedure seems to work perfectly for $G_1 \geq 6$. When the largest clusters are treated, it always under-rejects, but not severely.

One other result that is evident in Figure 1 is that both RI procedures work extremely well when the treated clusters are chosen at random. That makes sense, because the theory of randomization inference is based on treatment being assigned at random. In the context of DiD inference with non-experimental data, however, this result is actually quite misleading. In this case, clusters are not treated at random, and the investigator knows which clusters were actually treated.

The distribution of cluster sizes is the same for all the experiments. The good results for random treatment arise because we are averaging over 100,000 replications. In some of those replications, when small clusters happen to be treated, too many rejections occur. In others, when large clusters happen to be treated, too few rejections occur. Only when clusters of intermediate size (or an appropriate mix of small and large clusters) are treated do both RI procedures actually work well before averaging.

Figure 2 illustrates this point. The figure shows rejection frequencies for the RI-t procedure when $G_1 = 1$ and $G = 40$.[9] The horizontal axis shows the rank of the treated cluster, ordered from smallest to largest. There are five curves, which correspond to five values of γ. Each point on the curve represents a rejection frequency for a different treated cluster. The higher the value of γ, the more variable are the cluster sizes. As expected, the tests over-reject when the treated cluster is small and under-reject when it is large. Both over-rejection and under-rejection become more severe as γ increases.

Because the relative sizes of treated and control clusters evidently matter greatly when the smallest clusters are treated, we perform a second set of experiments in which $N = 5000$, $G = 20$, and the parameter γ in equation (16) is varied between 0 and 4 at intervals of 0.5. This is done for three values of G_1 (1, 2, and 3) for both RI procedures. As can be seen in Figure 3, they both work very well when $\gamma = 0$, and they both over-reject to a greater and greater extent as γ increases. However, the RI-t procedure over-rejects less and less severely as G_1 increases from 1 to 2 to 3, while the RI-β procedure over-rejects much more severely for $G_1 = 2$ and $G_1 = 3$ than for $G_1 = 1$. This is consistent with the theoretical results in Subsection 2.7.

It is of interest to compare the performance of the RI-β and RI-t procedures with that of the wild cluster bootstrap. Figure 4 shows rejection frequencies for the restricted wild cluster bootstrap for the same experiments as Figure 1.[10] As the theoretical and simulation results in MacKinnon and Webb (2016) suggest, the WCR bootstrap severely under-rejects for very small values of G_1, but it performs very well for $G_1 \geq 6$. Unlike the two RI procedures, its performance is very similar in all three cases, although the region of severe under-rejection is smallest when the treated clusters are the largest ones.

[9] Results for the RI-β procedure are not reported because, as the theory of Subsection 2.7 suggests, they are very similar to the ones in Figure 2.

[10] The figure does not show results for the WCU bootstrap, because they are so different from the ones for the WCR bootstrap. When $G_1 = 1$, the rejection frequencies are 0.615, 0.758, and 0.861 for the largest-first, random, and smallest-first cases.

3.3 Why the RI Procedures Can Fail

Figure 5 provides more intuition about why both RI procedures can fail with unbalanced clusters. The figure plots the distributions of $\hat{\beta}$ and of the corresponding t statistic for $G_1 = 1$ and $G_1 = 2$ with $N = 2000$, $G = 40$, and $\gamma = 2$ for a pure treatment model. The smallest cluster has 16 observations, and the largest has 134. Within each panel, distributions are plotted for three cases: the case in which G_1 randomly-chosen clusters are treated, the case in which the treated clusters are chosen from the smallest 10 clusters, and the case in which the treated clusters are chosen from the largest 10 clusters. The first case corresponds to the unconditional distribution of either $\hat{\beta}$ or $\hat{t}$, and the other two cases correspond to distributions conditional on the treated clusters being either small or large.

Recall that, for randomization inference to be valid, the distribution of the test statistic must be invariant to randomization. An implication of this is that the two conditional distributions should be indistinguishable from the unconditional distribution. For $G_1 = 1$, however, both conditional distributions differ greatly from the unconditional distribution for both $\hat{\beta}$ and $\hat{t}$. For $G_1 = 2$, the conditional distributions of $\hat{\beta}$ differ just as greatly from the unconditional one. However, the conditional distributions of $\hat{t}$ are much closer to the unconditional distribution, although they are still distinct. We also obtained results, not shown, for $G_1 = 3, 4, \ldots, 8$. As G_1 increases, the conditional distributions of $\hat{\beta}$ never converge to the unconditional distribution, while those of $\hat{t}$ do converge quite rapidly.

Since researchers always know the sizes, and usually the identities, of the clusters that are treated, it generally makes no sense to pretend that the treated clusters are chosen at random. Failing to condition on what the researcher knows about the treated and control clusters inevitably results in unreliable inference, especially for RI-β. Ferman and Pinto (2015) eloquently makes this point in the context of aggregate data.

3.4 RI Procedures with Heteroskedasticity

In the experiments reported so far, the distributions of the coefficients can differ across clusters only because cluster sizes may vary. However, that is not the only possible reason for those distributions to differ. Another possibility is that the error terms for the treated clusters may have larger or smaller variances than those of the controls. To investigate this possibility, we performed an additional set of experiments in which the standard error for the treated clusters was λ times the standard error for the controls. We would expect over-rejection when $\lambda > 1$ and under-rejection when $\lambda < 1$. This is easiest to see for the extreme case in which $G_1 = 1$. From equations (14) and (15), it is evident that the larger the variance of the error terms for the treated cluster, the larger will be the variances of $\hat{\beta}_2$ and $\hat{t}_2$ for the treated cluster when $G_1 = 1$. As G_1 increases, the problem should go away for RI-t but not for RI-β.

Figure 6 shows rejection frequencies for the RI-β and RI-t procedures for a DiD model with 40 equal-sized clusters and 800 observations. Results are shown for three values of λ, namely, $\lambda = 2.0$, $\lambda = 1.25$, and $\lambda = 0.5$. As expected, both procedures over-reject when $\lambda > 1$ and under-reject when $\lambda < 1$. When $G_1 = 1$, both the over-rejection for $\lambda = 2.0$ and the under-rejection for $\lambda = 0.5$ are very severe. For all values of λ, they are almost identical for RI-β and RI-t. As G_1 increases, the performance of RI-t initially improves quite quickly, while that of RI-β improves very slowly. However, the rate of improvement for RI-t slows

down greatly as G_1 increases. It still over-rejects noticeably for $G_1 = 10$ when $\lambda = 2.0$ and under-rejects noticeably when $\lambda = 0.5$.[11]

4 Randomization Inference and Interval P Values

The most natural way to calculate an RI P value is probably to use the equivalent of equation (11). Let S denote the number of repetitions, which would be G_0 when $G_1 = 1$ and the minimum of ${}_G\mathrm{C}_{G_1} - 1$ and B when $G_1 > 1$. Then the analog of (11) is

$$\hat{p}_1^* = \frac{1}{S}\sum_{j=1}^{S} \mathrm{I}\big(|t_j^*| > |\hat{t}|\big). \tag{17}$$

However, this method of computing a P value is arbitrary. A widely-used alternative is

$$\hat{p}_2^* = \frac{1}{S+1}\left(1 + \sum_{j=1}^{S} \mathrm{I}\big(|t_j^*| > |\hat{t}|\big)\right). \tag{18}$$

Both procedures are valid, as would be any procedure that yields a number between $\hat{p}_1^*$ and $\hat{p}_2^*$, because P values based on a finite number of simulations are interval-identified rather than point-identified. Evidently, $\hat{p}_1^*$ and $\hat{p}_2^*$ tend to the same value as $S \to \infty$, but they can yield quite different results when S is small.

Figure 7 shows analytical rejection frequencies for tests at the .05 level based on equations (17) and (18). The tests would reject exactly 5% of the time if S were infinite, but the figure is drawn for values of S between 7 and 103. In the figure, R denotes the number of times that $\hat{t}$ is more extreme than t_j^*, so that $\hat{p}_1^* = R/S$ and $\hat{p}_2^* = (R+1)/(S+1)$. It is evident that $\hat{p}_1^*$ always rejects more often than $\hat{p}_2^*$, except when $S = 19$, 39, 59, and so on.[12] Even for fairly large values of S, the difference between the two rejection frequencies can be substantial.[13]

This phenomenon does not cause a serious problem for bootstrap inference. We can obtain identical inferences from the two varieties of P value by choosing the number of bootstraps B (which is equivalent to S) so that $\alpha(B+1)$ is an integer, where α is the level of the test. In many cases, it is feasible to make B large, so that the interval between $\hat{p}_1^*$ and $\hat{p}_2^*$ must be very small whether or not $\alpha(B+1)$ is an integer.

Unfortunately, neither of these solutions works for randomization inference. As an extreme example, suppose the data come from Canada, which has just ten provinces. If one

[11] Using a different experimental design, Canay, Romano and Shaikh (2014) also study the performance of RI-β (and several other tests) when the treated clusters have greater variance than the untreated ones. They find even more severe overrejection for $\lambda = 2.0$ than we do.

[12] The figure is drawn under the assumption that we reject whenever either P value is equal to or less than 0.05. This is the only correct procedure for $\hat{p}_2^*$. However, for $\hat{p}_1^*$ it might be more natural to reject only when $\hat{p}_1^* < 0.05$. If that were done, the results for $\hat{p}_1^*$ with $S = 20$, 40, 60, and so on would be identical to the results for $\hat{p}_2^*$ with those values of S. The remainder of the figure would be unchanged.

[13] It is possible to obtain an exact test by using a draw from the U[0, 1] distribution. The procedure proposed in Racine and MacKinnon (2007) simply replaces the 1 after the large left parenthesis in (18) with such a draw. A similar procedure, which allows for ties, is used in Young (2015a). However, these procedures have the unfortunate property that the outcome of the test depends on the realization of a single random variable drawn by the investigator.

province is treated, then $G_1 = 1$, $G_0 = 9$, and the P value can lie in only one of nine intervals: 0 to 1/10, 1/9 to 2/10, 2/9 to 3/10, and so on. Even if $R = 0$, it would never be reasonable to reject at the .01 or .05 levels. The problem with P values not being point-identified is discussed at length in Webb (2014).

4.1 Wild Bootstrap Randomization Inference

In this subsection, we suggest a simple way to overcome the problem discussed in the previous one. We propose a procedure that we refer to as wild bootstrap randomization inference, or WBRI. The WBRI procedure essentially nests the RI-t procedure of Subsection 2.6 within the wild cluster bootstrap of Subsection 2.3. The procedure for generating the t^* statistics is as follows:

1. Estimate equation (1) by OLS and calculate $\hat{t}$ for the coefficient of interest using CRVE standard errors.

2. Construct a bootstrap sample, $\boldsymbol{y}_b^*$, using the restricted wild cluster bootstrap procedure discussed in Subsection 2.3. Then estimate equation (1) using $\boldsymbol{y}_b^*$ and calculate a bootstrap t statistic t_b^* using CRVE standard errors.

3. Re-estimate equation (1) using $\boldsymbol{y}_b^*$, sequentially changing the "treated" group(s) to all possible sets of G_1 groups, except the set that was actually treated. When $G_1 = 1$, this is done by cycling the "treated" group across all G_0 control groups. Calculate a t statistic t_{bj}^* for each randomization using CRVE standard errors.

4. Repeat steps 2 and 3 B times, constructing a new $\boldsymbol{y}_b^*$ in each step 2.

5. Perform inference by comparing $\hat{t}$ to the $B \times {}_G\mathrm{C}_{G_1}$ bootstrap statistics. These include B bootstrap statistics t_b^* that correspond to the G_1 actually treated groups and are drawn from exactly the same distribution as the t statistics in the restricted wild cluster bootstrap procedure, along with $B({}_G\mathrm{C}_{G_1} - 1)$ bootstrap statistics t_{bj}^* in which each set of groups other than the actual one is "treated" in turn.

The WBRI procedure can be used to generate as many t^* statistics as desired by making B large enough. Thus it can solve the problem of interval P values. However, it does not remedy the failure of the RI-t procedure when G_1 is small. Thus we cannot expect it to yield reliable inferences in that case when clusters are heterogeneous.

Since every possible set of G_1 clusters is "treated" in the bootstrap samples, the number of test statistics is $B \times {}_G\mathrm{C}_{G_1}$. Unless G is quite small, this will be a large number for $G_1 > 2$ even when $B = 1$. In general, it makes sense to use the WBRI procedure only when the RI-t procedure does not provide enough t_j^* for the interval P value problem to be negligible. As a rule of thumb, we suggest using WBRI when $G_1 = 1$ and $G < 500$, or $G_1 = 2$ and $G < 45$, or $G_1 = 3$ and $G < 20$. We suggest choosing B so that $B \times {}_G\mathrm{C}_{G_1}$ is at least 1000. If G is sufficiently small, one may want to enumerate (that is, pick every possible value from) the Rademacher distribution, or use an alternative bootstrap weight distribution such as the 6-point distribution suggested in Webb (2014).

4.2 Monte Carlo Results for WBRI

The WBRI procedure described in Subsection 4.1 is designed to avoid the problem of interval P values. Based on Figure 8, it seems to be quite effective at doing so. The figure deals with the case in which $G_1 = 1$, which is when the interval P value problem is most severe. Every cluster has 100 observations, and the number of clusters varies from 10 to 60, which implies that the number of controls varies from 9 to 59. Thus when $G = 20$, 40, and 60, the two RI P values must yield the same outcomes. In every other case, however, $\hat{p}_1^* = R/S$ must reject more often than $\hat{p}_2^* = (R+1)/(S+1)$. As expected, the observed rejection frequencies for the two RI tests look very similar to the theoretical ones in Figure 7.

In Figure 8, the WBRI rejection frequencies are almost always between the two RI rejection frequencies and are always quite close to 5% except when G is very small. This is what we would like to see. However, it must be remembered that the figure deals with a very special case. The WBRI procedure cannot be expected to work any better than the RI-t procedure when the treated clusters are smaller or larger than the untreated clusters, or when their error terms have different variances. There is at present no proof that it will work well even when the only problem is that the number of control groups is small and α times one plus that number is not an integer.

5 Empirical Examples

In this section, we consider two empirical examples. In the first of them, $G_1 = 2$, so that no method can be expected to work very well. In the second, $G_1 = 10$, so that the WCR bootstrap should work well and randomization inference should not work better. We include the second example because it was used in Conley and Taber (2011).

5.1 Birth Control Pills

Bailey (2010) examines the relationship between the introduction of the birth control pill and the decrease in fertility in the United States since about 1957. The paper uses state-by-state variation in "Comstock laws," which prohibited, among other things, the advertising and sale of the birth control pill. The practice of using these laws to restrict the sale of birth control pills was essentially ended by the U.S. Supreme Court's 1965 Griswold v. Connecticut decision.

Part of the analysis in Bailey (2010) shows that women in states with sales restrictions on the birth control pill were indeed less likely to have taken the pill by 1965. The analysis employs a DiD regression using data on married, white women from the *National Fertility Surveys* for the years 1965 and 1970. The women come from 47 states, and clustering is done at the state level.

Bailey estimates a probit regression in which the dependent variable is an indicator variable that equals 1 in 1965 or 1970 if the respondent had ever taken the birth control pill by that year. The key regressors are an indicator variable `Salesban` that equals 1 if the state had a sales ban on the birth control pill in 1960, and `Salesban` interacted with a dummy variable `D1970` for observations from 1970. Estimated coefficients and standard errors for these two regressors are presented in her Table 2, Column 1. Other regressors include `D1970`, three regional dummies, an indicator variable equal to 1 if the state had a physician exemption to the sales ban, and each of these variables interacted with `D1970`.

There is no real need to use a probit model in this case. Because all regressors are indicator variables, and the mean of the dependent variable (which is 0.515) is far from the limits of 0 and 1, using ordinary least squares inevitably produces results almost identical to the probit ones. In fact, the probit t statistics for `Salesban` and `Salesban`×`D1970` are −2.76 and 1.46, and the OLS ones are −2.71 and 1.37; these are all based on cluster-robust standard errors.[14]

Prior to the "Griswold" decision, several states repealed their previously existing sales bans. In particular, Illinois and Colorado repealed their Comstock laws in 1961. It is of interest to ask whether women in these early-repeal states were more or less likely to use the pill than women in other states with a sales ban. We therefore created an indicator variable `rep61` equal to 1 for those two states and added `rep61`×`D1965` and `rep61`×`D1970` to the base specification. Results for the four coefficients of interest are shown in Table 1.

Table 1: Effects of Sales Ban and Early Repeal, Full Sample

	Coef.	Std. Err.	CR t-stat	WCR p^*	RI-β p^*	RI-t p^*
`Salesban`	−0.0418	0.0156	−2.677	0.0275		
`Salesban`×`D1970`	0.0290	0.0274	1.059	0.3196		
`rep61`×`D1965`	−0.1253	0.0231	−5.432	0.5455	0.0629	0.0555
`rep61`×`D1970`	−0.0427	0.0287	−1.488	0.4577	0.6152	0.4450

Taken at face value, the cluster-robust t statistic for `rep61`×`D1965` in column 3 of Table 1 appears to be telling us that living in an early-repeal state very significantly lowered the probability of using the pill in 1965. However, because there are only two such states, the analysis of Section 2.1 suggests that this t statistic is probably much too large. In contrast, the WCR bootstrap (based on $B = 99{,}999$) yields a P value of about 0.55, which the analysis of MacKinnon and Webb (2016) suggests is probably much too conservative. Thus the cluster-robust t statistic and the bootstrap P value yield wildly contradictory results, which could have been expected before even computing them, and are therefore of no real use in this case.

We also compute two randomization inference P values for each regressor. Because $G = 47$, the value of S is $47 \cdot 46/2 - 1 = 1080$. We report RI P values computed using equation (18), because they are slightly more conservative than ones based on equation (17). The two RI procedures yield results that are very similar to each other, with P values just a little greater than .05. Although the RI P values do not entirely resolve the uncertainty about whether the coefficient on `rep61`×`D1965` is significant, they at least yield sensible results that could not have been predicted in advance.

One way to investigate the robustness of these results is to limit attention to the 23 states that had sales bans in 1960. This reduces the sample size to 3780 observations and requires us to drop the variables `Salesban` and `Salesban`×`D1970`. Results for the two coefficients of interest are shown in Table 2.

[14] Although we attempted to use the same sample as Bailey, our sample has 6929 observations, and hers has 6950. We are unable to explain this minor discrepancy. Bailey does not explicitly report t statistics. Calculating them from coefficients and standard errors reported to only two decimal places, her t statistics are similar enough to our probit ones that they could actually be equal.

Table 2: Effects of Early Repeal, Sales Ban Sample

	Coef.	Std. Err.	CR t-stat	WCR p^*	RI-β p^*	RI-t p^*
`rep61×D1965`	−0.1199	0.0244	−4.917	0.4390	0.0395	0.0791
`rep61×D1970`	−0.0457	0.0316	−1.445	0.4602	0.3162	0.2885

The results in Table 2 are very similar to the ones in Table 1. The most noticeable difference is that the RI-β P value is now just 0.0395, while the RI-t P value is almost exactly twice as large. Since these are based on $S = 23 \cdot 22/2 - 1 = 252$, they would have been noticeably smaller (0.0357 and 0.0754) if we had used equation (17) rather than equation (18) to compute them. Because it is difficult to understand why early repeal would have reduced pill use in 1965, we believe the RI-t results to be more plausible than the RI-β ones. This accords with the results in Figures 1, 3, and 5, all of which suggest that RI-t should be somewhat less unreliable than RI-β when $G_1 = 2$.

Although this example is not one where randomization inference can be expected to work well, because there are only two treated clusters, RI certainly yields results that are much more plausible, and much less predictable, than using either cluster-robust t statistics or the wild cluster bootstrap.

5.2 Merit Scholarships

In this subsection, we consider an empirical example studied in Conley and Taber (2011). It deals with the impact of state-level merit scholarships initiated during the 1989-2000 period. These programs generally offered scholarships for students to attend college in their home state conditional on being above some academic threshold. The details differ state by state, but they are not important for our purposes.

Conley and Taber (2011) attempts to determine whether the 10 merit scholarships that were in operation by the end of 2000 had any impact on college enrollment by estimating the following DiD regression using data from 1989-2000:

$$\begin{aligned}\text{college}_{ist} = \beta_0 + \beta_1 \text{merit}_{ist} + \beta_2 \text{male}_{ist} + \beta_3 \text{black}_{ist} + \beta_4 \text{asian}_{ist} \\ + \sum_{j=2}^{51} \gamma_j \text{state}_{ist}^j + \sum_{k=2}^{12} \delta_k \text{year}_{ist}^k + \epsilon_{ist}.\end{aligned}$$

Here college_{ist} is the outcome of interest, a binary indicator for whether individual i in state s and year t was enrolled in college, and the treatment variable merit_{ist} equals 1 if state s offered a merit scholarship in year t. The remaining variables are all binary indicator variables. The dataset has $N = 42{,}161$ observations taken from all states, including the District of Columbia, so that $G = 51$. The paper presents estimates of β_1 along with several different confidence intervals in Column C of Table II.

The table reports that $\hat{\beta}_1 = 0.034$, along with a 95% CRVE confidence interval of [0.008, 0.059]. Although it is not explicitly reported, we calculated the P value for the test of $\beta_1 = 0$ based on the $t(50)$ distribution to be 0.010. However, using a method that essentially inverts RI-β P values, the paper estimates a 95% confidence interval for β_1 of

$[-0.003,\ 0.093]$.[15] Thus, unlike the conventional CRVE confidence interval, the Conley-Taber 95% confidence interval contains 0.

We use the Conley-Taber data and modify their Stata code to conduct inference on β_1 using both RI-β and RI-t.[16] With 3000 randomizations and symmetric P values, we obtain an RI-t P value of 0.0317 and an RI-β P value of 0.11667. Like the Conley-Taber confidence interval, the RI-β P value fails to reject the null at the 5% level. In contrast, our RI-t P value of 0.0317 suggests that there is a statistically significant effect at the 5% level.

We also calculate the WCR P value for $\beta_1 = 0$, based on $B = 99,999$ bootstraps. It is 0.021, which is quite similar to the RI-t P value. With 10 treated states, the WCR P value should be quite reliable. In view of this result and the fact that, in all our Monte Carlo experiments, the RI-t procedure was closer to the desired size than RI-β, we conclude that the merit scholarship programs did have a statistically significant impact.

6 Conclusion

We compare several new and existing procedures for inference with few treated clusters, focusing on ones based on randomization inference (RI). There are three main findings, which are obtained theoretically for a simple model in Subsection 2.7 and confirmed by simulation results in Section 3.

The first result is that none of the procedures works well when there are very few treated clusters and those clusters are atypical in terms of either the number of observations or the variance of the error terms. The second is that all the RI procedures appear to work well when the treated clusters are typical or chosen at random. The third is that the performance of procedures based on randomization inference for coefficients (RI-β) improves slowly or not at all as G_1 (the number of treated clusters) increases, while the performance of procedures based on randomization inference for t statistics (RI-t) generally improves quite rapidly. Thus the latter can often be used safely when G_1 is fairly small, but not extremely small. We also introduce a bootstrap-based modification of randomization inference which appears to solve the problem of interval P values when there are few control groups.

[15] The procedure searches separately for both the upper and lower limit of the confidence interval, by re-randomizing treatment amongst 10 of the 51 states.

[16] We thank the authors for making their code and data easily available.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010) 'Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program.' *Journal of the American Statistical Association* 105(490), 493–505

Bailey, Martha A. (2010) '"Momma's got the pill": How Anthony Comstock and Griswold v. Connecticut shaped US childbearing.' *American Economic Review* 100(1), 98–129

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004) 'How much should we trust differences-in-differences estimates?' *The Quarterly Journal of Economics* 119(1), pp. 249–275

Bester, C. Alan, Timothy G. Conley, and Christian B. Hansen (2011) 'Inference with dependent data using cluster covariance estimators.' *Journal of Econometrics* 165(2), 137–151

Cameron, A. Colin, and Douglas L. Miller (2015) 'A practitioner's guide to cluster robust inference.' *Journal of Human Resources* 50, 317–372

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2008) 'Bootstrap-based improvements for inference with clustered errors.' *The Review of Economics and Statistics* 90(3), 414–427

Canay, Ivan A, Joseph P Romano, and Azeem M Shaikh (2014) 'Randomization tests under an approximate symmetry assumption.' *Technical Report No. 2014-13, Stanford University*

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald (2015) 'Asymptotic behavior of a t test robust to cluster heterogeneity.' Technical Report, University of California, Santa Barbara

Conley, Timothy G., and Christopher R. Taber (2011) 'Inference with "Difference in Differences" with a small number of policy changes.' *The Review of Economics and Statistics* 93(1), 113–125

Davidson, Russell, and Emmanuel Flachaire (2008) 'The wild bootstrap, tamed at last.' *Journal of Econometrics* 146(1), 162 – 169

Donald, Stephen G, and Kevin Lang (2007) 'Inference with difference-in-differences and other panel data.' *The Review of Economics and Statistics* 89(2), 221–233

Dufour, Jean-Marie (2006) 'Monte carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics.' *Journal of Econometrics* 133(2), 443–477

Ferman, Bruno, and Christine Pinto (2015) 'Inference in differences-in-differences with few treated groups and heteroskedasticity.' Technical Report, Sao Paulo School of Economics

Fisher, R.A. (1935) *The Design of Experiments* (Oliver and Boyd)

Ibragimov, Rustam, and Ulrich K. Müller (2016) 'Inference with few heterogeneous clusters.' *Review of Economics & Statistics* 98, to appear

Imbens, Guido W., and Michal Kolesar (2016) 'Robust standard errors in small samples: Some practical advice.' *Review of Economics and Statistics* 98, to appear

Lehmann, E. L., and Joseph P. Romano (2008) *Testing Statistical Hypotheses* Springer Texts in Statistics (Springer New York)

Liang, Kung-Yee, and Scott L. Zeger (1986) 'Longitudinal data analysis using generalized linear models.' *Biometrika* 73(1), 13–22

MacKinnon, James G., and Matthew D. Webb (2016) 'Wild bootstrap inference for wildly different cluster sizes.' *Journal of Applied Econometrics* 31, to appear

Racine, Jeffrey S., and James G. MacKinnon (2007) 'Simulation-based tests that can use any number of simulations.' *Communications in Statistics: Simulation and Computation* 36(2), 357–365

Rosenbaum, Paul R (1996) '6 observational studies and nonrandomized experiments.' *Handbook of Statistics* 13, 181–197

Webb, Matthew D. (2014) 'Reworking wild bootstrap based inference for clustered errors.' Working Papers 1315, Queen's University, Department of Economics, August

Yates, F. (1984) 'Tests of significance for 2×2 contingency tables.' *Journal of the Royal Statistical Society. Series A (General)* 147(3), 426–463

Young, Alwyn (2015a) 'Channelling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results.' Technical Report, London School of Economics

Young, Alwyn (2015b) 'Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections.' Technical Report, London School of Economics

Figure 1: Rejection Frequencies for Randomization Inference

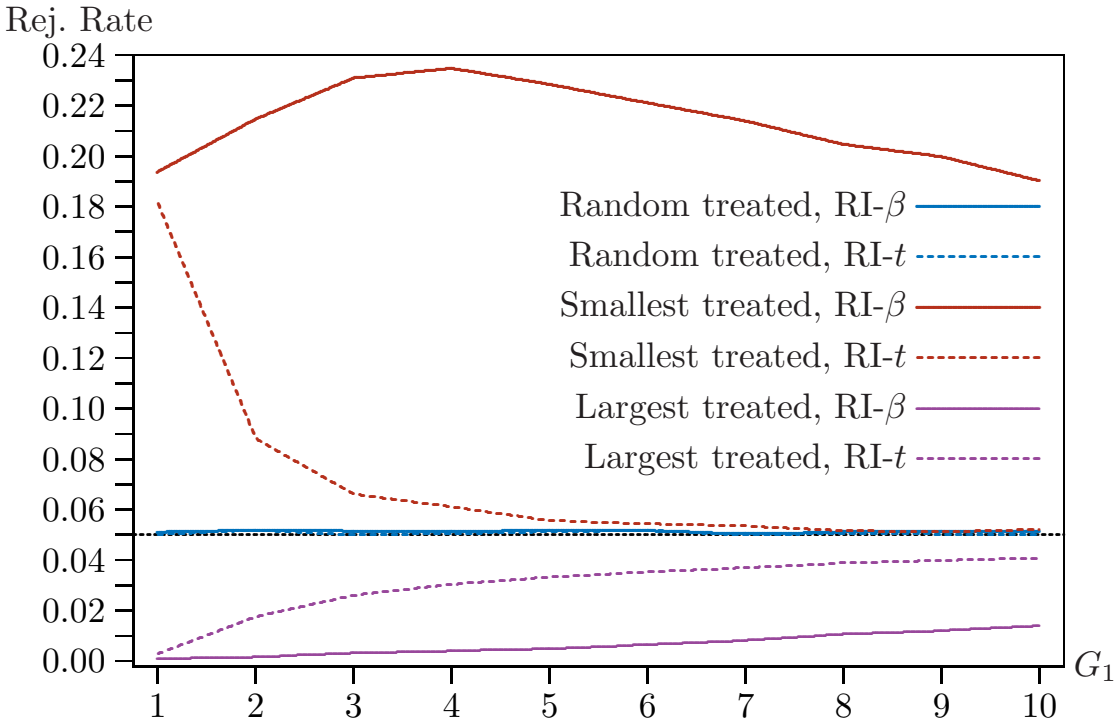


Figure 2: Rejection Frequencies for t Statistic RI when $G_1 = 1$

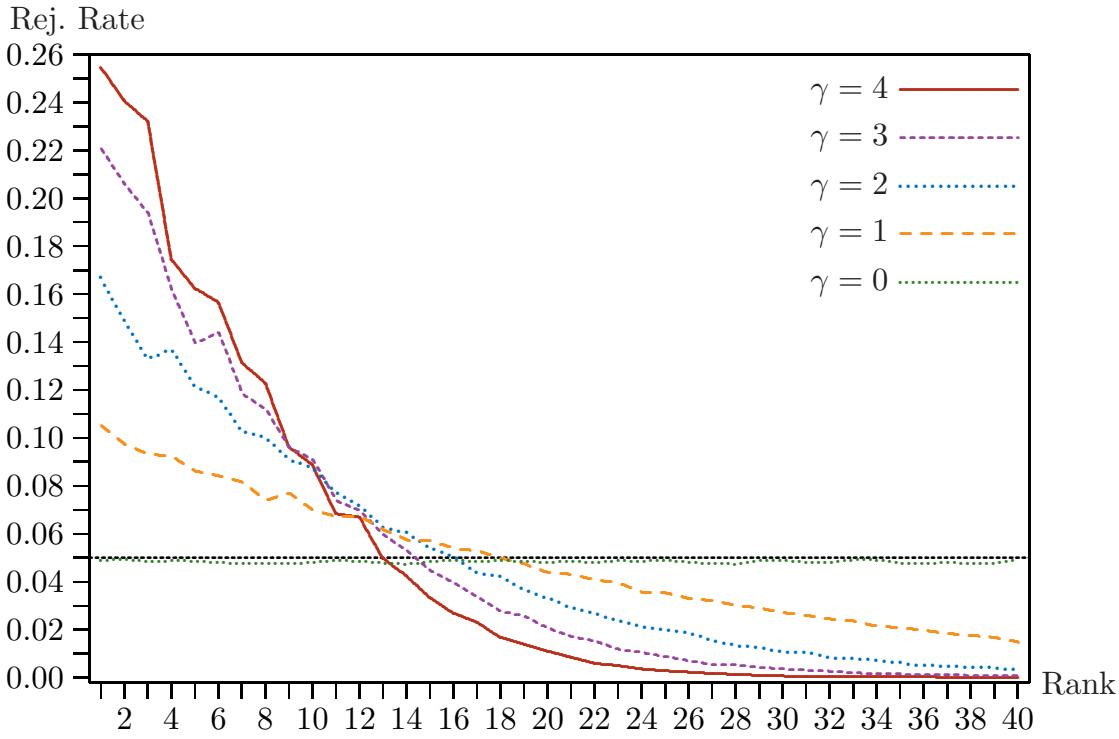


Figure 3: Rejection Frequencies for Two Types of Randomization Inference

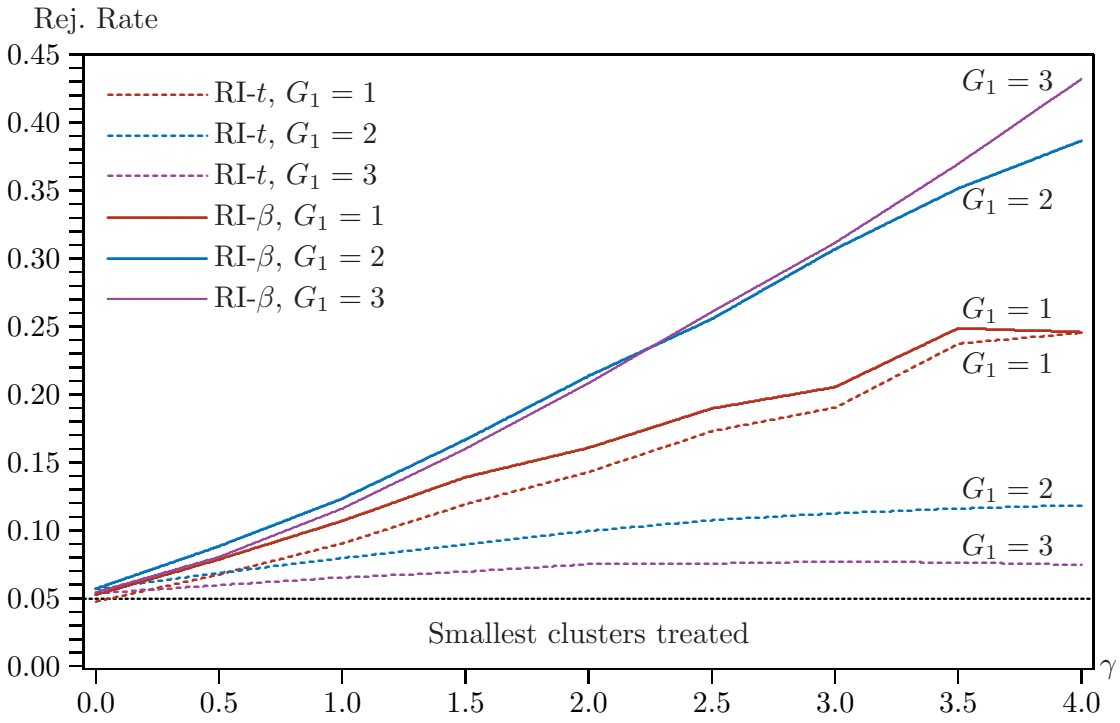


Figure 4: Rejection Frequencies for the WCR Bootstrap

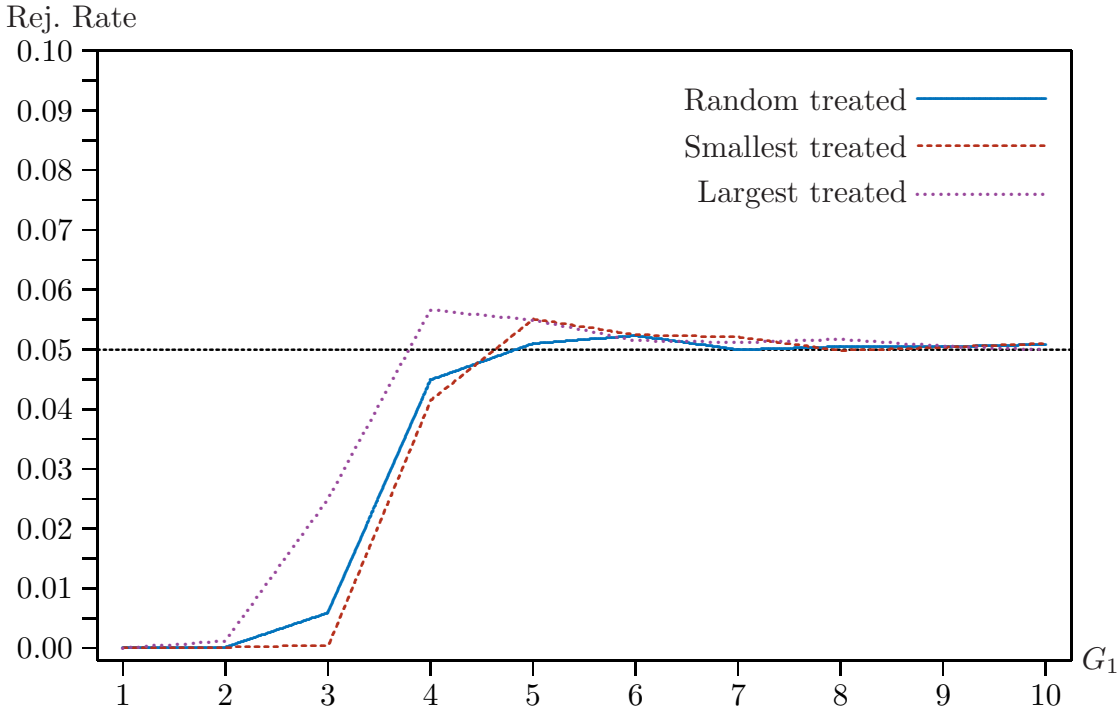
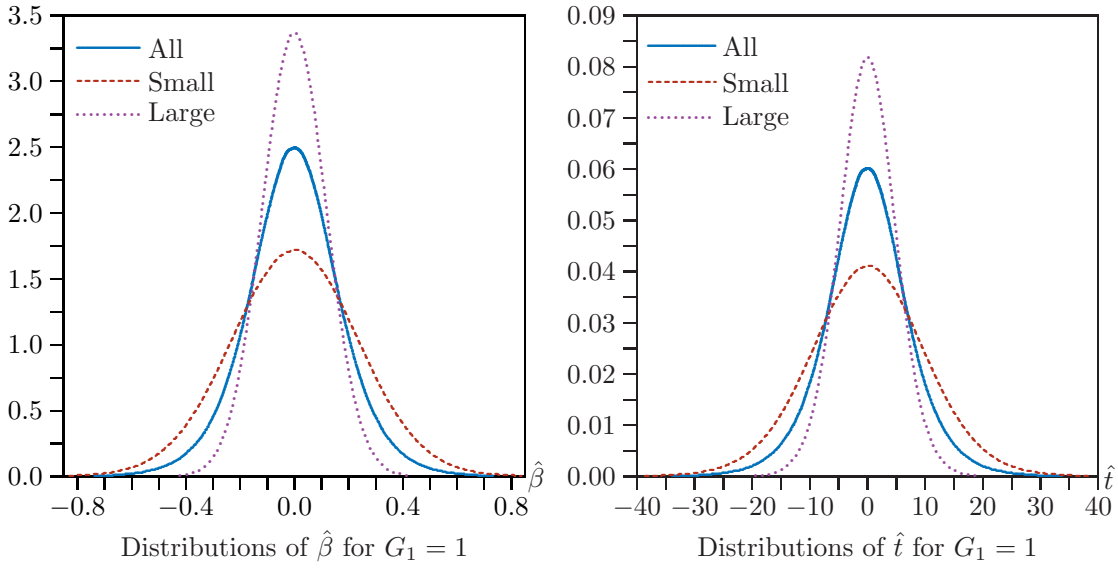
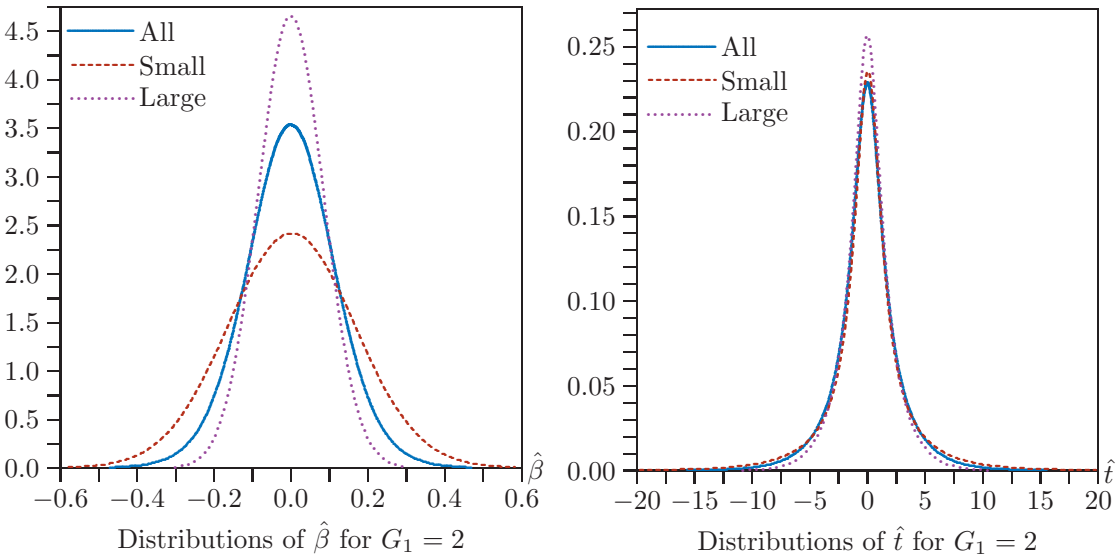


Figure 5: Empirical Distributions of $\hat{\beta}$ and $\hat{t}$



Distributions of $\hat{\beta}$ for $G_1 = 1$ | Distributions of $\hat{t}$ for $G_1 = 1$



Distributions of $\hat{\beta}$ for $G_1 = 2$ | Distributions of $\hat{t}$ for $G_1 = 2$

Figure 6: Rejection Frequencies for RI Procedures with Heteroskedasticity

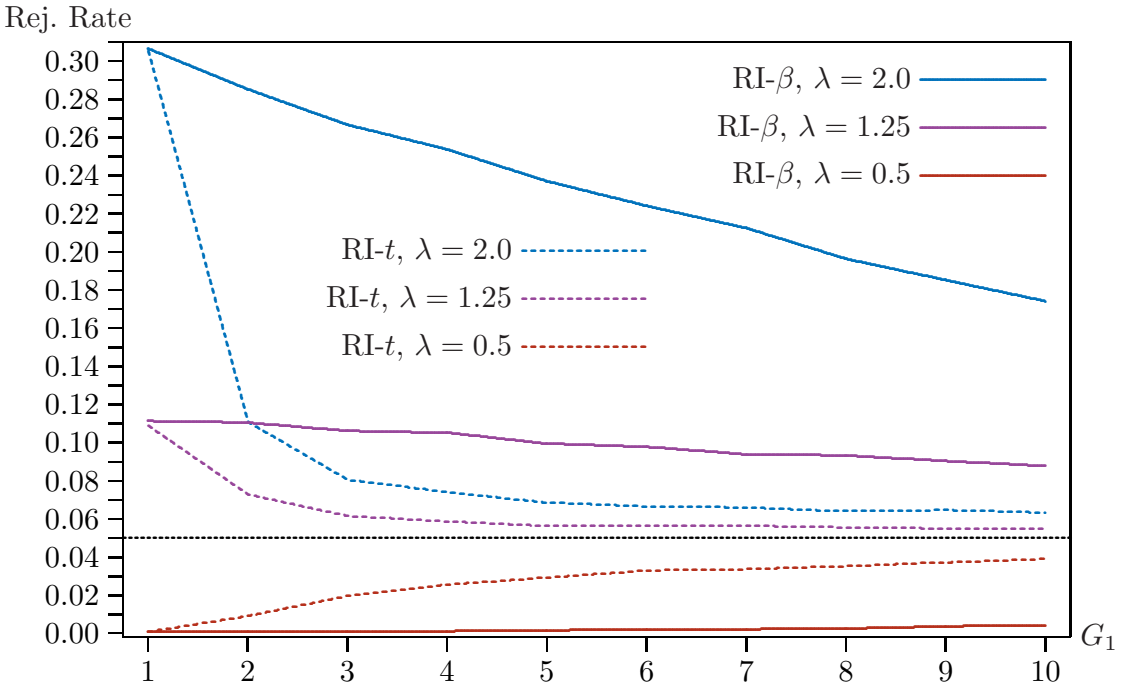


Figure 7: Rejection Frequencies and Number of Simulations

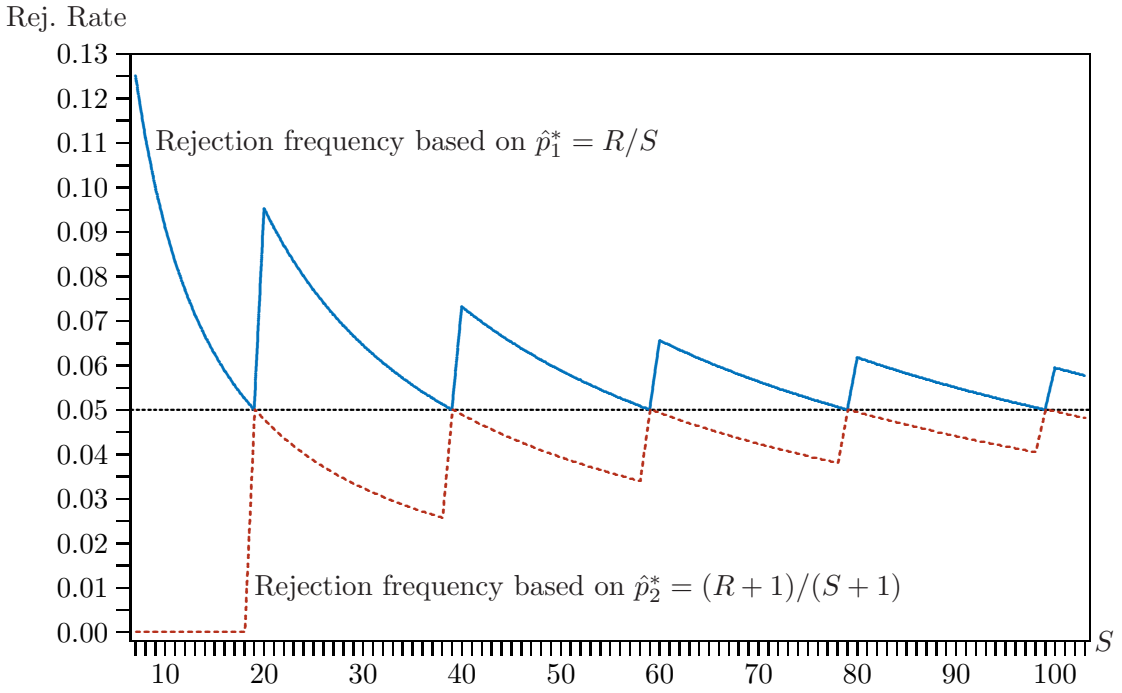


Figure 8: WBRI Rejection Frequencies and RI Intervals