Asymptotic Inference with Clustering

Consider the t-statistic for $\boldsymbol{a}^\top\boldsymbol{\beta} = \boldsymbol{a}^\top\boldsymbol{\beta}_0$:

$$t_a = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{(\boldsymbol{a}^\top\hat{\boldsymbol{V}}\boldsymbol{a})^{1/2}}. \tag{1}$$

Often just one element of $\boldsymbol{a}$ equals 1, and the remaining elements equal 0. Then (1) is simply $\hat{\beta}_j - \beta_{j0}$ divided by its standard error.

Here $\hat{\boldsymbol{V}}$ denotes one of CV_1, CV_2, or CV_3.

When there are $r > 1$ linear restrictions of the form $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, inference can be based on the Wald statistic

$$W = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top(\boldsymbol{R}\hat{\boldsymbol{V}}\boldsymbol{R}^\top)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}). \tag{2}$$

When $r = 1$, the t-statistic (1) is just the signed square root of a particular Wald statistic with $\boldsymbol{R} = \boldsymbol{a}^\top$ and $\boldsymbol{r} = \boldsymbol{a}^\top\boldsymbol{\beta}_0$.

Asymptotic analysis involves letting the sample size become arbitrarily large, so that all but the highest-order terms vanish.

With clustered data, there is more than one natural way to let the sample size become large.

We can make various assumptions about what happens to G and the N_g as we let N tend to infinity.

Two key asymptotic results must hold.

1. A central limit theorem (CLT) must apply to the sum of the score vectors $\boldsymbol{s}_g$. The vector $\sum_{g=1}^{G} \boldsymbol{s}_g$ needs to follow a multivariate normal distribution with variance matrix $\sum_{g=1}^{G} \boldsymbol{\Sigma}_g$.
2. A law of large numbers (LLN) must apply to the matrices $\sum_{g=1}^{G} \hat{\boldsymbol{s}}_g \hat{\boldsymbol{s}}_g^\top$, $\sum_{g=1}^{G} \grave{\boldsymbol{s}}_g \grave{\boldsymbol{s}}_g^\top$, or $\sum_{g=1}^{G} \acute{\boldsymbol{s}}_g \acute{\boldsymbol{s}}_g^\top$ that appear in the expressions for CV_1, CV_2, or CV_3, so that they converge to $\sum_{g=1}^{G} \boldsymbol{\Sigma}_g$.

For asymptotic inference to be reliable, we need both the CLT and the LLN to provide good approximations. That is not always the case!

There are currently two quite different types of assumptions on which the asymptotic theory of cluster-robust inference can be based.

1. Let the number of clusters tend to infinity (**large-*G***), without restricting them to be small.
2. Hold the number of clusters fixed and let the number of observations within each cluster tend to infinity (**fixed-*G***).

The large-G approach generally seems preferable, because the fixed-G approach involves very strong and unrealistic assumptions.

However, both approaches provide valuable insights.

- Inference based on asymptotic theory can perform well, but it performs poorly in many commonly-encountered situations.
- This is particularly true for CV_1. The cluster jackknife (CV_3) is always more conservative and usually works better.
- Bootstrap methods may work even better, especially some versions of the **wild cluster bootstrap**.

Large Number of Clusters

The easiest approach is to assume that every cluster has a fixed number of observations, say M.

Then $N = MG$, and both N and G go to infinity at the same rate.

In this special case, the appropriate normalizing factor for the parameter estimator is either $\sqrt{G}$ or $\sqrt{N}$.

It is not difficult to show that $\sqrt{G}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically multivariate normal with variance matrix equal to the plim of

$$G(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\Bigl(\sum_{g=1}^{G}\boldsymbol{\Sigma}_g\Bigr)(\boldsymbol{X}^\top\boldsymbol{X})^{-1}. \tag{3}$$

This is evidently $O(1)$ when $N = MG$, since the filling in the sandwich is $O(G)$, and the two inverse matrices are $O(1/N)$.

We can relax the assumption that $G/N = M$ by allowing G to be only approximately proportional to N, so that G/N is roughly constant as $N \to \infty$.

Djogbenou, MacKinnon, and Nielsen (JoE, 2019) and Hansen and Lee (JoE, 2019) take a more flexible approach.

- They allow some, and possibly all, of the $N_g \to \infty$ as $N \to \infty$, but more slowly than N itself does.
- Although a key assumption is that $G \to \infty$, the appropriate normalizing factor for $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ is usually not $\sqrt{G}$.
- This factor depends on the regressors, the relative cluster sizes, the intra-cluster correlation structure, and interactions among these.

The *t*-statistic defined in (1) is shown to be asymptotically standard normal even though the rate at which $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ tends to zero is unknown. This is a case of **self-normalizing asymptotics**.

There are technical conditions that must be satisfied if $\hat{\boldsymbol{\beta}}$ is to be consistent and asymptotically normal.

The closer these conditions are to being violated, the less well we would expect asymptotic inference to perform.

- We cannot allow a single cluster to dominate the sample, in the sense that its size is proportional to N.
- When the scores are uncorrelated, the size of the largest cluster must increase no faster than the square root of the sample size. However, it can increase faster than this when there is a lot of intra-cluster correlation.
- Asymptotic inference tends to be unreliable when N_g are highly variable, especially when a very few clusters are unusually large.
- Asymptotic inference also tends to be unreliable when the clusters are heterogeneous (unbalanced) in other respects.

To see whether asymptotic inference is likely to be reliable, we should see how large G is, how much the N_g vary, and how much variation there is in leverage and partial leverage.

Small Number of Large Clusters

Some authors have assumed that G remains fixed (i.e., is “small”) as $N \rightarrow \infty$, while the cluster sizes diverge (i.e., are “large”).

Bester, Conley, and Hansen (JoE, 2011) proved that, for CV_1, the t-statistic (1) follows the $t(G-1)$ distribution asymptotically under very strong assumptions:

- All the clusters are assumed to be the same size M.
- A CLT must apply to the normalized score vectors $M^{-1/2}\boldsymbol{s}_g$ for all $g = 1, \ldots, G$, as $M \rightarrow \infty$.

This second assumption limits the amount of dependence within each cluster and requires it to diminish quite rapidly as $M \rightarrow \infty$. It rules out the random-effects model.

The $t(G-1)$ distribution is the default in `Stata` for CV_1-based inference. It can lead to noticeably more accurate, and more conservative, inferences than the $t(N-k)$ distribution.

When Asymptotic Inference Can Fail

Inference is "reliable" if:

1. tests at level α reject approximately α% of the time under the null;
2. confidence intervals at level $1-\alpha$ cover the true value approximately $(1-\alpha)$% of the time.

It is sometimes claimed that asymptotic inference based on CV_1 is reliable when $G \geq 50$.

This is false. There is no magic value of G that is always big enough.

In very favorable cases, inference based on CV_1 and the $t(G-1)$ distribution can be fairly reliable when $G=20$, but in unfavorable ones it can be unreliable even when $G=200$ or more.

Inference based on CV_3 tends to be more reliable, sometimes much more reliable, than inference based on other CRVEs. It can sometimes be too conservative.

Cluster Heterogeneity

The number of clusters G and the extent to which the distribution of the score vectors vary across clusters determines the quality of the asymptotic approximation.

- When cluster sizes vary a lot, it is very likely that the score vectors will also vary greatly across clusters.
- They can also vary due to heteroskedasticity of the disturbances at the cluster level and systematic variation across clusters in the distribution of the regressors.
- In view of the rate condition in the proofs for the large-G case, asymptotic approximations will surely become worse as $\max N_g$ increases relative to N/G.

For CV_1 and CV_2, t-tests always over-reject when the approximation is poor. For CV_3, they can either over-reject or under-reject.

Cluster-robust *t*-tests and Wald tests are at risk of over-rejecting to an extreme extent in two situations.

- One or a few clusters are unusually large.
- Only a few clusters are treated.

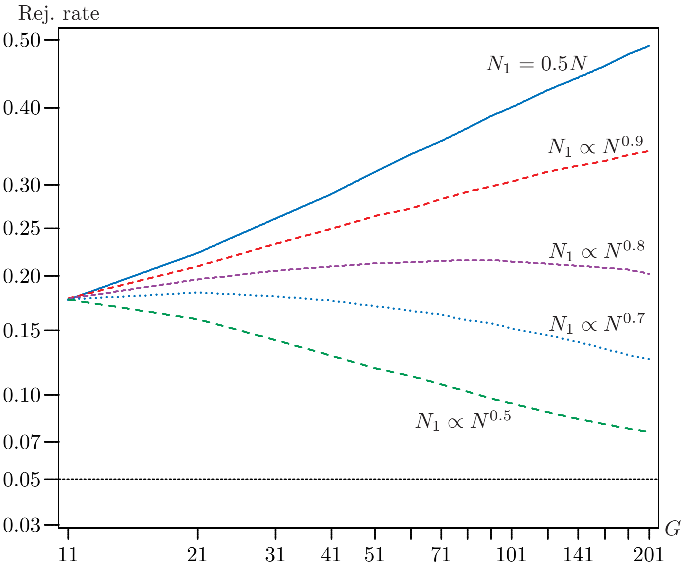
In these cases, one cluster, or just a few of them, have high leverage, in the sense that omitting one of these clusters has the potential to change the OLS estimates substantially.

- Both of these situations can occur even when G is not small.
- When one cluster is unusually large, the distribution of its score vector is much more spread out than the ones for other clusters.

Djogbenou, MacKinnon, and Nielsen (JoE, 2019) studies a case where up to half the sample is in one cluster. Rejection rates for CV_1 t-tests at .05 level increase as G increases.

This is empirically relevant, since more than half of all incorporations in the U.S. are in Delaware!

Figure 1: Rejection rates for CV$_1$ t-tests when there is one big cluster



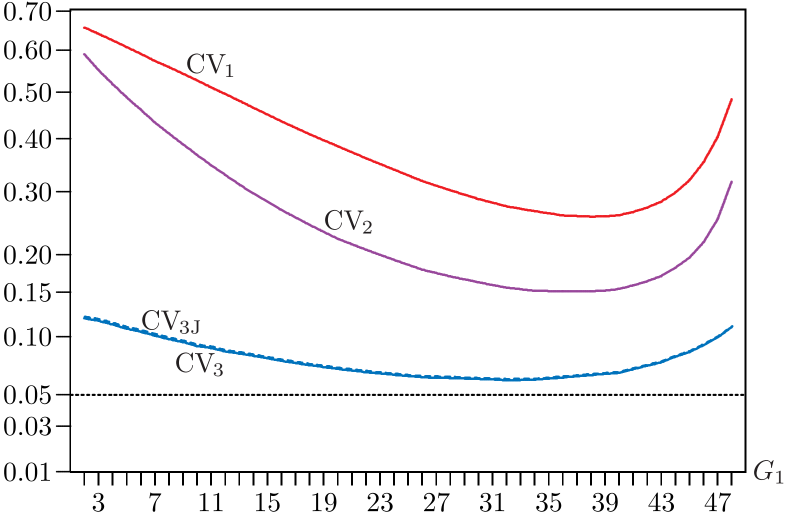
Although CV_3 works far from perfectly, it performs very much better than CV_1 and CV_2 in a simulation experiment with cluster sizes proportional to incorporations in U.S. states.

Figure 2 is from MacKinnon, Nielsen, and Webb (JAE, 2023). Delaware is always treated, along with 1 to 47 other states.

Having some extremely small clusters in a sample generally does not cause any problems, so long as there is not too much heterogeneity in the remainder of the sample.

- Suppose that a sample consists of, say, 25 large clusters, each with roughly 200 observations, and 15 tiny clusters, each with just one or a handful of observations.
- Except in very unusual cases, coefficient estimates and t-statistics would hardly change if we were to drop the tiny clusters, so this sample is better thought of as having 25 equal-sized clusters.
- The asymptotic approximations would perform just about the same whether or not the tiny clusters were included.

Figure 2: Rejection rates with cluster sizes proportional to incorporations



Tests at .05 level. 52.8% of the observations are for Delaware, which is always treated.

Treatment and Few Treated Clusters

Conventional inference fails when the regressor of interest is a treatment dummy, and treatment occurs only for observations in a small number of clusters.

Empirical score vectors for the treated clusters, even when modified (CV_2, CV_3), can provide very poor estimates of the actual score vectors.

Suppose d_{gi} is the treatment dummy for observation gi, and s_g^d is the element of $\boldsymbol{s}_g$ corresponding to the dummy.

If only observations in cluster 1 are treated, then

$$\begin{aligned} s_g^d = \sum_{i=1}^{N_g} d_{gi} u_{gi} &= \sum_{i=1}^{N_1} d_{1i} u_{1i} \text{ for } g = 1 \\ &= 0 \text{ for all } g \neq 1. \end{aligned} \tag{4}$$

The scores for the treatment dummy equal 0 for all control clusters!

Because the treatment regressor must be orthogonal to the residuals, the empirical score for cluster 1, $\hat{s}_1^d$, equals 0.

Since the actual score $s_1^d \neq 0$, this implies that CV1 provides a dreadful estimate of $\boldsymbol{V}(\hat{\boldsymbol{\beta}})$, at least for the coefficient on the treatment dummy.

The CV_1 standard error of this coefficient can easily be too small by a factor of five or more.

When more than one cluster is treated, the problem is not as severe.

- The $\hat{s}_g^d$ sum to 0 over the observations in all treated clusters.
- This causes them to be too small, but not to the same extent as when just one cluster is treated.
- How well the $\hat{\boldsymbol{s}}_g$ mimic the $\boldsymbol{s}_g$ depends on the sizes of treated and control clusters, the values of other regressors, and the numbers of treated observations within treated clusters.
- As G_1 (the number of treated clusters) increases, the problem often goes away fairly rapidly.

Increasing G when G_1 is small and fixed does not help and may cause over-rejection to increase.

For asymptotic theory to work, we need G_1/G to tend to a constant as the sample size increases.

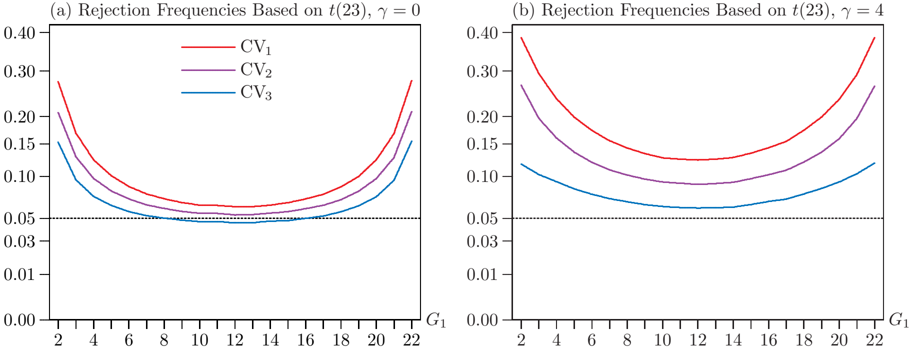
- When a cluster is either fully treated or not, having very few control clusters is as bad as having very few treated clusters.
- In fact, for models with balanced clusters, there is a perfect symmetry between $[G_1, G - G_1]$ and $[G - G_1, G_1]$.
- The situation is more complicated for DiD models, because then only some observations in the treated clusters are treated.

For detailed treatments of the few-treated problem, see MacKinnon and Webb (JAE, 2017; TPM, 2017; EJ, 2018).

When G_1 is small, tests based on CV_3 over-reject considerably less severely than ones based on CV_1. But they still over-reject.

Figure 3 is taken from MacKinnon, Nielsen, and Webb (JAE, 2023).

Figure 3: Rejection rates for t-tests as number of treated clusters varies



- Tests are at the .05 level. $N = 9600$ and $G = 24$.
- When $\gamma = 0$, $N_g = 400$ for all g.
- When $\gamma = 4$, the N_g vary from 32 to 1513.

The Pairs Cluster Bootstrap

This bootstrap DGP is applicable to every model for clustered data.

It is also sometimes referred to as the **cluster bootstrap**, the **block bootstrap**, the **pairwise bootstrap**, or **resampling by cluster**.

The pairs cluster bootstrap works by grouping the data for every cluster into a $[\boldsymbol{y}_g, \boldsymbol{X}_g]$ pair and then resampling from the G pairs.

Every bootstrap sample is constructed by choosing G pairs at random with equal probability $1/G$. Thus N varies across them, unless all cluster sizes are the same.

Null not imposed, so numerator of bootstrap t-statistic must be $\hat{\beta}^*_b - \hat{\beta}$.

The bootstrap samples may not mimic the actual sample well because:

- The largest (or smallest) clusters may be over-represented in some bootstrap samples and under-represented in others.
- The $\boldsymbol{X}^\top \boldsymbol{X}$ matrix is different for every bootstrap sample.

For linear regression models, we should form the matrices and vectors

$$\boldsymbol{X}_g^\top \boldsymbol{X}_g, \;\; \boldsymbol{X}_g^\top \boldsymbol{y}_g, \quad g = 1, \ldots, G. \tag{5}$$

Then we resample from the pairs $[\boldsymbol{X}_g^\top \boldsymbol{X}_g, \; \boldsymbol{X}_g^\top \boldsymbol{y}_g]$.

Summing these yields the bootstrap sample $[\boldsymbol{X}^{*\top} \boldsymbol{X}^*, \;\; \boldsymbol{X}^{*\top} \boldsymbol{y}^*]$, from which we obtain

$$\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*\top} \boldsymbol{X}^*)^{-1} \boldsymbol{X}^{*\top} \boldsymbol{y}^*. \tag{6}$$

We can also readily obtain the CV_1 and CV_3 variance estimators for $\hat{\boldsymbol{\beta}}^*$. CV_2 takes a bit more work.

As usual, we generate B bootstrap samples and use them to compute $\hat{\boldsymbol{\beta}}^{*b}$ for $b = 1, \ldots, B$.

Unless we are just calculating bootstrap standard errors, we also need to compute a test statistic, say τ_b^*, for $b = 1, \ldots, B$. The τ_b^* could be either t statistics or Wald statistics.

The Wild Cluster Bootstrap

The wild cluster bootstrap was first suggested by Cameron, Gelbach, and Miller (REStat, 2008). There are two variants: **restricted** and **unrestricted**, called WCR and WCU. For WCR,

- Estimate model subject to restriction(s) to obtain $\tilde{\boldsymbol{\beta}}$ and the $\tilde{\boldsymbol{u}}_g$.
- Multiply $\tilde{\boldsymbol{u}}_g$ by $v_g^* \sim iid(0,1)$ (often Rademacher, ±1 with prob. 1/2 each). Bootstrap regressand is $\boldsymbol{y}_g^* = \boldsymbol{X}_g\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{u}}_g v_g^*$.
- This preserves the intra-cluster correlations of the scores.
- Equivalently (faster!) generate the bootstrap scores directly as

  $$\boldsymbol{s}_g^{*b} = v_g^{*b}\tilde{\boldsymbol{s}}_g, \quad g = 1, \ldots, G, \tag{7}$$

  where $\tilde{\boldsymbol{s}}_g = \boldsymbol{X}_g^\top \tilde{\boldsymbol{u}}_g$ is the score vector for cluster g evaluated at $\tilde{\boldsymbol{\beta}}$.
- Unlike the pairs cluster bootstrap, the WCR bootstrap uses the actual $\boldsymbol{X}_g$ and thus retains the actual cluster sizes.
- WCR bootstrap can be extremely fast, unless G is very large.

Djogbenou, MacKinnon, and Nielsen (JoE, 2019) establishes the asymptotic validity of the WCR and WCU bootstraps.

For the WCU bootstrap, the unrestricted scores, the $\hat{\boldsymbol{s}}_g$, are used instead of the restricted ones, the $\tilde{\boldsymbol{s}}_g$, in the bootstrap DGP (7).

This means that the bootstrap test statistic must have numerator $\hat{\beta}^*_b - \hat{\beta}$, like the one for the pairs cluster bootstrap.

- In most cases, the WCU bootstrap does not perform as well in finite samples as the WCR one.
- However, it has the advantage that the bootstrap DGP does not depend on the restrictions to be tested.
- Same set of bootstrap samples can be used to perform tests on any restriction or set of restrictions and/or to construct confidence intervals for any coefficient of interest.
- These may be studentized bootstrap intervals, or ones that use bootstrap standard errors.

Classic versions of the WCR and WCU bootstraps are surprisingly inexpensive to compute. One method requires only the matrices $\boldsymbol{X}_g^\top \boldsymbol{X}_g$ and the vectors $\boldsymbol{X}_g^\top \boldsymbol{y}_g$; see MacKinnon (E&S, 2023).

Tricks employed by `boottest` for Stata make even faster computation possible; see Roodman, MacKinnon, Nielsen, and Webb (SJ, 2019).

`boottest` computes WCR bootstrap P values for both t-tests and Wald tests, and also WCR bootstrap confidence intervals.

Performance of the WCR bootstrap

It tends to deteriorate as G becomes smaller, as k increases, and as the clusters become more heterogeneous.

- It tends to over-reject when a very few clusters are much larger than average or have very high partial leverage.
- It tends to under-reject when the number of treated clusters G_1, or control clusters $G_0 = G - G_1$, is very small, even for large G.
- In the latter case, the WCU and Pairs bootstraps tend to over-reject, as do asymptotic tests.

When G is small, the WCR bootstrap encounters an important practical problem.

- For the Rademacher distribution, or any other two-point distribution, the number of possible bootstrap samples is just 2^G.
- Thus number of possible symmetric bootstrap P values is 2^{G-1}.

The 6-point distribution of Webb (CJE, 2023) largely solves this problem, because $6^G >> 2^G$. It works almost as well as Rademacher for most values of G, and sometimes much better when G is small.

Whenever either 2^G (for Rademacher) or 6^G (for six-point) is smaller than B, we can **enumerate** all possible bootstrap samples instead of drawing them at random.

This eliminates simulation randomness. `boottest` uses enumeration by default when B exceeds the number of possible bootstrap samples.

Could also use continuous distributions like $\mathrm{N}(0,1)$ or $\mathrm{U}(-\sqrt{3}, \sqrt{3})$, but they generally do not work as well as discrete distributions.

MacKinnon, Nielsen, and Webb (JAE, 2023) proposes three new variants of the WCR (and also of the WCU) bootstraps.

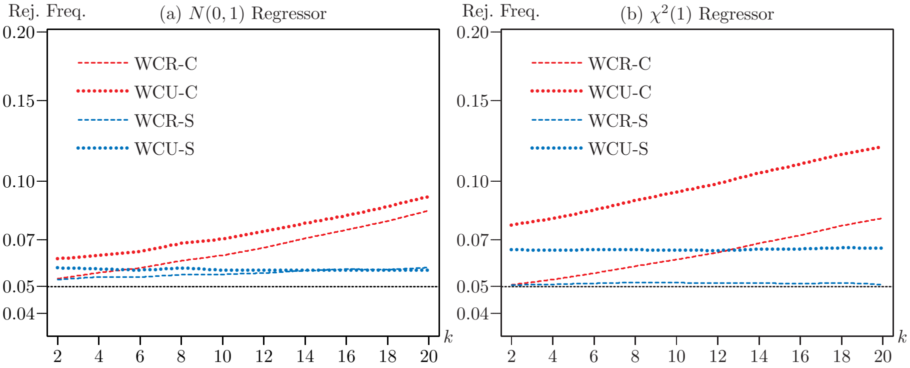
The classic WCR bootstrap uses (7) to generate bootstrap samples and CV_1 to calculate standard errors. It is now called WCR-C.

Two of the new variants use CV_3 and can be expensive.

The WCR-S variant uses CV_1 test statistic, but the bootstrap scores are generated using an alternative to (7) that transforms the restricted scores in a way similar to the $\acute{\boldsymbol{s}}_g$ used by CV_3.

- All three new variants outperform WCR-C in many cases.
- WCR-S and WCU-S are now in `boottest`. Both often work well, with WCU-S sometimes much better than WCU-C.
- Tendency to over-reject increases much more slowly for WCR-S than for WCR-C as k and/or cluster size variation increases.
- Unfortunately, new variants still tend to under-reject severely when the number of treated clusters is small.

Figure 4: Rejection frequencies for bootstrap tests as k varies



- $N = 9600$ and $G = 24$.
- Cluster sizes vary from 130 to 899 ($\gamma = 2$).
- There are 400,000 replications, and $B = 399$.