A More Robust t-Test*

Ulrich K. Müller
Princeton University
Department of Economics
Princeton, NJ, 08544

May 2020

Abstract

Standard inference about a scalar parameter estimated via GMM amounts to applying a t-test to a particular set of observations. If the number of observations is not very large, then moderately heavy tails can lead to poor behavior of the t-test. This is a particular problem under clustering, since the number of observations then corresponds to the number of clusters, and heterogeneity in cluster sizes induces a form of heavy tails. This paper combines extreme value theory for the smallest and largest observations with a normal approximation for the average of the remaining observations to construct a more robust alternative to the t-test. The new test is found to control size much more successfully in small samples compared to existing methods. Analytical results in the canonical inference for the mean problem demonstrate that the new test provides a refinement over the full sample t-test under more than two but less than three moments, while the bootstrapped t-test does not.

Keywords: t-statistic, extreme value distribution, refinement

*I am thankful for very helpful comments and advice from Angus Deaton, Hank Farber, Bo Honoré, Karsten Müller and participants at various workshops. Financial support from the National Science Foundation through grant SES-1919336 is gratefully acknowledged.

1 Introduction

The usual t-test for inference about the mean of a population from an i.i.d. sample is a key building block of statistics and econometrics. Not only does it have many direct applications, but also many other standard forms of inference reduce to the application of a t-test applied to a suitably defined population. For example, consider a linear regression with scalar regressor, $Y_i = X_i\beta + \varepsilon_i$, $\mathbb{E}[X_i\varepsilon_i] = 0$. A test of the null hypothesis $H_0 : \beta = \beta_0$ reduces to a test of $\mathbb{E}[W_i] = 0$ for $W_i = (Y_i - X_i\beta_0)X_i$, and the usual t-statistic computed from the i.i.d. sample W_i amounts to a specific version of the usual heteroskedasticity robust test suggested by White (1980). Under clustering that allows for arbitrary correlations between ε_j for all $j \in \mathcal{C}_i$, $i = 1, \ldots, n$, the effective observations become $W_i = \sum_{j\in\mathcal{C}_i}(Y_j - X_j\beta_0)X_j$. In the presence of additional controls $Y_i = X_i\beta + Z_i'\gamma + \varepsilon_i$, the equivalence to the inference for the mean problem holds approximately after projecting Y_i and X_i off Z_i. This further extends to instrumental variable regression, clustering, and GMM estimation.

The asymptotic validity of standard t-statistic based inference relies on two arguments. First, the law of large numbers implies that the variance estimator in the denominator has negligible estimation error. Second, a central limit theorem applied to the numerator yields approximate normality. Underlying populations with heavy tails are a threat to both. Even if the second moment exists, so that t-statistic based inference is asymptotically justified, large samples might be required before these approximations become accurate.

The effective sample size in empirical work is often considerably smaller than the raw number of observations. This can arise because researchers are interested in inference for smaller subgroups, or because nonparametric kernel estimators are employed that effectively depend only on relatively few observations, or because the relevant variation only concerns fewer observation, such as in studies about rare events. It is also very common for standard errors to be clustered, reducing the effective number of independent observations to the number of clusters, which tends to be only moderately large. What is more, in many applications clusters are of fairly heterogeneous size (think of the 50 states of the U.S., say). Even if none of the variables under study are heavy-tailed, a substantial portion of the parameter sampling variation will then stem from the randomness of the large clusters, inducing a form of heavy-tailedness in the resulting W_i variables. See Section 5.3 for an illustration.

This paper develops an alternative to the t-test that performs more reliably when the underlying population has potentially heavy tails. The focus is exclusively on the case of moderately heavy tails, that is, the first two moments of W_i exist, so that asymptotically, standard t-statistic based inference is valid. The aim is to devise an inference method that does not overreject if the underlying population has moderately heavy tails, without losing much in terms of efficiency if the underlying population has light tails. The theoretical development only concerns the canonical inference for the mean problem, but we show how to adapt the procedure to also obtain more reliable inference about scalar parameters estimated by GMM, including the case with clustering. Note that in such more general contexts, the quality of standard inference is poor if the induced W_i has heavy tails, for which it is neither necessary nor sufficient that the variables under study are individually heavy-tailed.

To describe the key idea, consider the hypothesis test of $H_0 : \mathbb{E}[W_i] = 0$ against $H_a : \mathbb{E}[W_i] \neq 0$ based on an i.i.d. sample W_i, $i = 1, \ldots, n$, from a population W with cumulative distribution function F. For expositional ease, suppose that F has a thin left tail, but a potentially heavy right tail. For some given k, let $\mathbf{W}^R = (W_1^R, W_2^R, \cdots, W_k^R)'$ be the k largest order statistics, with W_1^R the sample maximum. Conditional on $\mathbf{W}^R$, the remaining "small" observations W_i^s, $i = 1, \ldots, n-k$ are i.i.d. draws from the truncated distribution with c.d.f. $F(w)/F(W_k^R)$ for $w \leq W_k^R$. The mean of this truncated distribution under H_0 is no longer zero, however, but is given by $-m(W_k^R) < 0$, where

$$m(w) = -\mathbb{E}[W|W \leq w] = \frac{\mathbb{P}(W > w)\mathbb{E}[W|W > w]}{1 - \mathbb{P}(W > w)}.$$

Note that $m(w)$ for w large is determined by the properties of F in its right tail.

The idea now is to apply three asymptotic approximations. First, invoke standard extreme value theory to obtain an approximation for the distribution of $\mathbf{W}^R$ in terms of a (joint) extreme value distribution governed by three parameters describing location, scale and shape. Second, apply the central limit theorem to the conditional i.i.d. sample of remaining observations W_i^s from the truncated (and hence no longer heavy-tailed) distribution to argue that $(n-k)^{-1}\sum_{i=1}^{n-k} W_i^s$ is approximately normal with mean $-m(W_k^R)$ under H_0 (and arbitrarily different mean under the alternative H_a). Third, by the same arguments that justify extreme value theory, obtain an approximation to $m(w)$ in terms of the same three parameters that govern the distributional approximation of $\mathbf{W}^R$.

These approximations lead to a *parametric* approximate joint model of $k+1$ statistics: $\mathbf{W}^R$ is jointly extreme value, and $(n-k)^{-1}\sum_{i=1}^{n-k} W_i^s$ is normally distributed with a mean that, under H_0, depends on W_k^R, and the parameters of the extreme value distribution. For given k, this is a small sample nonstandard parametric testing problem, and one can construct tests that are exactly valid under the approximate parametric model. Specifically, we apply computational techniques similar to those developed in Elliott, Müller, and Watson (2015) to determine a powerful that is of level α in this parametric model. Once the test is applied to the original mean testing problem, it is not longer of level α by construction. But the explicit modelling of the potentially moderately heavy tail via extreme value theory might improve performance over the usual t-test.

The main theoretical result of this paper corroborates this conjecture by considering higher order improvements for populations with finite variance, but that do not possess a third moment, and for which extreme value theory applies. We consider asymptotics in which k is a fixed number that does not vary as a function of n. In this way, the asymptotics reflect that moderately large samples only contain limited information about the tail properties of the underlying population. We show that the approximation error of the parametric model for k fixed induces an error in the rejection probability in the mean testing problem that converges to zero faster than the error in rejection probability of the usual t-test. In that sense, the new approach yields a refinement over the usual t-test and provides theoretical support for the usefulness of the new perspective.

A natural alternative to obtain more accurate approximations is to consider the bootstrap. Bloznelis and Putter (2003) show that the percentile-t bootstrap provides a refinement whenever the underlying population has at least three moments. A second, apparently new theoretical result shows that the bootstrap *does not* provide a refinement when the underlying population has between two and three moments.

The new method readily generalizes to the case where both tails are potentially heavy. The approximate parametric model then consists of $2k + 1$ statistics, with k joint extreme value observations from the left tail governed by three parameters, k extreme value observations from the right tail governed by their own three parameters, and the conditionally normal average of the middle observations. Since in most applications, there are no compelling reasons to assume any constraints between the properties of the left and right tail, the approximate parametric problem is thus indexed by a six dimensional nuisance parameter. We use a version of the the algorithm of Elliott, Müller, and Watson (2015) to numerically determine a powerful test in this parametric problem for selected values of k. The large nuisance parameter space turns this into a major computational challenge. Once a powerful valid test has been determined, however, applying it in practice is entirely straightforward and does not pose any significant computational burden. This includes its use to obtain more reliable inference about scalar parameters estimated by GMM with potentially clustered errors; see Section 4.4 for details.

Our preferred default method uses $k = 8$ and is appropriate when the sample consists of at least 50 independent clusters or observations.[1] The tests were determined for various significance levels, enabling the construction of confidence intervals at the 90%, 95% and 99% level via test inversion, and the computation of p-values. Corresponding tables and STATA code is provided in the replication files.

Monte Carlo simulations show that the new approach leads to much better size control in moderately large samples compared to existing methods, at fairly small cost in terms of average confidence interval length for thin-tailed populations. This is true in the canonical inference for the mean case, as predicted by the theory, but also when comparing two means, and for inference about regression coefficients under clustering. In one design, the clusters are Metropolitan Statistical Areas, which are fairly heterogeneous in size. As discussed above, this heterogeneity induces the resulting W_i to be quite heavy-tailed, which leads to poor performance of standard cluster robust inference. A moderately large number of heterogenous clusters (say, no more than 100 or 200) is quite common in applied economics, making the new approach a potentially attractive alternative in such settings.

The remainder of the paper is organized as follows. The next section discusses the literature and reviews known results about extreme value theory and the approximate distribution of t-statistics. Section 3 more carefully describes the new theoretical results in the inference for the mean problem. Section 4 provides details on the construction of the new test. Small sample simulation results are reported in Section 5. Section 6 concludes.

2 Background

2.1 Relationship to Literature

The new method "robustifies" the usual t-test in the sense of providing more reliable inference under moderately heavy tails. The classic robustness literature (see, for instance, Huber (1996) for an overview) is based on a very different notion; in this literature, it is assumed that the observations are contaminated, with a small fraction not stemming from the population of interest. This in turn raises the question what type of estimands can still

[1] We also provide an alternative, even more robust test for $k = 4$ that is applicable to samples with as few as 25 indepedent clusters or observation.

be reliably learned about, and how to do so efficiently. In contrast, the estimands considered here are defined relative to the distribution that generated the data. Which of these views is appropriate depends on the application, and in particular whether relatively extreme observations are part of the population of interest, or rather are induced by measurement errors. It is also possible to combine the approaches, such as applying the new test in a regression of winsorized variables (with the estimand defined relative to a winsorized population).

Even under the assumption that the data is entirely uncontaminated, informative inference requires assumptions beyond the existence of moments: The classic impossibility result of Bahadur and Savage (1956) shows that one cannot learn about the population mean from i.i.d. samples of any size, even if all moments are assumed to exist. The substantial assumption pursued here is that the population tails are such that extreme value theory provides reasonable approximations. This effectively amounts to an assumption that the tails of the underlying distribution are approximately (generalized) Pareto. Given the theoretical prevalence and empirical success of extreme value theory for learning about the tail of distributions (for overviews and references, see, for instance, Embrechts, Klüppelberg, and Mikosch (1997) or de Haan and Ferreira (2007)), this seems a reasonably general starting point, especially given that some assumption must be made. What is more, the approximate Pareto tail is only imposed in the extreme tail with approximate mass of k/n for k fixed, which is enough to ensure that the largest (and smallest) k observations are governed by extreme value theory.

Müller and Wang (2017) pursue this "fixed-k" approach for the purpose of inference about tail properties, such as extreme quantiles. In contrast, the remaining literature on the modelling of tails considers asymptotics where $k = k_n$ diverges with the sample size. In large samples, k_n diverging asymptotics allow for consistent estimation of tail properties, at least pointwise for a fixed population. In practice, though, the approximations generated from k_n diverging asymptotics are not very useful for, say, samples of size $n = 50$ or $n = 100$, as there are only a handful of observations that can usefully be thought of as stemming from the tail, so that any approximation that invokes "consistency" of tail property estimators becomes misleading.

The separate analysis of the largest and remaining terms of a sum of independent random variables goes back to at least Csörgö, Haeusler, and Mason (1988); also see Zaliapin, Kagan, and Schoenberg (2005), Kratz (2014) and Müller (2019). The relatively closest precursors to this work are Peng (2001, 2004) and Johansson (2003). These authors are concerned with inference about the mean from an i.i.d. sample under very heavy tails, that is, the underlying population has less than two moments. For such populations, the usual t-statistic does not converge to a normal distribution. Peng (2001, 2004) and Johansson (2003) suggest estimating the contribution of the two tails to the overall mean by consistently estimating the tail Pareto parameters using the smallest and largest k_n observations, with k_n diverging, and combining those estimates with the estimate of the mean of the remaining middle observations.

Another approach to overcome the Bahadur and Savage (1956) impossibility result is to assume a bounded support, with known bounds. Romano (2000), Schlag (2007) and Gossner and Schlag (2013) derive corresponding methods.

2.2 Extreme Value Theory

Let $W_1^R \geq W_2^R \geq ... \geq W_k^R$ denote the largest k order statistics from an i.i.d. sample from a population with distribution F. Suppose the right tail of F is approximately Pareto in the sense that for some scale parameter $\sigma > 0$ and tail index $\xi > 0$

$$\lim_{w \to \infty} \frac{1 - F(w)}{(w/\sigma)^{-1/\xi}} = 1 \tag{1}$$

so that the second moment of W exists if and only if $\xi < 1/2$. Then W is in the maximum domain of attraction of the Fréchet limit law

$$n^{-\xi} W_1^R \Rightarrow \sigma X_1 \tag{2}$$

where $X_1^{-1/\xi} \sim E_1$ with E_1 an exponentially distributed random variable.

As is well known (see, for instance, Theorem 2.8.2 of Galambos (1978)), (2) implies that extreme value theory also holds jointly for the first k order statistics

$$n^{-\xi}\mathbf{W}^R = n^{-\xi} \begin{pmatrix} W_1^R \\ \vdots \\ W_k^R \end{pmatrix} \Rightarrow \sigma \mathbf{X} = \sigma \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}. \tag{3}$$

The distribution of $\mathbf{X}$ satisfies $\{X_j^{-1/\xi}\}_{j=1}^k \sim \{\sum_{l=1}^j E_l\}_{j=1}^k$, where E_l are i.i.d. exponential random variables.

Since the new theoretical results of this paper concern rates of convergence, a suitable strengthening of the approximate Pareto tail assumption (1) is needed. Falk, Hüsler, and Reiss (2004) define the δ-neighborhood of the Pareto distribution with index ξ as follows.

Condition 1 *For some $\delta, w_0 > 0$, F admits a density for $w > w_0$ of the form*

$$f(w) = (\xi\sigma)^{-1}(\frac{w}{\sigma})^{-1/\xi-1}(1+h(w)) \tag{4}$$

with $|h(w)|$ uniformly bounded by $Cw^{-\delta/\xi}$ for some finite C.

Theorem 5.5.5 of Reiss (1989) shows that under Condition 1, (3) provides accurate approximations in the sense that

$$\sup_B |\mathbb{P}(n^{-\xi}\mathbf{W}^R \in B) - \mathbb{P}(\sigma\mathbf{X} \in B)| = O(n^{-\delta}) \tag{5}$$

for $\delta \leq 1$, where the supremum is taken over all Borel sets $B \subseteq \mathbb{R}^k$.

Many heavy-tailed distributions satisfy Condition 1: for the right tail of a student-t distribution with ν degrees of freedom, $\xi = 1/\nu$ and $\delta = 2\xi$, for the tail of a Fréchet or generalized extreme value distribution with parameter α, $\xi = 1/\alpha$ and $\delta = 1$, and for an exact Pareto tail, δ may be chosen arbitrarily large. But there also exist heavy-tailed distributions in the domain of attraction of a Fréchet limit law that do not satisfy Condition 1, such as density of the form (4) with $h(x) = 1/\log(1+x)$, for example. Under some additional regularity conditions, Theorem 3.2 of Falk and Marohn (1993) shows Condition 1 to be necessary to obtain an error rate of extreme value approximations of order $n^{-\delta}$ for $\delta > 0$. Roughly speaking, Condition 1 thus formalizes the assumption that extreme value theory provides accurate approximations.

2.3 Approximations to the t-Statistic

Let T_n be the usual t-statistic computed from an i.i.d. sample $W_1, \ldots, W_n$, where $W_i \sim W$. If $\mathbb{E}[W] = 0$ and $\mathbb{E}[W^2] < \infty$, then $T_n \Rightarrow \mathcal{N}(0,1)$. A seminal paper by Bentkus and Götze (1996) establishes a bound on the rate of this convergence which does not require the third moment of W to exist. In particular, Bentkus and Götze (1996) show that for some $C > 0$ that does not depend on F, and $\mathbb{E}[W^2] = 1$,

$$\sup_t |\mathbb{P}(T_n < t) - \Phi(t)| \leq C\mathbb{E}[W^2\mathbf{1}[|W| > n^{1/2}]] + Cn^{-1/2}\mathbb{E}[|W|^3\mathbf{1}[|W| \leq n^{1/2}]] \tag{6}$$

where $\Phi(t) = \mathbb{P}(Z < t)$, $Z \sim \mathcal{N}(0,1)$. (The explicit claim of uniformity with respect to F is made in Bentkus, Bloznelis, and Götze (1996).) This result is key for the new theoretical results of this paper.

To put (6) into perspective, recall that the classic Berry-Esseen bound shows a $n^{-1/2}$ rate of the approximation $Z_n = n^{-1/2}\sum_{i=1}^{n} W_i/\sqrt{\mathbb{E}[W^2]} \Rightarrow \mathcal{N}(0,1)$, with a constant that involves the third moment of W. If $\mathbb{E}[|W|^3]$ does not exist, then the classic Berry-Esseen bound is inapplicable. Theorem 5 of Petrov (1975) provides the analogue of (6) with T_n replaced by Z_n. The contribution of Bentkus and Götze (1996) is thus to show that the random norming inherent in the t-statistic does not affect Petrov's (1975) result.

Subsequent research by Hall and Wang (2004) provides a sharp bound on the rate of convergence: If $\mathbb{E}[W^2] < \infty$, their results imply that

$$\frac{\sup_t |\mathbb{P}(T_n < t) - \Phi(t)|}{n\mathbb{P}(|W| > n^{-1/2}) + n^{1/2}|\mathbb{E}[WA_n]| + n^{-1/2}\mathbb{E}[|W|^3 A_n] + n^{-1}\mathbb{E}[|W|^4 A_n]} \tag{7}$$

with $A_n = \mathbf{1}[|W| \leq n^{1/2}]$ is bounded away from zero and infinity uniformly in n.

A final relevant result from the literature concerns the bootstrap approximation to the distribution of the t-statistic. Let $\mathbf{W} = (W_1, \ldots, W_n)$, and let T_n^* be a bootstrap draw of T_n from the demeaned empirical distribution of W_i, conditional on $\mathbf{W}$. Bloznelis and Putter (2003) show that if F is non-lattice and $\mathbb{E}[|W|^3] < \infty$, then

$$\sup_t |\mathbb{P}(T_n^* < t|\mathbf{W}) - \mathbb{P}(T_n < t)| = o(n^{-1/2}) \text{ a.s.} \tag{8}$$

while, for $\mathbb{E}[W^3] \neq 0$, $\liminf_{n\to\infty} n^{1/2} \sup_t |\mathbb{P}(T_n < t) - \Phi(t)| > 0$. In other words, as long as W has finite non-zero third moment, the error in the bootstrap approximation to the distribution of the t-statistic is of smaller order than the normal approximation, and the bootstrap provides a refinement over the usual t-statistic.

3 New Theoretical Results

To ease exposition, we focus in this section on the case where the left tail of W is light, as in the introduction. The analogous results also hold when both tails are moderately heavy with tail index smaller than 1/2; we provide an analogue of Theorem 2 in Appendix A.2.

3.1 Properties of Bootstrapped t-Statistic under $1/3 < \xi < 1/2$

Theorem 1 *Suppose (1) holds for* $1/3 < \xi < 1/2$, *and* $\int_{-\infty}^{0} |w|^3 dF(w) < \infty$. *Then*

(a) $\liminf_{n\to\infty} n^{1/(2\xi)-1} \sup_t |\mathbb{P}(T_n < t) - \Phi(t)| > 0$ *and*

(b) $n^{3(1/2-\xi)} \sup_t |\mathbb{P}(T_n^* < t|\mathbf{W}) - \Phi(t)| = O_p(1)$.

Since $3(1/2 - \xi) > 1/(2\xi) - 1$ for $1/3 < \xi < 1/2$, the triangle inequality implies that $\sup_t |\mathbb{P}(T_n^* < t|\mathbf{W}) - \mathbb{P}(T_n < t)| = O_p(n^{1-1/(2\xi)})$, so Theorem 1 shows that the bootstrap does not provide a refinement if the underlying population has between two and three moments, at least as long as the population has an approximate Pareto tail. This result is apparently new, but it is not difficult to prove. From Markov's inequality, $\int_{-\infty}^0 |w|^3 dF(w) < \infty$ implies that also $|W|$ has a Pareto tail with index $1/3 < \xi < 1/2$ in the sense of (1). Part (a) now simply follows from evaluating the sharp bound on the rate of convergence in (7). Part (b) follows from applying the Bentkus and Götze (1996) bound (6) to the empirical distribution of $\bar{W}_i = W_i - n^{-1}\sum_{j=1}^n W_j$: By (3), $n^{-\xi}\max_i |W_i|$ converges in distribution, so $\max_i |\bar{W}_i| = O_p(n^{\xi})$. Since $\xi < 1/2$, this implies $n^{-1}\sum_{i=1}^n \bar{W}_i^2 \mathbf{1}[|\bar{W}_i| > \sqrt{n}] \xrightarrow{p} 0$. Furthermore, $|W_i|^3$ has a Pareto tail of index $3\xi > 1$. Thus $n^{-3\xi}\sum_{i=1}^n |\bar{W}_i|^3$ converges in distribution to a stable distribution (see, for instance, LePage, Woodroofe, and Zinn (1981), who elucidate the connection between extreme value theory and stable limit laws), so that $n^{-3/2}\sum_{i=1}^n |\bar{W}_i|^3 = O_p(n^{3\xi-3/2})$, and the result follows.

The existence of three moments, corresponding to a tail index of $\xi < 1/3$, is necessary to obtain the first term of an Edgeworth expansion that underlies the proof of Bloznelis and Putter (2003). More intuitively, recall that under $\xi < 1/3$, the Berry-Esseen bound shows that the central limit theorem has an approximation quality of order $n^{-1/2}$. Now under (1), $\mathbb{P}(W_1^R > \sigma\sqrt{n})$ is of order $(1 - n^{-1/(2\xi)})^n \approx n^{1-1/(2\xi)}$. Thus, for $\xi > 1/3$, the largest observation is of order $\sqrt{n}$ with a probability that is an order of magnitude larger than $n^{-1/2}$. Non-normal observations of order $\sqrt{n}$ are not negligible in the central limit theorem, so the rare large values of W_1^R under $\xi > 1/3$ are responsible for a deterioration of the central limit theorem approximation compared to the $\xi < 1/3$ case (cf. Hall and Wang (2004)). But from (2) W_1^R is of order n^{ξ} in nearly all samples, so the bootstrap approximation misses this effect, and systematically underestimates the heaviness of the tail.

3.2 New Asymptotic Approximation

We first discuss the approximate parametric problem in more detail. Under the Pareto tail assumption (1), we find from a straightforward calculation that for large w, $m(w) = -\mathbb{E}[W|W \leq w] \approx \sigma^{1/\xi}w^{1-1/\xi}/(1-\xi)$. Let $s_n^2 = (n-k)^{-1}\sum_{i=1}^{n-k}(W_i^s - \bar{W}^s)^2$ be the usual variance estimator from the $n-k$ smallest observations. With k fixed, s_n^2 still converges in probability to the unconditional variance of W, $s_n^2 \xrightarrow{p} \mathrm{Var}[W]$. Since the ultimate test we derive is scale invariant, there is no loss of generality of normalizing $\mathrm{Var}[W] = 1$. From the convergence to the joint extreme value distribution in (3), $n^{-\xi}\mathbf{W}^R \overset{a}{\sim} \sigma\mathbf{X}$, where we write $\overset{a}{\sim}$ for "is approximately distributed as." Furthermore, under local alternatives $\mathbb{E}[W] = n^{-1/2}\mu$, the t-statistic

$$T_n^s = \frac{\sum_{i=1}^{n-k} W_i^s}{\sqrt{(n-k)s_n^2}}$$

computed from $\{W_i^s\}_{i=1}^{n-k}$ is approximately normal with mean $\mu - n^{-1/2}m(W_k^R) \approx \mu - n^{-1/2}\sigma^{1/\xi}(W_k^R)^{1-1/\xi}/(1-\xi)$. Combining these two approximations yields

$$\mathbf{Y}_n = \begin{pmatrix} \mathbf{W}^R/\sqrt{(n-k)s_n^2} \\ T_n^s \end{pmatrix} \overset{a}{\sim} \begin{pmatrix} \eta_n \mathbf{X} \\ Z + \mu - \eta_n \frac{1}{1-\xi} X_k^{1-1/\xi} \end{pmatrix} = \mathbf{Y}_n^* \tag{9}$$

with $\eta_n = \sigma n^{-(1/2-\xi)}$ and $Z \sim \mathcal{N}(0,1)$ independent of $\mathbf{X}$. The first k elements of $\mathbf{Y}_n$ are the largest k order statistics divided by the denominator of the $k+1$ element T_n^s, so that $\mathbf{Y}_n$ is invariant to changes in scale $\{W_i\}_{i=1}^n \to \{cW_i\}_{i=1}^n$ for $c > 0$. The approximate parametric model on the right-hand side of (9) treats these as jointly extreme value with scale η_n and tail index ξ, and conditionally normally distributed with some (negative) mean that is a function of X_k and the parameters η_n and tail index ξ under $\mu = 0$.

As discussed in the introduction, the central idea of this paper is to use the parametric model $\mathbf{Y}_n^*$ to determine a test $\varphi : \mathbb{R}^{k+1} \mapsto \{0,1\}$ of $H_0 : \mu = 0$ that satisfies $\mathbb{E}[\varphi(\mathbf{Y}_n^*)] \leq \alpha$ by construction for all $\xi < 1/2$, at least for all $n \geq n_0$ and some appropriate upper bounds on σ. We discuss the construction of such tests in the next section. Any such test φ may then be applied to the left-hand side of (9), $\varphi(\mathbf{Y}_n)$, to test $H_0 : \mathbb{E}[W] = 0$ from the observations $W_1, \ldots, W_n$.

Our main theoretical result is the following.

Theorem 2 *For $k > 1$, let $r_k(\xi) = \frac{3(1+k)(1-2\xi)}{2(1+k+2\xi)}$. Suppose Condition 1 holds with $\delta \geq r_k(\xi)$, $\int_{-\infty}^0 |w|^p dF(w) < \infty$ for all $p > 0$ and that $\varphi : \mathbb{R}^{k+1} \mapsto \{0,1\}$ is such that for some finite m_φ, $\varphi : \mathbb{R}^{k+1} \mapsto \{0,1\}$ can be written as an affine function of $\{\varphi_j\}_{j=1}^{m_\varphi}$, where each φ_j is of the form*

$$\varphi_j(\mathbf{y}, y_0) = \mathbf{1}[\mathbf{y} \in \mathcal{H}_j]\mathbf{1}[y_0 \leq b_j(\mathbf{y})]$$

with $b_j : \mathbb{R}^k \mapsto \mathbb{R}$ a Lipschitz continuous function and $\mathcal{H}_j$ a Borel measurable subset of $\mathbb{R}^k$ with boundary $\partial\mathcal{H}_j$. For $\mathbf{u} = (1, u_2, \ldots, u_k)' \in \mathbb{R}^k$ with $1 \geq u_2 \geq u_3 \geq \ldots \geq u_k$, let $\mathcal{I}_j(\mathbf{u}) = \{s > 0 : s\mathbf{u} \in \partial\mathcal{H}_j\}$. Assume further that for some $L > 0$, and Lebesgue almost all $\mathbf{u}$, $\mathcal{I}_j(\mathbf{u})$ contains at most L elements in the interval $[L^{-1}, \infty)$.

Then under $H_0 : \mu = 0$, for $\frac{1+k}{1+3k} < \xi < 1/2$ and any $\epsilon > 0$

$$|\mathbb{E}[\varphi(\mathbf{Y}_n)] - \mathbb{E}[\varphi(\mathbf{Y}_n^*)]| \leq Cn^{-r_k(\xi)+\epsilon}.$$

Recall from Theorem 1 (a) above that the rate of convergence of the normal approximation to the distribution of the t-statistic is $n^{1/(2\xi)-1}$. Since for $\frac{1+k}{1+3k} < \xi < 1/2$, $r_k(\xi) > 1/(2\xi) - 1$, the theorem shows that the difference in the rejection rates of φ in the parametric model $\mathbb{E}[\varphi(\mathbf{Y}_n^*)]$ and in the original inference for the mean problem $\mathbb{E}[\varphi(\mathbf{Y}_n)]$ is of smaller order. In this sense, the new approximation provides a refinement for underlying populations that have between two and three moments.

The Bentkus and Götze (1996) bound (6) implies that conditional on $\mathbf{W}^R$, T_n^s is well approximated by a standard normal distribution, since the W_i^s form an i.i.d. sample from a truncated distribution with less heavy tails compared to the original population. Furthermore, under Condition 1, it follows from (5) that the distribution of $\mathbf{W}^R/\sqrt{(n-k)}$ is well approximated by the distribution of $\eta_n \mathbf{X}$. The difficulty in the proof of Theorem 2 arises from the presence of s_n^2 in the scale normalization of $\mathbf{W}^R$ in $\mathbf{Y}_n$. While it is easy to show that $s_n^2 \xrightarrow{p} \mathrm{Var}[W] = 1$, the proof of Theorem 2 requires this convergence to be sufficiently fast, and this complication leads to the presence of k in the rate r_k (intuitively, larger k lead to more truncation, so s_n^2 is estimated from a distribution with a lighter tail).

Note that the usual full sample t-statistic of $H_0 : \mathbb{E}[W] = 0$, T_n, is approximated in terms of $\mathbf{Y}_n = (Y_{1n}, \ldots, Y_{kn}, T_n^s)'$ by

$$\tilde{T}_n = \frac{T_n^s + \sum_{i=1}^k Y_{in}}{\sqrt{1 + \sum_{i=1}^k Y_{in}^2}} \tag{10}$$

up to a $O_p(n^{-1/2})$ term. Under (9), the distribution of $\tilde{T}_n$ is approximated by

$$T(\mathbf{Y}_n^*) = \frac{Z + \mu - \eta_n \frac{1}{1-\xi} X_k^{1-1/\xi} + \eta_n \sum_{i=1}^k X_i}{\sqrt{1 + \eta_n^2 \sum_{i=1}^k X_i^2}}. \tag{11}$$

Application of Theorem 2 with $\varphi(\mathbf{Y}_n^*) = \mathbf{1}[T(\mathbf{Y}_n^*) < t]$ shows that this approximation has a faster rate of convergence compared to the usual standard normal approximation. Müller (2019) has shown that one can combine extreme value theory to improve the rates of approximation to sums of i.i.d. random variables compared to the central limit theorem under $\xi > 1/3$. On implication of Theorem 2 is thus a corresponding result for the case of self-normalized sums (10) and (11).

In principle, one could use this implication also to construct an alternative test φ that simply amounts to a t-test with appropriately increased critical value to ensure size control in the approximate model, $\mathbb{E}[\varphi(\mathbf{Y}_n^*)] \leq \alpha$. This is woefully inefficient, however, since the much larger critical value is only needed for samples where ξ and η_n are large, which would defeat the objective of obtaining a test that remains close to efficient for populations with thin tails.

4 Construction of a New Test

4.1 Generalized Parametric Model

To obtain accurate approximations in small samples also for potentially thin-tailed distributions, it makes sense to extend the parametric approximation to populations with an approximate generalized Pareto tail. The c.d.f. F of such populations satisfies

$$F(w) \approx 1 - (1 + \xi(w/\sigma - \nu))^{-1/\xi}, \quad \xi \in (-\infty, 1/2] \tag{12}$$

for all w close to the upper bound of the support of F, and here and in the following, expressions of the form $(1 + \xi x)^{-1/\xi}$ are understood to equal e^{-x} for $\xi = 0$. The Pareto tail assumption (1) of Section 2.2 is recovered as a special case for $\xi > 0$ with $\nu = 1/\xi$ and σ rescaled by ξ.[2]

Assumption (12) accommodates infinite support thin-tailed distributions, such as the exponential distribution, with $\xi = 0$, as well as distributions with finite upper bound on their support, such as the uniform distribution with $\xi = -1$. From the seminal work of Balkema and de Haan (1974) and Pickands (1975) (also see Theorem 5.1.1 of Reiss (1989)), it follows that under an appropriate formalization of (12), there exist real sequences a_n and κ_n such that

$$\frac{\mathbf{W}^R}{a_n} - \kappa_n \Rightarrow \mathbf{X} = (X_1, \ldots, X_k)' \tag{13}$$

is (jointly) generalized extreme value distributed, so that $\{(\xi X_j + 1)^{-1/\xi}\}_{j=1}^k \sim \{\sum_{l=1}^j E_l\}_{j=1}^k$ with E_l i.i.d. exponential random variables. If F is exactly generalized Pareto in the sense

[2] To avoid notational clutter, this section redefines some of the notation previously introduced in Sections 2 and 3 as appropriate for the more general model.

of (12), then Corollary 1.6.9 of Reiss (1989) implies

$$\left\{ \left(\xi \left(\frac{W_j^R}{a_n} - \kappa_n \right) + 1 \right)^{-1/\xi} \right\}_{j=1}^{k} \sim \left\{ \left(\frac{n}{\sum_{l=1}^{n+1} E_l} \right) \sum_{l=1}^{j} E_l \right\}_{j=1}^{k} \tag{14}$$

with $a_n = \sigma n^{\xi}$ and $\xi \kappa_n = 1 + n^{-\xi}(\xi\nu - 1)$, so that $\sum_{l=1}^{n+1} E_l/n \approx 1$ is the only approximation involved in (13). If only the right tail of F of mass $p_R > 0$ is exactly generalized Pareto, then (14) holds conditionally on the event $\sum_{l=1}^{j} E_l / \sum_{l=1}^{n+1} E_l \leq n p_R$, whose probability is larger than 99% for $k = 8$ and all $p_R \geq 16/n$, $n \geq 50$.

Under (12) and (13), from the same logic that led to (9), we obtain the approximate model

$$\mathbf{Y}_n = \begin{pmatrix} \mathbf{W}^R / \sqrt{(n-k)s_n^2} \\ T_n^s \end{pmatrix} \stackrel{a}{\sim} \begin{pmatrix} \eta_n(\mathbf{X} + \kappa_n \mathbf{e}) \\ Z + \mu - \eta_n m^*(\mathbf{X}, \kappa_n, \xi) \end{pmatrix} = \mathbf{Y}_n^* \tag{15}$$

where $\mathbf{e}$ is a $k \times 1$ vector of ones, $\eta_n = n^{-1/2} a_n$ and

$$m^*(\mathbf{X}, \kappa_n, \xi) = (1 + \xi X_k)^{-1/\xi} \left(\kappa_n + \frac{1 + \xi X_k}{\xi(1-\xi)} - \frac{1}{\xi} \right).$$

With $\kappa_n \to 1/\xi$ for $\xi > 0$, it is tempting to employ the additional approximation $\kappa_n = 1/\xi$ to eliminate the location parameter in (15), and this is implicitly applied in standard extreme value theory as reviewed in Section 2.2. However, unless n is very large, this leads to a considerably deterioration of the approximation in (13), and hence (15), so we do not do so in the following.

For practical implementations it is important to allow for the possibility that both tails are potentially moderately heavy. This is straightforward under an assumption that also the left-tail of F is approximately generalized Pareto in the sense of (12): Let $\mathbf{W}^L$ be the set of smallest k order statistics. Further let W_i^m be the $n - 2k$ "middle" order statistics $k+1, \ldots, n-k-1$, and let s_n^2 be the sample variance of W_i^m. Then in analogy to (15),

$$((n-2k)s_n^2)^{-1/2} \begin{pmatrix} \mathbf{W}^R \\ -\mathbf{W}^L \\ \sum_{i=1}^{n-2k} W_i^m \end{pmatrix} \stackrel{a}{\sim} \begin{pmatrix} \eta_n^R(\mathbf{X}^R + \kappa_n^R \mathbf{e}) \\ \eta_n^L(\mathbf{X}^L + \kappa_n^L \mathbf{e}) \\ Z - \eta_n^R m^*(\mathbf{X}^R, \kappa_n^R, \xi^R) + \eta_n^L m^*(\mathbf{X}^L, \kappa_n^L, \xi^L) \end{pmatrix} = \mathbf{Y}_n^*. \tag{16}$$

where $\mathbf{X}^L$ and $\mathbf{X}^R$ are independent and generalized extreme value distributed with tail index ξ^L and ξ^R, respectively, and independent of $Z \sim \mathcal{N}(0,1)$.

The scale and location parameters η_n and κ_n in this generalized model depend on the known sample size n. But they also depend on the tail parameters of the underlying population: Recall that $\eta_n = n^{-1/2}a_n = \sigma n^{\xi-1/2}$ in (14). With σ unknown, this product can in principle take on any positive value, even with (n,ξ) known, and the same holds for the parameter κ_n. For this reason, we will now drop the index n in the nuisance parameter $\theta = (\kappa^L,\eta^L,\xi^L,\kappa^R,\eta^R,\xi^R) \in \Theta_0$ and in the $2k+1$ dimensional observation $\mathbf{Y}^* = \mathbf{Y}^*_n$ from the approximate parametric model in (16). In this notation, the problem becomes the construction of a powerful test $\varphi(\mathbf{Y}^*)$ of $H_0 : \mu = 0$ against $H_a : \mu \neq 0$ that satisfies

$$\sup_{\theta\in\Theta_0} \mathbb{E}_\theta[\varphi(\mathbf{Y}^*)] \leq \alpha, \tag{17}$$

where

$$\mathbf{Y}^* = \begin{pmatrix} \eta^R(\mathbf{X}^R + \kappa^R\mathbf{e}) \\ \eta^L(\mathbf{X}^L + \kappa^L\mathbf{e}) \\ Z + \mu - \eta^R m^*(\mathbf{X}^R,\kappa^R,\xi^R) + \eta^L m^*(\mathbf{X}^L,\kappa^L,\xi^L) \end{pmatrix} = \begin{pmatrix} \mathbf{Y}^R \\ \mathbf{Y}^L \\ Y_0 \end{pmatrix} \tag{18}$$

and $\mathbf{Y}^J = (Y_1^J,\ldots,Y_k^J)'$ for $J \in \{L,R\}$.

From the representation of the joint generalized extreme value distribution in terms of i.i.d. exponentially distributed random variables, it follows that the density of $\mathbf{Y}^*$ is given by

$$f(\mathbf{y}^*|\theta,\mu) = f_T(\mathbf{y}^R|\theta^R) f_T(\mathbf{y}^L|\theta^L)\phi(y_0 - \mu + M^*(\mathbf{y}^R,\theta^R) - M^*(\mathbf{y}^L,\theta^L)) \tag{19}$$

where $\theta^J = (\kappa^J,\eta^J,\xi^J)$, ϕ is the density of a standard normal, $M^*(\mathbf{y},\theta^S) = \eta m^*(\mathbf{y}/\eta - \mathbf{e}\kappa,\kappa,\xi)$, and f_T is the "tail" density

$$f_T(\mathbf{y}|\theta^S) = \mathbf{1}[1+\xi x_k > 0]\mathbf{1}[1+\xi x_1 > 0]\eta^{-k}\exp\left[-(1+\xi x_k)^{-1/\xi} - (1+\xi^{-1})\sum_{i=1}^k \log(1+\xi x_i))\right]$$

with $\theta^S = (\kappa,\eta,\xi)$ the parameter of a "single tail" and $x_i = y_i/\eta - \kappa$ in obvious notation.

4.2 Nuisance Parameter Space

Allowing for arbitrary values of the location and scale parameters in the testing problem (17) is not fruitful: An unreasonably large nuisance parameter space Θ_0 leads to excessively conservative inference, and it renders the computational determination of powerful tests prohibitively difficult. With that in mind, in the default construction, we consider a nuisance parameter space Θ_0 that is partially motivated by a desire to obtain good size control in samples from an exact Pareto population when $n \geq n_0 = 50$. In the description of Θ_0, we refer to the extreme value approximation extended to the most extreme n_0 observations, $\{W_i^J\}_{i=1}^{n_0} \overset{a}{\sim} \{Y_i^J\}_{i=1}^{n_0}$ for $J \in \{L, R\}$. As noted in (14), this remains an good approximation for an exact generalized Pareto population even if $n = n_0$.

In particular, for $J \in \{L, R\}$, we impose

(a). $\xi^J < 1/2$

(b). $\kappa^J \leq 1/\xi^J$ for $\xi^J > 0$

(c). $\sum_{i=1}^{n_0} \mathbb{E}[Y_i^J] \geq 0$

(d). $\sum_{i=k+1}^{n_0-k} \mathbb{E}[(Y_i^J)^2] \leq 2.$

Restriction (a) imposes that the tails are such that at least two moments exists. Restriction (b) says that any potential tail shift is only inward relative to the Pareto default. Note that arbitrarily large inward shifts are incompatible with the population having mean zero. Restriction (c) puts a corresponding lower bound on the inward shift: For the right tail, it requires that the sum of the largest n_0 observations still has positive mean. To motivate restriction (d), note that the normalization by s_n implies that the sum of squared demeaned middle observations cannot be larger than unity. Ignoring the demeaning, taking expectations and approximating the distribution of these observations by again extending the extreme value distribution yields restriction (d) with a right-hand side of unity. We relax the upper bound to equal 2 to accommodate approximating errors in this argument.

We further impose cross restrictions between the two tails:

(e). $\mathbb{E}[Y_k^R] \geq -\mathbb{E}[Y_k^L]$

(f). $\sum_{i=1}^{n_0/2} \mathbb{E}[Y_i^L] > \sum_{i=1}^{n_0/2} \mathbb{E}[Y_i^R]$ implies $\mathbb{E}[Y_{n_0/2}^R] > 0$

(g). $\sum_{i=1}^{n_0/2} \mathbb{E}[Y_i^R] > \sum_{i=1}^{n_0/2} \mathbb{E}[Y_i^L]$ implies $\mathbb{E}[Y_{n_0/2}^L] > 0$

(h). $\sum_{i=k+1}^{n_0/2} \mathbb{E}[(Y_i^L)^2] + \sum_{i=k+1}^{n_0/2} \mathbb{E}[(Y_i^R)^2] \leq 2$

Restriction (e) amounts to an assumption that the two tails don't overlap. Under an extended tail assumption up to the most extreme $n_0/2$ observations, the middle observations take on values between $-Y_{n_0/2}^L$ and $Y_{n_0/2}^R$, leading to restrictions (f)-(g) under the null hypothesis of the overall mean being zero. Finally, restriction (h) is the analogous version of restriction (d) for each tail.

While restriction (c) involves the extreme value approximation for the most extreme n_0 observations, note that this approximation is only used to motivate a lower bound on κ^J, and for no other purpose. Consider, for instance, a sample of size $n = n_0 = 50$ from a mean-zero population that is a 50-50 mixture between a right skewed Pareto distribution and a uniform distribution, with overall continuous density. Since the uniform distribution is relatively more spread out compared to the left-tail of a right-skewed Pareto distribution, the overall mean-zero constraint leads to a Pareto tail that is shifted outward compared to a full Pareto distribution. Thus, restriction (c) is satisfied for this mixture, and as long as k is smaller than $n_0/2 = 25$, the approximate parametric model (16) can still be a good approximation

At the same time, one might argue that if the sample size n is much larger than n_0, this default parameter space Θ_0 is artificially large, and more powerful inference could be obtained by suitably reducing it. Note, however, that for any sample size n, the tails could be as large as they are in a sample of size $n_0 = 50$. For instance, consider a sample of size $n = 5000$ from a population that is a mixture between a point mass at zero and a Pareto distribution, with 99% mass on the point mass at zero. Then only approximately 50 observations in the sample will be non-zero, and those follow the Pareto distribution, so Θ_0 is again appropriate, and mechanical reduction of Θ_0 as a function of n leads to a poorly performing test in this problem.

Ultimately, inference about the mean requires a substantial assumption about the tails, and the stronger the assumptions, the more powerful the potential inference. The restriction Θ_0 as described here is one such choice, and as will be shown below, it yields informative inference while maintaining a high degree of robustness under moderately heavy tails.

4.3 Numerical Determination of Powerful Tests

Our approach is a variant of the algorithms in Elliott, Müller, and Watson (2015), denoted by EMW in the following, and Müller and Watson (2018); see Müller and Watson (in preparation) for a detailed survey. This approach yields a likelihood ratio-type test of the form

$$\varphi(\mathbf{y}^*) = \mathbf{1}\left[\frac{f_a(\mathbf{y}^*)}{\sum_{i=1}^{M}\lambda_i f(\mathbf{y}^*|\theta_i, 0)} > 1\right] \tag{20}$$

where the M values of $\theta_i \in \Theta_0$ and associated positive weights λ_i are iteratively determined so that the discrete mixture of θ taking on the values θ_i with probability $\lambda_i / \sum_{j=1}^{M} \lambda_j$ forms an approximate least favorable distribution for testing H_0 against the alternative H_a :"the density of $\mathbf{Y}^*$ is f_a". Here f_a is chosen to equal $f_a(\mathbf{y}^*) = \int f(\mathbf{y}^*|\theta,\mu)dF_a(\theta,\mu)$ for some weighting function F_a that determines against what kind of alternatives the resulting test is designed to be particularly powerful.

4.3.1 Specification of Weighting Function

We choose $F_a(\theta,\mu)$ to be an improper weighting function with density that is proportional to

$$\mathbf{1}[-1/2 \le \xi^L \le 1/2]\mathbf{1}[-1/2 \le \xi^R \le 1/2]/(\eta^R \eta^L) \tag{21}$$

so that the implied density on μ, κ^L and κ^R is flat. This choice is numerically convenient, as it leads to the product form $f_a(\mathbf{y}^*) = f_a^S(\mathbf{y}^L) f_a^S(\mathbf{y}^R)$ with $f_a^S(\mathbf{y}) = \int_{-1/2}^{1/2} f_{a|\xi}^S(\mathbf{y}|\xi)d\xi$ and $f_{a|\xi}^S$ proportional to the density of the scale and location maximal invariant considered in Müller and Wang (2017). By the same arguments as employed there, $f_{a|\xi}^S$ can be obtained by one dimensional Gaussian quadrature, and we approximate f_a^S by an average of those over a grid of values for $\xi \in [-1/2, 1/2]$.

The lower bound of $-1/2$ on (ξ^L, ξ^R) in (21) plays no important role, since for values of ξ^J that imply a thin tail, with very high probability the test is constrained to be of a form that does not involve f_a, as discussed next.

4.3.2 Switching

A key ingredient in the algorithm of EMW is an importance sampling estimate of the null rejection probability $\mathrm{RP}(\theta) = \mathbb{E}_\theta[\varphi^c(\mathbf{Y}^*)] = \int \varphi^c(\mathbf{y}^*) f(\mathbf{y}^*|\theta,0)d\mathbf{y}^*$ of a candidate test $\varphi^c : \mathbb{R}^{2k+1} \mapsto \{0,1\}$ under θ,

$$\widehat{\mathrm{RP}}(\theta) = N^{-1} \sum_{l=1}^{N} \varphi^c(\mathbf{Y}_{(l)}^*) \frac{f(\mathbf{Y}_{(l)}^*|\theta,0)}{\bar{f}(\mathbf{Y}_{(l)}^*)} \tag{22}$$

where $\mathbf{Y}_{(l)}^*$, $l = 1,\ldots,N$ are i.i.d. draws from the proposal density $\bar{f}$ (so that by the LLN, $\widehat{\mathrm{RP}}(\theta) \to \mathbb{E}_{\bar{f}}[\varphi^c(\mathbf{Y}^*) f(\mathbf{Y}^*|\theta,0)/\bar{f}(\mathbf{Y}^*)] = \mathrm{RP}(\theta)$ in obvious notation), where an appropriate $\bar{f}$ may be obtained by the algorithm in Müller and Watson (2018). Clearly, the larger Θ_0, the larger the number of importance sampling draws N needs to be for $\widehat{\mathrm{RP}}$ to be of satisfactory accuracy uniformly in $\theta \in \Theta_0$.

The nuisance parameter space Θ_0 of the last section is unbounded: the restrictions there did not put any lower bound on the scale parameters η^J or the shape parameters ξ^J, $J \in \{L, R\}$. Since the distribution of $\mathbf{Y}^*$ is highly informative about the scale of the tails, it is not possible to obtain uniformly accurate approximations via $\widehat{\mathrm{RP}}$ over Θ_0, even with arbitrary computational resources. It is therefore necessary to choose the test φ in a way that does not require a computational check of $\mathbb{E}_\theta[\varphi(\mathbf{Y}^*)] \leq \alpha$ over the entirety of Θ_0.

The solution to this challenge suggested in Elliott, Müller, and Watson (2015) is to *switch* to a default test with known size control under $\Theta_{00} \subset \Theta_0$, where the switching rule is such that the default test is employed with probability very close to one whenever $\mathbf{Y}^*$ is generated from Θ_{00}. If for $J \in \{L, R\}$, ξ^J is very small or κ^J is very large, then the resulting observation Y^J is highly compressed in the sense that Y_1^J/Y_k^J is positive and not much larger than one. Also, if Y_1^J (and thus the entire vector $\mathbf{Y}^J$) is small, then even if the tail is heavy in the sense of Y_1^J/Y_k^J being large, the tail still only makes a minor contribution to the overall variation of the data. We operationalize this by introducing the switching index

$$\chi(\mathbf{Y}^J) = \max(0, \min(Y_1^J - \rho_1, \mathbf{1}[Y_k^J > 0](Y_1^J/Y_k^J - 1 - \rho_r)) \tag{23}$$

for positive values of ρ_r, ρ_1 close to zero, so that $\chi(\mathbf{Y}^J) = 0$ implies that either Y_1^J is small or Y_1^J/Y_k^J is close to unity. If $\chi(\mathbf{Y}^J) = 0$ for one tail, but not the other, then the problem is heuristically close to knowing that only one of the tails is moderately heavy. For example, suppose $\chi(\mathbf{Y}^L) = 0$, so the left-tail seems thin. Under approximation (15), the sum of all observations that are not in the right tail equals $\tilde{Y}_0^L = Y_0 - \sum_{i=1}^k Y_i^L$, with corresponding approximate variance equal to $\tilde{V}^L = 1 + \sum_{i=1}^k (Y_i^L)^2$. It hence makes sense to switch to a "single tail" test $\varphi^S : \mathbb{R}^{k+2} \mapsto \{0, 1\}$ that treats $\mathbf{Y}^R$ as the extreme observations from the potentially heavy tail, and $\tilde{Y}_0^L$ to be approximately normal with mean $-M^*(\mathbf{Y}^R, \theta^R)$ and variance $\tilde{V}^L$. In analogy to (20), such a test is of the form

$$\varphi^S(\mathbf{y}^R, \mathbf{y}^L, y_0) = \mathbf{1}\left[\frac{f_a^S(\mathbf{y}^R)}{\sum_{i=1}^{M^S} \lambda_i^S f^S(\mathbf{y}^R, \mathbf{y}^L, y_0|\theta_i^S)} > 1\right] \tag{24}$$

where $\theta_i^S \in \mathbb{R}^3$ and $\lambda_i^S > 0$ form again a numerically determined approximate least favorable distribution, and

$$f^S(\mathbf{y}^R, \mathbf{y}^L, y_0|\theta_i^S) = f_T(\mathbf{y}|\theta^S)\phi((\tilde{y}_0^L + M^*(\mathbf{y}^R, \theta^S)/\sqrt{\tilde{v}^L})/\sqrt{\tilde{v}^L}. \tag{25}$$

The corresponding test for a thin right tail is given by $\varphi^S(\mathbf{y}^L, \mathbf{y}^R, -y_0)$.

If both tails seem thin, then one would expect that the distribution of the analogue to the full-sample t-statistic (cf. (11) from Section 3.2)

$$T(\mathbf{Y}^*) = \frac{Y_0 + \sum_{i=1}^k Y_i^R - \sum_{i=1}^k Y_i^L}{\sqrt{1 + \sum_{i=1}^k (Y_i^R)^2 + \sum_{i=1}^k (Y_i^L)^2}} \tag{26}$$

to be reasonably well approximated by a standard normal distribution, especially if $\sum_{i=1}^k (Y_i^R)^2 + \sum_{i=1}^k (Y_i^L)^2$ is small.

These considerations, a numerical analysis, and the sequential structure of the eventual algorithm presented in Section 4.3.4 below motivate the restriction of tests φ to reject, $\varphi(\mathbf{Y}^*) = 1$, only if all of the following four conditions hold:

1. $|T(\mathbf{Y}^*)| > \mathrm{cv}_T(\mathbf{Y}^*)$, where $\mathrm{cv}_T(\mathbf{Y}^*) = w_{\mathrm{cv}}(\mathbf{Y}^*)\,\mathrm{cv}_\alpha^Z + (1 - w_{\mathrm{cv}}(\mathbf{Y}^*))\,\mathrm{cv}_\alpha^T$ with $w_{\mathrm{cv}}(Y^*) = 1/(1 + \sum_{i=1}^k (Y_i^R)^2 + \sum_{i=1}^k (Y_i^L)^2)$ and $(\mathrm{cv}_\alpha^Z, \mathrm{cv}_\alpha^T)$ the $1 - \alpha/2$ quantiles of a standard normal and student-t distribution with degrees of freedom equal to $80 + 10\log(\alpha)$, respectively;

2. $\varphi_\chi^S(\mathbf{Y}^R, \mathbf{Y}^L, Y_0) = 1$;

3. $\varphi_\chi^S(\mathbf{Y}^L, \mathbf{Y}^R, -Y_0) = 1$;

4. the "two-tail" test $\varphi^{TT} : \mathbb{R}^{2k+1} \mapsto \{0, 1\}$ of the form (20) rejects, $\varphi^{TT}(\mathbf{Y}^*) = 1$;

where

$$\varphi_\chi^S(\mathbf{y}^R, \mathbf{y}^L, y_0) = \mathbf{1}\left[\frac{\exp[5\chi(\mathbf{y}^L)] \cdot f_a^S(\mathbf{y}^R)}{\sum_{i=1}^{M^S} \lambda_i^S f^S(\mathbf{y}^R, \mathbf{y}^L, y_0 | \theta_i^S)} > 1\right]. \tag{27}$$

The additional term $\exp[5\chi(\mathbf{y}^L)]$ in (27) compared to (24) ensures that conditions 2 and 3 are not binding whenever the corresponding switching index $\chi(\mathbf{y}^L)$ is large. Its continuity in $\mathbf{y}^L$ avoids the sharp change of the form of the rejection region as a function of $\mathbf{y}^*$ that would be induced by a simpler hard threshold rule $\varphi_\chi^S(\mathbf{y}^R, \mathbf{y}^L, y_0) = \mathbf{1}[\chi(\mathbf{y}^L) = 0]\varphi^S(\mathbf{y}^R, \mathbf{y}^L, y_0)$.

Condition 1 implies that φ never rejects if the analogue $T(\mathbf{Y}^*)$ of the usual t-statistic does not reject; in that sense, we seek to "robustify" the usual t-statistic to obtain better size control. Condition 1 has the additional appeal that sums of the form (22) then effectively only involve $\mathbf{Y}_{(l)}^*$ for which $|T(\mathbf{Y}_{(l)}^*)| \geq \mathrm{cv}_T(\mathbf{Y}_{(l)}^*)$, with an associated gain in computing speed.

We stress that the definition of a "thin tail" in (23), the approximate normality of (26) and so forth are purely heuristic and do not enter the evaluation of $\mathbb{E}_\theta[\varphi(\mathbf{Y}^*)]$ by the algorithm; this probability is always computed from the distribution (18) of $\mathbf{Y}^*$. The heuristics merely motivate the particular form of φ just described. As discussed, it is not possible to numerically check that $\mathbb{E}_\theta[\varphi(\mathbf{Y}^*)] \leq \alpha$ for *all* $\theta \in \Theta_0$. But we employ extensive numerical analysis to ensure that $\mathbb{E}_\theta[\varphi(\mathbf{Y}^*)] \leq \alpha$ over a very large set, including values of θ that lead to the events $\chi(\mathbf{Y}^J) = 0$ for $J \in \{L, R\}$ with probability close to zero, close to one or in between. The simple form that φ takes on with very high probability in the remainder of the parameter space makes it plausible that $\mathbb{E}_\theta[\varphi(\mathbf{Y}^*)] \leq \alpha$ for all $\theta \in \Theta_0$, or at the least, very nearly so.

4.3.3 Recombining Tails in Importance Sampling

Even though the switching rule of the last subsection reduces the numerically relevant parameter space to a bounded set, this set still turns out to be so large that a very large number N of importance sampling draws are necessary to obtain adequate approximations. The computationally expensive part in the evaluation of $\widehat{\mathrm{RP}}(\theta)$ in (22) for different θ is the evaluation of $f(\mathbf{Y}^*_{(l)}|\theta, 0)$ (since all $\bar{f}(\mathbf{Y}^*_{(l)})$ can be computed once and stored).

These evaluations can be dramatically sped up by recombining two "single tails" in different combinations: For a given $\theta^S = (\kappa, \eta, \xi)$, let $\mathbf{Y}^e \in \mathbb{R}^{k+1}$ be an "extended" single tail with distribution

$$\mathbf{Y}^e = \begin{pmatrix} \eta(\mathbf{X} + \kappa\mathbf{e}) \\ Z/\sqrt{2} - \eta m^*(\mathbf{X}, \kappa, \xi) \end{pmatrix} = \begin{pmatrix} \mathbf{Y}^S \\ Y_0^e \end{pmatrix}$$

where $\mathbf{X}$ is distributed as as in (13), independent of $Z \sim \mathcal{N}(0, 1)$. Denote the density of $\mathbf{Y}^e$ by $f^e(\mathbf{y}^e|\theta^S)$. Given two independent vectors $\mathbf{Y}^e_{(1)}$ and $\mathbf{Y}^e_{(2)}$ distributed according to $\theta_1^S = \theta^L$ and $\theta_2^S = \theta^R$, respectively, note that their combination into the "both tails" observation $(\mathbf{Y}^{S\prime}_{(1)}, \mathbf{Y}^{S\prime}_{(2)}, Y^e_{0,(1)} - Y^e_{0,(2)})' \in \mathbb{R}^{2k+1}$ has the same distribution as $\mathbf{Y}^*$ in (18), since the difference of two independent normals of variance $1/2$ is again standard normal. Thus, with $\mathbf{Y}^e_{(l)}$ i.i.d. draws from a suitable proposal density $\bar{f}^e$, one obtains the alternative estimator

$$\widetilde{\mathrm{RP}}(\theta) = (KN)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{N} \varphi^c((\mathbf{Y}^{S\prime}_{(l)}, \mathbf{Y}^{S\prime}_{(l+k)}, Y^e_{0,(l)} - Y^e_{0,(l+k)})') \frac{f^e(\mathbf{Y}^e_{(l)}|\theta^L) f^e(\mathbf{Y}^e_{(l+k)}|\theta^R)}{\bar{f}^e(\mathbf{Y}^e_{(l)}) \bar{f}^e(\mathbf{Y}^e_{(l+k)})} \quad (28)$$

that recombines each extended single tail with K different other extended single tails, for a total of KN importance draws. Yet evaluation of (28) only requires a simple product of the

$(K + N)$ values $f^e(\mathbf{Y}^e_{(l)}|\theta^S)$ for $\theta^S \in \{\theta^L, \theta^R\}$. We let $K = 128$ and $N = 640,000$ for a total of nearly 82 million importance sampling draws.

4.3.4 Implementation

The overall algorithm proceeds in four stages. To describe these stages, let $\Theta^S_0 \subset \mathbb{R}^3$ be the set of parameters satisfying the constraints (a)-(d) of Section 4.2 on one tail. Let $\Theta^S_s \subset \Theta^S_0$ be such that for $\theta^S \in \Theta^S_s$, the event that the switching index is zero, $\chi(\mathbf{Y}^J) = 0$, happens with at least 90% probability, and $\Theta^S_{ss} \subset \Theta^S_s$ be such that $\chi(\mathbf{Y}^J) = 0$ with probability of exactly 90%.

1. Choose (ρ_r, ρ_1) such that for $\theta^L, \theta^R \in \Theta^S_{ss}$, $\mathbb{E}_\theta[\mathbf{1}[|T(\mathbf{Y}^*)| > \mathrm{cv}_T(\mathbf{Y}^*)]] \leq \alpha$ under $\theta = (\theta^{L\prime}, \theta^{R\prime})'$.

2. Use the algorithm of EMW to numerically determine φ^S via $\{\lambda^S_i\}^{M_S}_{i=1}$ and $\{\theta^S_i\}^{M_S}_{i=1}$ in (24) so that
$$\mathbb{E}_\theta[\mathbf{1}[|T(\mathbf{Y}^*)| > \mathrm{cv}_T(\mathbf{Y}^*)]\varphi^S(\mathbf{Y}^R, \mathbf{Y}^L, Y_0)] \leq \alpha$$
for all $\theta = (\theta^{L\prime}, \theta^{R\prime})' \in \Theta_0$ with $\theta^L \in \Theta^S_{ss}$ and $\theta^R \in \Theta^S_0 \backslash \Theta^S_s$.

3. Use the algorithm of EMW to determine φ^{TT} via $\{\lambda_i\}^M_{i=1}$ and $\{\theta_i\}^M_{i=1}$ so that the overall test φ of the form described in Section 4.3.2 satisfies $\mathbb{E}_\theta[\varphi(\mathbf{Y}^*)] \leq \alpha$ under all $\theta = (\theta^{L\prime}, \theta^{R\prime})' \in \Theta_0$ for $\theta^L, \theta^R \in \Theta^S_0 \backslash \Theta^S_s$.

4. Spot-check that φ indeed satisfies $\mathbb{E}_\theta[\varphi(\mathbf{Y}^*)] \leq \alpha$ for all $\theta \in \Theta_0$, including $\theta = (\theta^{L\prime}, \theta^{R\prime})'$ with $\theta^L, \theta^R \in \Theta^S_s$.

Note that the parameter set under consideration becomes consecutively larger in Steps 1-3, and the form imposed on tests described in Section 4.3.2 ensures that any potential remaining overrejections of a stage (say, due to estimation errors of rejection probabilities) can be corrected by the subsequent stage, which increases the numerical stability of the algorithm. Null rejection probabilities are estimated throughout with the importance sampling estimator of Section 4.3.3. This estimator has an importance sampling standard error (appropriately adjusted for the dependence in (28)) of no more than 0.05%, 0.15% and 0.2% for $\alpha = 1\%, 5\%, 10\%$, respectively.

We apply this algorithm to the default nuisance parameter space with $n_0 = 50$ of Section 4.2 for $k = 8$ and $\alpha \in \{0.002, 0.004, \ldots, 0.008, 0.01, 0.02, \ldots, 0.10, 0.12, \ldots,$

$0.20, 0.25, 0.30, 0.40, 0.50\}$, and also for $k = 4$ to the larger nuisance parameter space where $n_0 = 25$ (with $n_0/2$ interpreted as 12) in the constraints of Section 4.2. By adding the obvious additional constraints to the form of the tests for $\alpha \neq 0.05$, we ensure that the 95% and 99% level confidence intervals obtained via test inversion[3] always contain the 90% and 95% level intervals, respectively, and that the p-value is always coherent with the level of the reported confidence interval. After trivial modifications that decrease their rejection probability by an arbitrarily small amount, these tests satisfy the condition of the two-tailed analogue of Theorem 2; see Appendix A.2 for details. For comparison purposes, we also generate tests with $k \in \{4, 12\}$ for $\alpha \in \{0.01, 0.05\}$ in the default parameter space. For $k = 8$ and a given level α, the computations take about one hour on a modern 24 core workstation in a Fortran implementation, and about 3 hours for $k = 12$. Once the values for $\{\lambda_i^S, \theta_i^S\}_{i=1}^{M_S}$ and $\{\lambda_i, \theta_i\}_{i=1}^{M}$ are determined in this fashion, the evaluation of the resulting test φ, as required in applications, is computationally trivial.

4.4 Application to GMM

Suppose we estimate the parameter vector $\vartheta = (\beta, \gamma')' \in \mathbb{R}^q$ by Hansen's (1982) Generalized Method of Moments using the $r \times 1$ moment condition $\mathbb{E}[g(\vartheta, z)] = 0$ from data z_j, $j = 1, \ldots, n_z$ and $r \times r$ positive definite weighting function $\hat{\Psi}$. Suppose further that the data z_j is i.i.d. across clusters defined by the partition $\{\mathcal{C}_i\}_{i=1}^n$ of $\{j : 1 \leq j \leq n\}$ (so that $\mathcal{C}_i = \{i\}$ and $n = n_z$ under i.i.d. sampling of z_j). Then as $n \to \infty$, under standard regularity conditions, $\hat{\vartheta} = (\hat{\beta}, \hat{\gamma}')'$ satisfies

$$\sqrt{n}(\hat{\vartheta} - \vartheta) = (\Gamma' \Psi \Gamma)^{-1} \Gamma' \Psi \cdot n^{-1/2} \sum_{i=1}^{n} G_i + o_p(1) \tag{29}$$

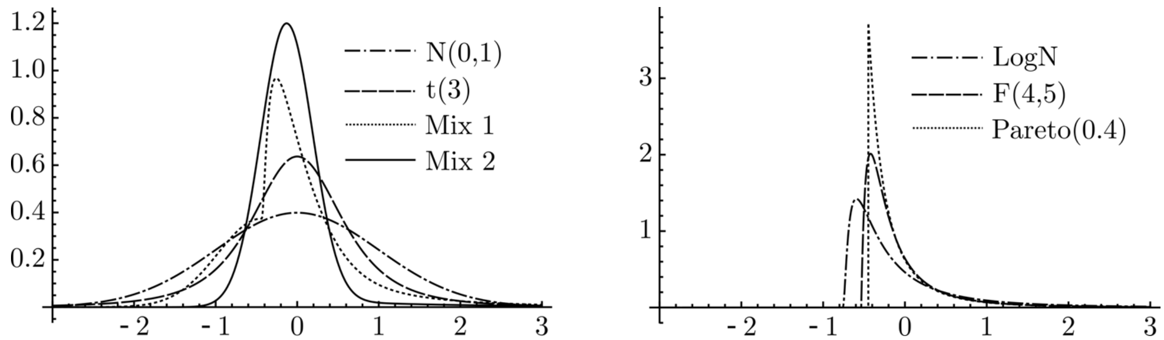
where $G_i = \sum_{j \in \mathcal{C}_i} g(\vartheta, z_j)$ are i.i.d., $\hat{\Gamma} = n^{-1} \sum_{j=1}^{n_z} \partial g(\vartheta, z_j)/\partial \vartheta'|_{\vartheta = \hat{\theta}} \xrightarrow{p} \Gamma$ and $\hat{\Psi} \xrightarrow{p} \Psi$ with Γ and Ψ non-stochastic, so that the large sample variability of $\hat{\vartheta}$ is entirely driven by the average of i.i.d. observations G_i. Correspondingly, the standard GMM hypothesis test of $H_0 : \beta = \beta_0$ is numerically equivalent to the usual t-test of $H_0 : \beta = \beta_0$ computed from the n observations

$$\hat{W}_i = \hat{\beta} + \iota_1' (\hat{\Gamma}' \hat{\Psi} \hat{\Gamma})^{-1} \hat{\Gamma}' \hat{G}_i, \; i = 1, \ldots, n \tag{30}$$

where $\hat{G}_i = \sum_{j \in \mathcal{C}_i} g(\hat{\vartheta}, z_j)$ and ι_1 is the $q \times 1$ vector $(1, 0, \ldots, 0)'$.

[3] In the rare samples where test inversion yields disconnected sets, we set the confidence interval equal to the smallest interval that contains all non-rejections.

Figure 1: Population Densities in Monte Carlo Experiments



Thus, to the extent that G_i follows a distribution with moderately heavy tails, one would expect that small sample inference is improved by applying the new test to the observations $\{\hat{W}_i\}_{i=1}^n$. A corresponding analytical refinement result analogous to Theorem 2 is beyond the scope of this paper.

5 Small Sample Results

This section presents six sets of small sample results: two for inference about the mean from an i.i.d. sample, two for the difference of population means from two independent samples, and two for a regression coefficient with clustered standard errors. In all three cases, the data is either generated from analytical distributions, or from draws with replacement from a large data set. We focus on tests of nominal 5% level in the main text; results for 1% level tests are reported in the appendix and exhibit broadly similar patterns.

5.1 Inference for the Mean

We initially compare our default test with $k = 8$ ("NEW DEFAULT") with three standard tests for the population mean: standard t-statistic based inference with critical value from a student-t distribution with $n - 1$ degrees of freedom "T-STAT"; the percentile-t bootstrap based on the absolute value of the t-statistic "SYM-BOOT"; and the percentile-t bootstrap based on the signed t-statistic "ASYM-BOOT". The data is generated from one of seven populations: the standard normal distribution N(0,1), the log-normal distribution LogN, the F-distribution with 4 degrees of freedom in the numerator and 5 in the denominator

Table 1: Small Sample Results in Inference for the Mean

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
T-STAT	5.0|**0.99**	10.0|0.74	13.5|0.65	4.7|**1.01**	13.6|0.62	7.4|0.88	18.8|0.60
SYM-BOOT	5.0|**1.00**	7.8|1.07	10.8|1.27	4.1|**1.11**	10.6|1.35	6.9|1.12	18.1|1.44
ASYM-BOOT	5.2|**1.00**	6.9|0.96	8.9|1.03	7.4|1.06	8.6|1.07	8.1|1.02	17.6|1.08
NEW DEFAULT	3.8|**1.10**	3.2|**0.93**	4.5|**0.77**	3.3|**1.44**	5.2|**0.72**	3.3|**1.11**	12.2|0.66
				$n = 100$			
T-STAT	4.9|**1.00**	8.2|0.83	10.9|0.73	4.6|**1.01**	11.5|0.71	6.9|0.91	15.4|0.60
SYM-BOOT	5.0|**1.00**	6.7|1.04	9.1|1.21	4.2|**1.08**	9.2|1.19	6.4|1.06	14.1|1.17
ASYM-BOOT	5.1|**1.00**	6.5|0.97	7.5|1.02	6.6|1.05	7.7|1.00	7.4|1.00	13.4|0.95
NEW DEFAULT	4.8|**1.01**	3.1|**1.26**	3.6|**1.04**	3.8|**1.37**	3.6|**1.00**	3.3|**1.31**	7.9|0.75
				$n = 500$			
T-STAT	5.0|**1.00**	5.9|**0.95**	7.8|0.87	4.8|**1.01**	7.9|0.86	5.8|**0.97**	9.6|0.77
SYM-BOOT	5.0|**1.00**	5.4|**1.01**	6.9|1.10	4.7|**1.03**	6.8|1.19	5.4|**1.01**	8.1|1.04
ASYM-BOOT	5.0|**1.00**	5.5|**1.00**	7.0|1.01	6.1|1.02	6.4|1.05	6.0|1.00	7.7|0.95
NEW DEFAULT	4.9|**1.00**	4.1|**1.18**	4.3|**1.21**	4.5|**1.13**	4.1|**1.22**	4.4|**1.18**	3.2|**1.21**

Notes: Entries are the null rejection probability in percent, and the average length of confidence intervals relative to average length of confidence intervals based on size corrected t-statistic (bold if null rejection probability is smaller than 6%) of nominal 5% level tests. Based on 20,000 replications.

F(4,5), the student-t distribution with 3 degrees of freedom t(3), an equal probability mixture between a N(0,1) and LogN distribution Mix1, and a 95 / 5 mixture between a N(0,1/25) and a LogN distribution Mix2. All population distributions are normalized to have mean zero and unit variance; the corresponding densities are plotted in Figure 1. Technically, only the Pareto distribution, the t-distribution and the F-distribution exhibit heavy Pareto like tails in the sense of (3) with tail indices $\xi = 0.4$, $\xi = 1/3$ and $\xi = 0.4$, respectively, but as a practical matter, also the log-normal and the two mixture distributions are right-skewed enough to make small sample inference challenging.

Table 1 reports null rejection probabilities, along with the average length of the resulting confidence interval, expressed as a multiple of the average length of the infeasible confidence interval that is based on the t-statistic, but applies the size adjusted critical value. As can be seen from Table 1, the new method comes much closer to controlling size under moderately heavy-tailed distributions. For the thin-tailed normal population, the new method only leads to 10% longer intervals for $n = 50$, and essentially no excessive length for $n \in \{100, 500\}$. For other populations, the intervals of the new method are often much longer than those from other methods; but since the other methods do not come close to controlling size, that comparison is not meaningful (entries in bold indicate where tests are close to valid with a null rejection probability below 6%). Remarkably, for $n = 50$, the new method yields shorter confidence intervals than the size corrected t-statistic for some populations while still controlling size. The explicit modelling of the tails can also yield efficiency gains, since under a Pareto-like tail, the sample mean is not the efficient estimator of the population mean.

An exception to the good performance of the new method is the student-t population with three degrees of freedom. Even though it has fairly heavy tails, with the third moment not existing, its symmetry enables T-STAT and SYM-BOOT to control size at much less cost to average length compared to the new method.[4]

Table 2 compares different versions of the new method across the same set of seven populations. We consider $k \in \{4, 8, 12\}$ for the default parameter space with $n_0 = 50$, and also include the even more robust test with $k = 4$ constructed from the parameter space in Section 4.2 with $n_0 = 25$. For $n = 25$, only the test with $n_0 = 25$ comes close to controlling size for non-thin tailed populations (and the test with $k = 12$ cannot be applied at all, since there is only a single "middle" observation). For larger n, the tests with $k = 4$ are even more successful in controlling size compared to the default method, but at a non-negligible cost in terms of longer confidence intervals. In contrast, the test for $k = 12$ does not yield an additional substantial reduction in average length, and has worse size control for $n = 50$. These results underlie our choice of the test with $k = 8$ and $n_0 = 50$ as the default, and in the following, we exclusively focus on this variant.

A potential objection to this first set of Monte Carlo results is that the underlying populations have smooth tails, which might overstate the effectiveness of the new method "in practice". To address this concern, consider a population that is equal to the (discrete) distribution from a large economic data set. We use the income data of 2016 mortgage applicants as reported by banks under the Home Mortgage Disclosure Act (HMDA). From this database of more than 16 million applications, we create subpopulations that condition

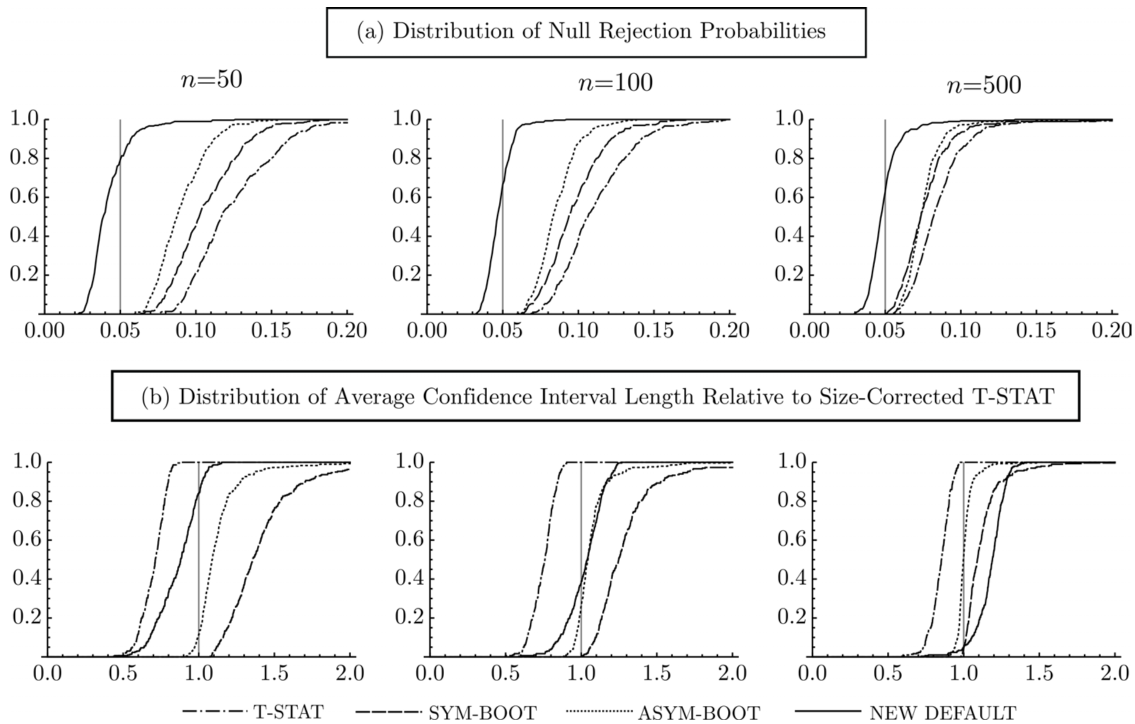
[4] The analytical result by Bakirov and Székely (2005) shows that the usual 5% level t-test remains small sample valid under arbitrary scale mixtures of normals, which includes all t-distributions.

Table 2: Small Sample Results of New Methods for Inference for the Mean

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 25$			
DEF: $k = 8, n_0 = 50$	4.7|**1.00**	13.1|0.64	16.5|0.56	4.0|**1.02**	17.8|0.53	8.0|0.82	18.7|0.62
$k = 4, n_0 = 50$	2.7|**1.25**	7.4|0.70	10.1|0.61	3.1|**1.16**	11.6|0.57	5.2|**0.93**	11.7|0.67
$k = 12, n_0 = 50$	NA	NA	NA	NA	NA	NA	NA
$k = 4, n_0 = 25$	2.3|**1.42**	4.3|**0.80**	5.7|**0.67**	2.1|**1.49**	7.3|0.61	3.1|**1.09**	9.3|0.72
				$n = 50$			
DEF: $k = 8, n_0 = 50$	3.8|**1.10**	3.2|**0.93**	4.5|**0.77**	3.3|**1.44**	5.2|**0.72**	3.3|**1.11**	12.2|0.66
$k = 4, n_0 = 50$	3.9|**1.15**	2.6|**0.99**	3.6|**0.81**	3.6|**1.46**	4.3|**0.76**	3.2|**1.15**	11.1|0.66
$k = 12, n_0 = 50$	3.5|**1.15**	4.3|**0.86**	6.1|0.73	2.8|**1.43**	8.0|0.68	2.8|**1.08**	10.9|0.67
$k = 4, n_0 = 25$	3.7|**1.15**	2.1|**1.15**	2.8|**0.94**	3.2|**1.70**	3.2|**0.87**	3.2|**1.32**	11.6|0.73
				$n = 100$			
DEF: $k = 8, n_0 = 50$	4.8|**1.01**	3.1|**1.26**	3.6|**1.04**	3.8|**1.37**	3.6|**1.00**	3.3|**1.31**	7.9|0.75
$k = 4, n_0 = 50$	5.0|**1.02**	3.0|**1.23**	3.4|**1.00**	3.8|**1.51**	3.8|**0.95**	3.5|**1.30**	5.6|**0.71**
$k = 12, n_0 = 50$	4.0|**1.06**	2.8|**1.27**	3.5|**1.07**	3.6|**1.33**	3.4|**1.04**	2.9|**1.33**	8.7|0.75
$k = 4, n_0 = 25$	5.1|**1.01**	2.6|**1.37**	3.3|**1.13**	4.2|**1.56**	3.7|**1.07**	3.6|**1.43**	6.2|0.82
				$n = 500$			
DEF: $k = 8, n_0 = 50$	4.9|**1.00**	4.1|**1.18**	4.3|**1.21**	4.5|**1.13**	4.1|**1.22**	4.4|**1.18**	3.2|**1.21**
$k = 4, n_0 = 50$	5.1|**1.00**	4.4|**1.31**	4.6|**1.26**	4.4|**1.31**	4.4|**1.24**	4.2|**1.32**	3.2|**1.10**
$k = 12, n_0 = 50$	5.1|**1.00**	3.7|**1.18**	4.2|**1.17**	3.1|**1.16**	4.2|**1.16**	3.7|**1.19**	2.7|**1.28**
$k = 4, n_0 = 25$	5.0|**1.00**	4.2|**1.34**	4.2|**1.38**	4.5|**1.28**	4.3|**1.32**	4.1|**1.36**	3.1|**1.27**

Notes: See Table 1.

Figure 2: Small Sample Results for HMDA Populations



on U.S. state and the gender of the applicant, as well as the purpose of the mortgage (home purchase, home improvement or refinancing) and whether or not the unit is owner-occupied. We eliminate all records with missing data, and only retain subpopulations with at least 5000 observations. For each of the resulting 300 subpopulations, we compare the performance of alternative methods for inference about the mean, based on i.i.d. samples of size n (that is, sampling is with replacement).

Panel (a) of Figure 2 plots the cumulative distribution function of the null rejection probabilities over the 300 subpopulations for each tests considered in Table 1, estimated from 20,000 draws from each subpopulation. Nominally, all mass should be to the left of the 5% line, but the traditional tests don't comes close. For instance, for $n = 100$, the usual t-statistic has null rejection probability of less than 10% for only approximately 40% of the 300 subpopulation. In comparison, the new test controls size much more successfully.

Panel (b) of Figure 2 plots the cumulative distribution function of the average length of the confidence intervals, relative to the average length of the size corrected t-statistic based interval. For $n = 50$, the new method not only controls size better than the bootstrap tests, but it also leads to confidence intervals that are typically shorter on average. In fact, they are substantially shorter than what is obtained from the infeasible size corrected interval. For $n = \{100, 500\}$, this is no longer the case and the better size control of the new method comes at the cost of somewhat longer confidence intervals.

One might argue that in the HMDA example, one could avoid the complications of the heavy right tail of the income distribution by considering the logarithm of the applicants' income. But, of course, there is no robust way to transform a confidence interval for the population mean of log-income into a valid confidence interval for the population mean income. What is more, in many contexts, the policy relevant parameter is the population mean (and not, say, the median) of some potentially heavy-tailed distribution: think of health care costs, or flood damage, or asset returns.

5.2 Difference between Two Population Means

Our second set of Monte Carlo experiment concerns inference about the difference of two population means $\mathbb{E}[W^{\mathrm{I}}] - \mathbb{E}[W^{\mathrm{II}}]$ based on two independent equal-sized i.i.d. samples $W_i^j \sim W^j$, $i = 1, \ldots, n/2$, $j \in \{\mathrm{I,II}\}$. Casting this in terms of a linear regression and applying the general mapping (30) yields

$$\hat{W}_i = \begin{cases} \bar{W}^{\mathrm{I}} - \bar{W}^{\mathrm{II}} + 2(W_i^{\mathrm{I}} - \bar{W}^{\mathrm{I}}) & \text{for } i \leq n/2 \\ \bar{W}^{\mathrm{I}} - \bar{W}^{\mathrm{II}} - 2(W_{i-n/2}^{\mathrm{II}} - \bar{W}^{\mathrm{II}}) & \text{for } i > n/2 \end{cases}$$

where $\bar{W}^j = (n/2)^{-1} \sum_{i=1}^{n/2} W_i^j$ are the sample means for $j \in \{\mathrm{I,II}\}$.

We initially generate data according to

$$W_i^{\mathrm{I}} = \nu_i + \varepsilon_i^{\mathrm{I}}, \quad W_i^{\mathrm{II}} = \varepsilon_i^{\mathrm{II}} \tag{31}$$

for $i = 1, \ldots, n/2$, where $\varepsilon_i^j \sim iid\mathcal{N}(0, 1/10)$ across i and $j \in \{\mathrm{I,II}\}$, and ν_i is distributed according to one of the distributions of Table 1. Inference about $\mathbb{E}[W^{\mathrm{I}}] - \mathbb{E}[W^{\mathrm{II}}]$ can then be thought of as inference about the average treatment effect $\mathbb{E}[\nu_i]$, with the design amounting to a large but highly heterogeneous additive treatment effect.

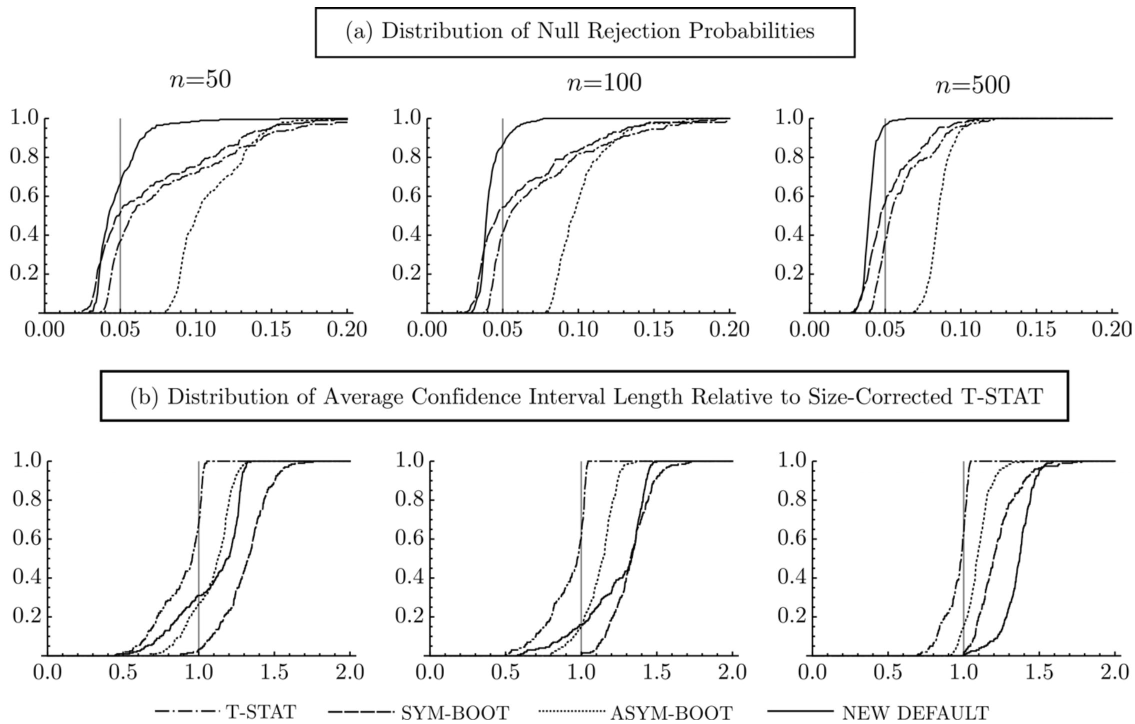
Table 3 compares the new method to standard t-statistic based inference and a symmetric and asymmetric percentile-t bootstrap, where now the bootstrap samples combine $n/2$ randomly selected observations with replacement from each of the two samples. In this exercise the design with $\nu_i \sim \mathcal{N}(0, 1)$ leads to a much longer confidence interval from the new method with $n = 50$. The reason is that with $\varepsilon_i^j \sim \mathcal{N}(0, 1/10)$ in (31), W_i^{I} has much

Table 3: Small Sample Results for Difference of Population Means

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
T-STAT	5.7|**0.96**	8.9|0.81	8.9|0.83	5.1|**0.99**	8.9|0.83	7.2|0.90	9.0|0.83
SYM-BOOT	5.7|**0.97**	8.3|1.02	8.6|1.12	4.7|**1.07**	8.6|1.12	6.8|1.06	8.9|1.22
ASYM-BOOT	5.9|**0.97**	8.8|0.93	9.1|0.98	7.4|1.03	9.1|0.98	8.6|0.98	10.0|1.02
NEW DEFAULT	2.0|**1.40**	3.8|**1.06**	4.7|**1.06**	2.3|**1.47**	4.8|**1.05**	3.5|**1.25**	6.2|0.99
				$n = 100$			
T-STAT	5.5|**0.98**	7.8|0.87	7.9|0.87	4.9|**1.00**	8.2|0.86	6.9|0.92	9.9|0.82
SYM-BOOT	5.4|**0.98**	7.0|1.04	7.4|1.13	4.4|**1.07**	7.7|1.14	6.4|1.05	9.6|1.21
ASYM-BOOT	5.4|**0.98**	7.6|0.98	7.9|1.01	6.9|1.04	8.3|1.01	7.6|0.99	10.6|1.02
NEW DEFAULT	4.4|**1.08**	3.4|**1.28**	4.2|**1.20**	3.7|**1.43**	4.4|**1.18**	4.0|**1.30**	7.8|1.01
				$n = 500$			
T-STAT	5.4|**0.98**	6.4|0.94	6.4|0.93	4.6|**1.01**	6.8|0.91	5.7|**0.97**	8.3|0.85
SYM-BOOT	5.5|**0.98**	5.8|**1.01**	6.0|1.12	4.4|**1.03**	6.3|1.10	5.3|**1.01**	7.7|1.08
ASYM-BOOT	5.4|**0.98**	6.3|0.99	6.6|1.04	5.8|**1.02**	6.6|1.02	6.2|1.00	8.3|0.98
NEW DEFAULT	5.3|**0.99**	4.2|**1.21**	4.1|**1.25**	4.2|**1.15**	4.2|**1.24**	4.2|**1.22**	4.2|**1.31**

Notes: See Table 1.

Figure 3: Small Sample Results for Two Samples from HDMA Populations



larger variance than W_i^{II}. The distribution of $\hat{W}_i$ is thus approximately equal to a 50-50 mixture of two normal distributions with very different variances, which is heavier tailed than a normal distribution. At the same time, for asymmetric ν_i, standard methods do not control size well, while the new method does so much more successfully.

As a second exercise, we generate W_i^j as $n/2$ i.i.d. draws of two randomly selected subpopulations of the HMDA data set considered in the last section. Inference about $\mathbb{E}[W^{\mathrm{I}}]-\mathbb{E}[W^{\mathrm{II}}]$ then corresponds to inference about the average treatment effect if the treatment induces a change from the distribution of income in one subpopulation to the distribution in another—maybe a plausible calibration for an intervention that affects individuals' incomes. Figure 3 reports the performance of the inference methods of Table 3 for 200 randomly selected pairs of subpopulations, in analogy to Figure 2 above. We find that also in this exercise, standard methods fail to produce reliable inference, while the new method is substantially more successful at controlling size.

5.3 Clustered Linear Regression

A third set of Monte Carlo experiments explores the performance of the new method for inference in a clustered linear regression

$$Y_{it} = \beta X_{it} + Z_{it}'\gamma + u_{it},\ t = 1, \ldots, T_i,\ i = 1, \ldots, n \tag{32}$$

with conditionally mean zero u_{it}, so that there are T_i observations in cluster i. Viewing linear regression as a special case of GMM inference, we obtain from the development of Section 4.4 and the Frisch-Waugh Theorem that

$$\hat{W}_i = \hat{\beta} + \left(n^{-1} \sum_{j=1}^{n} \sum_{t=1}^{T_j} \hat{X}_{jt} \right)^{-1} \sum_{t=1}^{T_i} \hat{X}_{it} \hat{u}_{it}$$

where $\hat{u}_{it}$ and $\hat{\beta}$ are the OLS estimates of u_{it} and β, and $\hat{X}_{it}$ are the residuals of a OLS regression of X_{it} on Z_{it}. We consider four tests of $H_0 : \beta = \beta_0$: The t-statistic implemented by STATA, which is nearly identical to a standard t-test applied to $\hat{W}_i$, except for degree of freedom corrections; the suggestion of Imbens and Kolesar (2016) to account for a potentially small number of heterogeneous clusters "Im-Ko" (we consider the variant that involves the data dependent degree of freedom adjustment K_{IK} in their notation); the wild cluster bootstrap that imposes the null hypothesis suggested by Cameron, Gelbach, and Miller (2008) "CGM"; and the new default test applied to $\hat{W}_i$ "new default".

We initially consider data generated from model (32) where

$$u_{it} = \nu_i X_{it} + \varepsilon_{it}, \tag{33}$$

ν_i is i.i.d. mean-zero with a distribution that is one of the seven populations considered in Table 1, one element of Z_{it} is a constant, and X_{it}, the 5 non-constant elements of Z_{it}, and ε_{it} are independent standard normal. We set $T_i = T = 10$ for all clusters. The presence of ν_i induces heteroskedastic correlations within each cluster of observations $\{Y_{it}\}_{t=1}^{T}$.

Table 4 reports the results. As in the inference about the mean problem, the new method is seen to control size much more successfully compared to the other methods, although at a cost in average confidence interval length that is more pronounced than in Table 1 for the thin-tailed $\nu_i \sim \mathcal{N}(0, 1)$. Intuitively, the product of two independent normals $\nu_i X_{it}$ has considerably heavier tails than a normal distribution, but it is still symmetric.

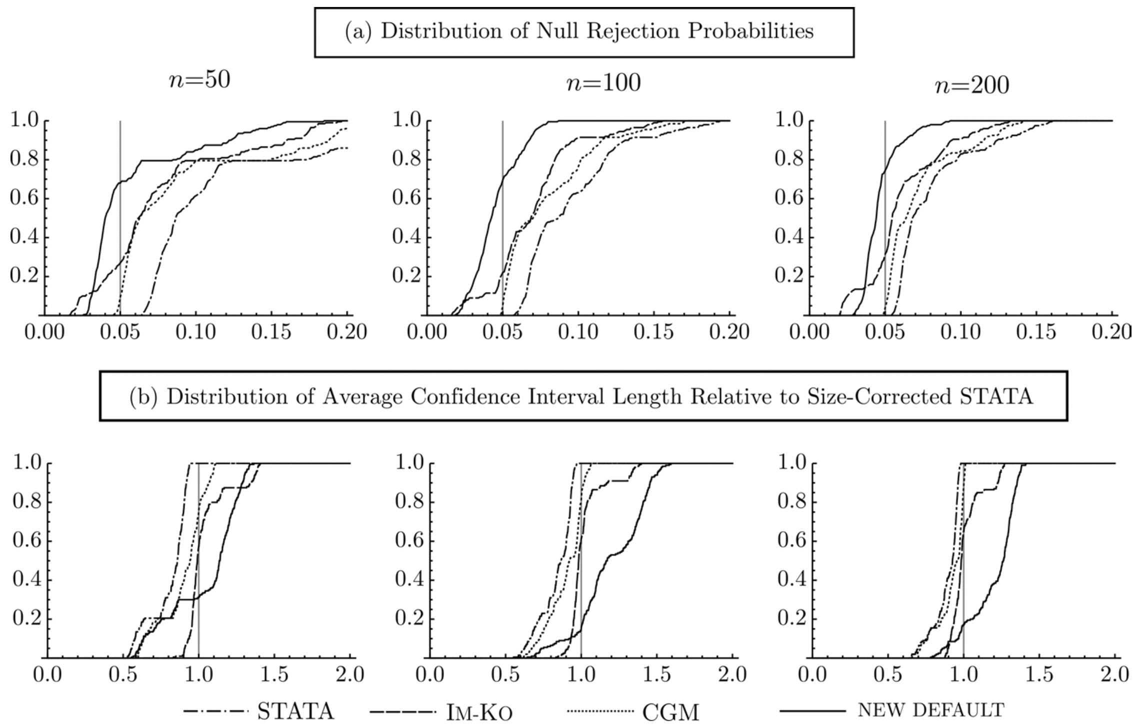
In the final Monte Carlo exercise we again consider a discrete population from a large economic data set. In particular, we consider a sample of all employed workers aged 18-65

Table 4: Small Sample Results in Clustered Regression Design

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
STATA	5.1|**1.00**	9.3|0.80	10.7|0.76	4.7|**1.01**	10.9|0.75	6.9|0.92	12.3|0.75
Im-Ko	4.9|**1.00**	9.1|0.81	10.5|0.77	4.5|**1.02**	10.7|0.75	6.7|0.92	12.0|0.75
CGM	5.0|**1.01**	9.4|0.77	10.8|0.72	5.0|**1.00**	11.0|0.70	7.0|0.89	12.3|0.68
NEW DEFAULT	3.3|**1.34**	3.5|**0.97**	4.4|**0.92**	2.8|**1.44**	4.5|**0.89**	3.3|**1.19**	7.3|0.88
				$n = 100$			
STATA	5.2|**0.99**	7.6|0.87	9.5|0.81	4.7|**1.01**	9.8|0.79	6.7|0.93	11.8|0.75
Im-Ko	5.1|**1.00**	7.5|0.87	9.4|0.81	4.7|**1.01**	9.8|0.80	6.6|0.93	11.7|0.75
CGM	5.0|**1.00**	7.7|0.85	9.6|0.77	4.9|**1.01**	9.9|0.75	6.6|0.91	11.9|0.69
NEW DEFAULT	4.5|**1.11**	3.2|**1.26**	4.2|**1.12**	4.0|**1.42**	4.4|**1.10**	3.8|**1.32**	7.0|0.96
				$n = 500$			
STATA	5.1|**1.00**	6.1|0.95	7.1|0.91	5.0|**1.00**	7.5|0.89	5.5|**0.97**	8.8|0.83
Im-Ko	5.1|**1.00**	6.1|0.95	7.1|0.91	5.0|**1.00**	7.4|0.89	5.5|**0.98**	8.8|0.83
CGM	5.0|**1.00**	6.1|0.94	7.3|0.87	5.1|**0.99**	7.7|0.85	5.6|**0.97**	9.0|0.80
NEW DEFAULT	5.0|**1.00**	4.1|**1.20**	4.3|**1.24**	4.7|**1.14**	4.4|**1.23**	4.0|**1.22**	3.5|**1.29**

Notes: Entries are the null rejection probability in percent, and the average length of confidence intervals relative to average length of confidence intervals based on size corrected STATA (bold if null rejection probability is smaller than 6%) of nominal 5% level tests.

Figure 4: Small Sample Results for CPS Clustered Regressions



from the 2018 merged outgoing rotation group sample of the Current Population Survey (CPS). We let the dependent variable Y_{it} be the logarithm of wages, and pick the regressor of interest X_{it} and the 5 non-constant controls Z_{it} as a random subset of potential regressor including gender, race, age and dummies for Hispanic, non-white, married, public sector employer, union membership and whether hours or the wage was imputed. The resulting coefficient β on X_{it} in the regression using the entire 145,838 individuals in the database is the population coefficient. We cluster at the level of 308 Metropolitan Statistical Areas (MSAs).[5] The four different methods of Table 4 are then employed to conduct inference about β based on a sample consisting of all individuals that reside in n randomly selected MSAs, where the MSAs are drawn with replacement. Note that in this data generating process, the clusters are i.i.d. and the population regression coefficient is equal to β by construction.

Figure 4 depicts the results over 200 populations generated in this manner, analogous to

[5]For the purposes of this exercise, we treat as additional MSAs the part of each U.S. state outside of any CBSA area.

Figure 3, for $n \in \{50, 100, 200\}$. (We consider $n = 200$ rather than $n = 500$ for the largest sample size to avoid that with high probability, samples contain many identical clusters.) In this design none of the methods come close to perfectly controlling size. Still, the new method is substantially more successful, albeit at the cost of considerably longer average confidence intervals for $n \in \{100, 200\}$.

The poor performance of the standard methods might come as a surprise given that none of the variables in the CPS exercise are heavy-tailed, and the number of clusters is not particularly small. Approximately (cf. equation (29)), the variability of the OLS estimator $\hat{\beta}$ is driven by the average of the n i.i.d. random variables

$$G_i = \sum_{t=1}^{T_i} \tilde{X}_{it} u_{it}$$

where u_{it} is the population regression error and $\tilde{X}_{it}$ is the residual of a population regression of X_{it} on Z_{it}. The distribution of G_i may be heavy-tailed because (i) u_{it} has a heavy-tailed component, as in (33) above; (ii) the joint distribution of $(\tilde{X}_{it}, u_{it})$ is such that $\tilde{X}_{it} u_{it}$ is heavy-tailed; (iii) $\tilde{X}_{it}$ is heavy-tailed; (iv) T_i is heterogeneous across i, so that large clusters with big T_i lead to G_i with high variance; or a combination of these effects. MSAs are highly heterogeneous in their size: the largest contains 6,163 individuals, and the smallest only 42. Effect (iv) is thus clearly present, and the suggestion by Imbens and Kolesar (2016) is designed to accommodate effects (iii) and (iv). But as reported in Table 4, if G_i is heavy-tailed due to effect (i), then the adjustment of Imbens and Kolesar (2016) does not help much. The CPS design seems to exhibit all four effects to some degree, making correct inference quite challenging, and the new method relatively most successful at controlling size.

6 Conclusion

Whenever researchers compare a t-statistic to the usual standard normal critical value they effectively assume that the central limit theorem provides a reasonable approximation. This is true when conducting inference for the mean from an i.i.d. sample, but it holds more generally for linear regression, GMM inference, and so forth. As is well understood, the central limit theorem requires that the contribution of each term to the overall variation is small. To some extent, this is empirically testable: one can simply compare the absolute values of each (demeaned) term with the sample standard deviation. The normal approximation then surely becomes suspect if the largest absolute term is, say, equal to half of a standard deviation.

One may view the new test suggested here as a formalization of this notion: the extreme terms are set apart, and if they are large, then the test automatically becomes more conservative. What is more, even if the sample realization from an underlying population with a heavy tail fails to generate a very large term, it still leaves a tell-tale sign in the large spacings between the largest terms. Correspondingly, the test also becomes more conservative if the largest observations are far apart from each other, even if the largest one isn't all that large—the new method seeks to infer the likelihood of a potential large outlier based on the spacings of the extreme terms. These adjustments are disciplined by an assumption of Pareto-like tails. But as was found in the small sample simulations, they help generate more reliable inference also when the underlying population is more loosely characterized by a moderately heavy tail.

It would be desirable to extend the new method also to F-type tests of null hypotheses that restrict more than one parameter. Such an extension is far from straightforward, though: extreme value theory for the joint extremes of vector-valued observations does not yield a tightly parametrized approximate model, and the approach pursued here to determine the appropriate adjustments is already very computationally challenging.

A Appendix

A.1 Proof of Theorem 2

We write C for a generic large enough positive constant, not necessarily the same in each instance. Without loss of generality, under Condition 1 we can choose w_0 large enough so that uniformly in $w \geq w_0$,

$$C^{-1}w^{-1/\xi} \leq 1 - F(w) \leq Cw^{-1/\xi} \tag{34}$$

$$f(w) \leq Cw^{-1/\xi-1}. \tag{35}$$

Also we normalize $\mathrm{Var}[W] = 1$. Define

$$\begin{aligned} m(w) &= -\mathbb{E}[W|W \leq m],\ m^*(w) = \sigma^{1/\xi}\frac{w^{1-1/\xi}}{1-\xi} \\ A_n &= \mathbf{1}[W_k^R > w_0],\ A_n^* = \mathbf{1}[n^\xi \sigma X_k > w_0], \end{aligned}$$

$V(w) = \mathrm{Var}[W|W \leq w]$ and $\Delta_n = |(1-k/n)^{-1/2}/s_n - 1/\sqrt{V(W_k^R)}|$.

The proof of Theorem 2 is based on a number of preliminary Lemmas. We assume throughout that the assumptions of Theorem 2 hold, and that $n > k+1$. All limits are taken as $n \to \infty$.

Lemma 1 *For any $p > 0$,*

(a) $n^p\mathbb{P}(W_k^R \leq w_0) \to 0$;

(b) $n^p\mathbb{P}(n^\xi\sigma X_k \leq w_0) \to 0$.

Proof. (a) Follows from $\mathbb{P}(W_k^R \leq w_0) = \sum_{i=0}^{k-1}\binom{n}{n-i}F(w_0)^{n-i}(1-F(w_0))^i \leq n^kF(w_0)^{n-k}$ and $F(w_0) < 1$.

(b) Follows from a direct calculation from the density of $(\sum_{l=1}^k E_l)^{-\xi} = X_k/\sigma$. ■

Lemma 2 *(a) For any $p < k/\xi$, $\mathbb{E}[A_n|n^{-\xi}W_k^R|^p] = O(1)$.*

(b) $\mathbb{E}[A_n||n^{-\xi}\mathbf{W}^R||] = O(1)$.

(c) For any $p < 0$, $\mathbb{E}[X_k^p] = O(1)$.

(d) $\mathbb{E}[||\mathbf{X}||^2] < \infty$.

Proof. (a) The density of $S = n^{-\xi}W_k^R$, for $s > n^{-\xi}w_0$, is given by

$$\frac{n!}{(n-k)!(k-1)!}(1-F(n^\xi s))^{k-1}F(n^\xi s)^{n-k}f(n^\xi s)n^\xi \leq F(w_0)^{-k}n^{k+\xi}(1-F(n^\xi s))^{k-1}F(n^\xi s)^n f(n^\xi s).$$

Using (34) and (35), we have $(1 - F(n^\xi s))^{k-1} \leq Cn^{1-k}s^{(1-k)/\xi}$ and $f(n^\xi s) \leq Cn^{-1-\xi}s^{-1/\xi-1}$. Furthermore, using $(1-a/n)^n \leq e^{-a}$ for all $0 \leq a \leq n$ and (34), we have uniformly in $s \geq n^{-\xi}w_0$

$$F(n^\xi s)^n \leq (1 - C^{-1}s^{-1/\xi}/n)^n \leq \exp(-C^{-1}s^{-1/\xi}).$$

Thus, the density of S_n is bounded above by $Cs^{-k/\xi-1}\exp(-C^{-1}s^{-1/\xi})$ on $s\in[n^{-\xi}w_0,\infty)$, and the result follows.

(b) $A_n||\mathbf{W}^R||\leq kW_1^R$, and, proceeding as in the proof of part (a), the density of $n^{-\xi}W_1^R$ for $s>n^{-\xi}w_0$ is bounded above by $Cs^{-1/\xi-1}\exp(-C^{-1}s^{-1/\xi})$ on $s\in[n^{-\xi}w_0,\infty)$, so the result follows.

(c) $\mathbb{E}[X_k^p]=\sigma^p\mathbb{E}[(\sum_{l=1}^k E_l)^{-p\xi}]<\infty$, where the last inequality follows a direct calculation.

(d) $||\mathbf{X}||\leq k\sigma E_1^{-\xi}$, and the result follows by a direct calculation. ■

Lemma 3 *For $w>w_0$, let $\tilde{W}^0$ be a random variable with c.d.f. equal to $F(\tilde{w})/F(w)$ for $\tilde{w}<w$, and equal to one otherwise, and let $\tilde{W}=\tilde{W}^0+m(w)$. Then, uniformly in $w>w_0$*

(a) $m(w)\leq Cw^{-1/\xi+1}$;

(b) $|V(w)-1|\leq Cw^{2-1/\xi}$;

(c) $|m(w)-m^*(w)|\leq Cw^{1-(1+\delta)/\xi}+Cw^{1-2/\xi}$;

(d) for any $\beta_0>1$ and $1<\beta<\beta_0$, $\mathbb{E}[|\tilde{W}|^\beta]\leq Cw^{\beta_0-1/\xi}$;

(e) $\mathbb{E}[\tilde{W}^2\mathbf{1}[\tilde{W}^2>V(w)n]]\leq Cw^2n^{-1/(2\xi)}$;

(f) $\mathbb{E}[|\tilde{W}|^3\mathbf{1}[\tilde{W}^2\leq V(w)n]]\leq Cw^{(1/2-r_k(\xi))/\xi}$.

Proof. (a) Follows from $m(w)=\mathbb{E}[W\mathbf{1}[W>w]]/F(w)$, (35) and $F(w)\geq F(w_0)>0$.

(b) $V(w)=\mathbb{E}[W^2\mathbf{1}[W<w]]/F(w)-m(w)^2$, so that

$$1-V(w)=\frac{F(w)-1}{F(w)}+\frac{1-\mathbb{E}[W^2\mathbf{1}[W<w]]}{F(w)}+m(w)^2.$$

Now for $w>w_0$, $F(w)^{-1}\leq F(w_0)^{-1}$, $1-F(w)\leq Cw^{-1/\xi}$ by (34), and $m(w)\leq Cw^{-1/\xi+1}$ from part (a). Furthermore, using (35)

$$\begin{aligned}1-\mathbb{E}[W^2\mathbf{1}[W<w]] &= \int_w^\infty f(s)s^2ds\leq C\int_w^\infty s^{-1/\xi+1}ds\\ &\leq Cw^{2-1/\xi}\end{aligned}$$

so the result follows.

(c) For $w>w_0$

$$\begin{aligned}|m(w)-m^*(w)| &= \left|\frac{\int_w^\infty sf(s)ds}{F(w)}-\frac{\sigma^{1/\xi}\frac{w^{1-1/\xi}}{1-\xi}}{F(w)}+\frac{\sigma^{1/\xi}\frac{w^{1-1/\xi}}{1-\xi}}{F(w)}-\sigma^{1/\xi}\frac{w^{1-1/\xi}}{1-\xi}\right|\\ &\leq F(w)^{-1}\left|\int_w^\infty sf(s)ds-\int_w^\infty(\xi\sigma)^{-1}(\frac{s}{\sigma})^{-1/\xi-1}ds\right|+\sigma^{1/\xi}\frac{w^{1-1/\xi}}{1-\xi}F(w)^{-1}(1-F(w))\end{aligned}$$

and

$$\left|\int_w^\infty sf(s)ds-\int_w^\infty(\xi\sigma)^{-1}(\frac{s}{\sigma})^{-1/\xi-1}ds\right| \leq C\int_w^\infty s^{-1/\xi}|h(s)|ds$$

$$
\begin{aligned}
&\leq C\int_w^\infty s^{-(\delta+1)/\xi}ds \\
&\leq Cw^{1-(1+\delta)/\xi}
\end{aligned}
$$

and $F(w)^{-1} \leq F(w_0)^{-1}$, $1-F(w) \leq Cw^{-1/\xi}$, so that

$$|m(w) - m^*(w)| \leq Cw^{1-(1+\delta)/\xi} + Cw^{1-2/\xi}.$$

(d) By the c_r inequality and the result of part (a)

$$
\begin{aligned}
\mathbb{E}[|\tilde{W}|^\beta] &= \mathbb{E}[|W+m(w)|^\beta\,|W<w] \\
&\leq C\mathbb{E}[|W|^{\beta_0-1/\xi}|W|^{1/\xi-\beta_0+\beta}\,|W<w] + C|m(w)|^\beta \\
&\leq Cw^{\beta_0-1/\xi}\mathbb{E}[|W|^{1/\xi-\beta_0+\beta}\,|W<w] + Cw^{-\beta/\xi+\beta} \\
&\leq Cw^{\beta_0-1/\xi} + Cw^{-\beta/\xi+\beta}
\end{aligned}
$$

where the last inequality follows from $\mathbb{E}[|W|^{1/\xi-\beta_0+\beta}] < \infty$.

(e) Note that $V(w) \geq V(w_0) > 0$, and by the result in part (a), $|\tilde{W}| \leq Cw$ uniformly in $w \geq w_0$ almost surely. Thus

$$
\begin{aligned}
\mathbb{E}[\tilde{W}^2\mathbf{1}[\tilde{W}^2 > V(w)n]] &\leq \mathbb{E}[\tilde{W}^2\mathbf{1}[\tilde{W}^2 > V(w_0)n]] \\
&\leq Cw^2\mathbb{P}(\tilde{W}^2 > V(w_0)n) \\
&\leq Cw^2\mathbb{P}(W^2 > \tfrac{1}{2}V(w_0)n) + Cw^2\mathbf{1}[m(w)^2 > \tfrac{1}{2}V(w_0)n]
\end{aligned}
$$

where the third inequality uses

$$
\begin{aligned}
\mathbb{P}(|\tilde{W}| > s) &= \mathbb{P}(|W+m(w)| > s\,|W<w) \\
&\leq \mathbb{P}(|W| > \tfrac{1}{2}s\,|W<w) + \mathbf{1}[m(w) > \tfrac{1}{2}s] \\
&\leq F(w_0)^{-1}\mathbb{P}(|W| > \tfrac{1}{2}s) + \mathbf{1}[m(w) > \tfrac{1}{2}s]
\end{aligned}
$$

for all $s > 0$. Now $\mathbf{1}[m(w)^2 > \frac{1}{2}V(w_0)n] = 0$ for all large enough n, since $m(w) \leq C$ uniformly in w from part (a). Finally, $\mathbb{P}(W^2 > \frac{1}{2}V(w_0)n) \leq Cn^{-1/(2\xi)}$ from (34).

(f) Apply part (d) with $\beta_0 = (1/2 - r_k(\xi))/\xi + 1/\xi > 3$ for $\xi \in [1/3, 1/2)$ to obtain

$$\mathbb{E}[|\tilde{W}|^3\mathbf{1}[\tilde{W}^2 \leq V(w)n]] \leq \mathbb{E}[|\tilde{W}|^3] \leq Cw^{\beta_0}.$$

■

Lemma 4 *For any $\epsilon > 0$, $\mathbb{E}[A_n\mathbf{1}[\Delta_n > Cn^{-r_k(\xi)+1/2-\xi}|\mathbf{W}^R] \leq Cn^{-r_k(\xi)+\epsilon}(1 + (n^{-\xi}W_k^R)^{k/\xi-\epsilon})$.*

Proof. We initially prove

$$\mathbb{E}[A_n\mathbf{1}[|s_n^2 - V(W_k^R)| > n^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R] \leq Cn^{-r_k(\xi)+\epsilon}(1 + (W_k^R/n^\xi)^{k/\xi-\epsilon}). \tag{36}$$

With $\tilde{W}$ as defined in Lemma 3, note that by the c_r inequality, for any $\beta > 0$

$$\begin{aligned}
\mathbb{E}[|\tilde{W}^2 - V(w)|^\beta] &\leq C\mathbb{E}[|\tilde{W}|^{2\beta}] + CV(w)^\beta \\
&\leq C\mathbb{E}[|\tilde{W}|^{2\beta}] \qquad (37)
\end{aligned}$$

since $V(w) \leq 1$ and $\mathbb{E}[|\tilde{W}|^{2\beta}] > 0$ uniformly in $w \geq w_0$. Let $\tilde{W}_i$, $i = 1, \ldots, n-k$ be i.i.d. and distributed like $\tilde{W}$, and define $\tilde{Q}_i = (n-k)^{-1}(\tilde{W}_i^2 - V(w))$. Note that $\mathbb{E}[\tilde{W}_i] = \mathbb{E}[\tilde{Q}_i] = 0$. By Rosenthal's (1970) inequality, for any $p > 2$

$$\mathbb{E}\left[\left|\sum_{i=1}^{n-k} \tilde{Q}_i\right|^p\right] \leq C(n-k)\mathbb{E}[|\tilde{Q}_1|^p] + C((n-k)\mathbb{E}[\tilde{Q}_1^2])^{p/2}.$$

Application of (37) and Lemma 3 (d) yields, for $w \geq w_0$, $n > k+1$ and any $p_0 > p > 2$

$$\begin{aligned}
(n-k)\mathbb{E}[|\tilde{Q}_1|^p] &\leq Cn^{1-p}w^{2p_0-1/\xi} \\
&= Cn^{2\xi p_0 - p}(w/n^\xi)^{2p_0-1/\xi} \\
((n-k)\mathbb{E}[\tilde{Q}_1^2])^{p/2} &\leq Cn^{-p/2}\mathbb{E}[|\tilde{W}|^4]^{p/2} \\
&\leq Cn^{-p/2}w^{2p_0-p_0/(2\xi)} \\
&\leq Cn^{2\xi p_0-(p+p_0)/2}(w/n^\xi)^{2p_0-p_0/(2\xi)}
\end{aligned}$$

so that uniformly in $w \geq w_0$

$$\mathbb{E}\left[\left|\sum_{i=1}^{n-k} \tilde{Q}_i\right|^p\right] \leq Cn^{2\xi p_0-p}(1 + (w/n^\xi)^{2p_0-1/\xi}).$$

By Markov's inequality, for any $\alpha \in \mathbb{R}$

$$\mathbb{P}\left(\left|\sum_{i=1}^{n-k} \tilde{Q}_i\right| > \tfrac{1}{2}n^\alpha\right) \leq 2^p \frac{\mathbb{E}\left[\left|\sum_{i=1}^{n-k} \tilde{Q}_i\right|^p\right]}{n^{p\alpha}}.$$

Thus, with $\alpha = -r_k(\xi) + 1/2 - \xi$, $p_0 = (k+1)/(2\xi) - \epsilon/2$ and $p = p_0 - \epsilon/2$, we obtain from some algebra that

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{i=1}^{n-k} Q_i\right| > \tfrac{1}{2}n^{-r_k(\xi)+1/2-\xi}\right) &\leq Cn^{-r_k(\xi)+\epsilon(3/2-2\xi-r_k(\xi))}(1 + (w/n^\xi)^{k/\xi-\epsilon}) \qquad (38) \\
&\leq Cn^{-r_k(\xi)+\epsilon}(1 + (w/n^\xi)^{k/\xi-\epsilon})
\end{aligned}$$

since $3/2 - 2\xi - r_k(\xi) \leq 1$ uniformly in $\xi \in [1/3; 1/2]$. Furthermore, by Markov's inequality

$$\mathbb{P}\left(\left|(n-k)^{-1}\sum_{i=1}^{n-k} \tilde{W}_i\right|^2 > \tfrac{1}{2}n^\alpha\right) \leq 2\frac{(n-k)^{-1}V(w)}{n^\alpha} \qquad (39)$$

$$\leq \quad Cn^{\alpha-1} \leq Cn^{-r_k(\xi)+\epsilon}.$$

Now note that conditional on $\mathbf{W}^R$, $\{W_i^s + m(W_k^R)\}_{i=1}^{n-k}$ has the same distribution as $\{\tilde{W}_i\}_{i=1}^{n-k}$ with $w = W_k^R$. Thus, conditional on $\mathbf{W}^R$, the distribution of

$$s_n^2 - V(W_k^R) = (n-k)^{-1}\sum_{i=1}^{n-k}((W_i^s + m(W_k^R))^2 - V(W_k^R)) - \left((n-k)^{-1}\sum_{i=1}^{n-k}(W_i^s + m(W_k^R))\right)^2$$

is equal to the distribution of $\sum_{i=1}^{n-k} Q_i - \left((n-k)^{-1}\sum_{i=1}^{n-k}\tilde{W}_i\right)^2$ for $w = W_k^R$, so (36) follows from (38) and (39).

To conclude the proof of the lemma, note that $0 < V(w) < \infty$ uniformly in $w \geq w_0$, so for a large enough finite C, $|1/s_n - 1/\sqrt{V(W_k^R)}| > Cn^{-r_k(\xi)+1/2-\xi}$ implies $|s_n^2 - V(W_k^R)| < n^{-r_k(\xi)+1/2-\xi}$. The result thus follows from (36) and $\sup_{w>w_0}|(1-k/n)^{1/2}-1|V(w) = O(n^{-1/2})$. ■

Lemma 5 *(a)* $\mathbb{E}[A_n^*|\mathbf{1}[n^{-1/2+\xi}\sigma\mathbf{X}/\sqrt{V(n^{\xi}\sigma X_k)} \in \mathcal{H}_j] - \mathbf{1}[n^{-1/2+\xi}\sigma\mathbf{X} \in \mathcal{H}_j]|] \leq Cn^{-r_k(\xi)}$*;*

(b) For all $\epsilon > 0$, $|\mathbb{E}[A_n(\mathbf{1}[\mathbf{W}^R/\sqrt{(n-k)s_n^2} \in \mathcal{H}_j] - \mathbf{1}[n^{-1/2}\mathbf{W}^R/\sqrt{V(W_k^R)} \in \mathcal{H}_j])]| \leq Cn^{-r_k(\xi)+\epsilon}$.

Proof. (a) By a first order Taylor expansion $|V(w)^{-1/2} - 1| \leq C|1 - V(w)|$ uniformly in $w \geq w_0$. For $\mathbf{s} \in \mathbb{R}^k$ and $\mathcal{H} \subset \mathbb{R}^k$, let $d(\mathbf{s}, \mathcal{H})$ be the Euclidian distance of the point $\mathbf{s}$ from the set $\mathcal{H}$. We have

$$\begin{aligned}
& \mathbb{E}[A_n^*|\mathbf{1}[n^{-1/2+\xi}\sigma\mathbf{X}/\sqrt{V(n^{\xi}\sigma X_k)} \in \mathcal{H}_j] - \mathbf{1}[n^{-1/2+\xi}\sigma\mathbf{X} \in \mathcal{H}_j]|] \\
\leq \; & \mathbb{E}[A_n^*\mathbf{1}[d(n^{-1/2+\xi}\sigma\mathbf{X}, \partial\mathcal{H}_j) \leq Cn^{-1/2+\xi}||\mathbf{X}|| \cdot |1 - V(n^{\xi}\sigma X_k)|]] \\
\leq \; & \mathbb{E}[A_n^*\mathbf{1}[d(n^{-1/2+\xi}\sigma\mathbf{X}, \partial\mathcal{H}_j) \leq Cn^{-3/2+3\xi}||\mathbf{X}|| \cdot X_k^{2-1/\xi}]] \\
\leq \; & \mathbb{E}[A_n^*\mathbf{1}[d(n^{-1/2+\xi}\sigma\mathbf{X}, \partial\mathcal{H}_j) \leq Cn^{-3/2+3\xi}X_1^{3-1/\xi}]] \\
\leq \; & \mathbb{E}[A_n^*\mathbf{1}[d(n^{-1/2+\xi}\sigma\mathbf{X}, \partial\mathcal{H}_j) \leq Cn^{-r_k(\xi)}(1+X_1)]]
\end{aligned}$$

where the second inequality follows from Lemma 3 (b), and the last inequality holds because $-3/2 + 3\xi \leq -r_k(\xi)$ and $X_1^{3-1/\xi} \leq X_1$ for all $X_1 \geq 1$ for $\xi \in [1/3, 1/2]$.

Furthermore, with $\mathbf{U} = (U_1, \ldots, U_k)' = \mathbf{X}/X_1$,

$$\begin{aligned}
& \mathbb{E}[\mathbf{1}[d(n^{-1/2+\xi}\sigma\mathbf{X}, \partial\mathcal{H}_j) \leq Cn^{-r_k(\xi)}X_1]] \\
\leq \; & \mathbb{E}[\mathbb{E}[d(n^{-1/2+\xi}\sigma(X_1\mathbf{U}), \partial\mathcal{H}_j) \leq Cn^{-r_k(\xi)}(1+X_1)|\mathbf{U}]] \\
= \; & \mathbb{E}[\sum_{s\in\mathcal{I}_j(\mathbf{U})}\mathbb{E}[\mathbf{1}[|n^{-1/2+\xi}\sigma X_1 - s| \leq Cn^{-r_k(\xi)}(1+X_1)|\mathbf{U}]]
\end{aligned}$$

A calculation shows that the conditional density of X_1 given $\mathbf{U}$ is proportional to

$$u_k^{-k/\xi}s^{-k/\xi-1}\exp\left[-(su_k/\sigma)^{-1/\xi}\right],$$

a decreasing function of s for all large enough s, and the density of U_k is bounded above by $Cu_k^{1/\xi-1}$. Thus, for all large enough n,

$$
\begin{aligned}
& \mathbb{E}[\sum_{s\in\mathcal{I}_j(\mathbf{U})} \mathbb{E}[\mathbf{1}[|n^{-1/2+\xi}\sigma X_1 - s| \le Cn^{-r_k(\xi)}(1+X_1)|\mathbf{U}]] \\
\le\ & \mathbb{E}[L\mathbb{E}[\mathbf{1}[|n^{-1/2+\xi}\sigma X_1 - L^{-1}| \le Cn^{-r_k(\xi)}(1+X_1)|\mathbf{U}]] \\
\le\ & \mathbb{E}[L\mathbb{E}[\mathbf{1}[|n^{-1/2+\xi}\sigma X_1 - L^{-1}| \le Cn^{-r_k(\xi)}X_1|\mathbf{U}]] \\
\le\ & C(n^{1/2-\xi})^{-k/\xi-1}n^{-r_k(\xi)}n^{1/2-\xi}\int_0^1 \exp\left[-(L^{-1}n^{1/2-\xi}u_k/\sigma)^{-1/\xi}\right]u_k^{-(k-1)/\xi-1}du_k \\
\le\ & Cn^{-r_k(\xi)-(1/2-\xi)k/\xi}\cdot n^{(1/2-\xi)(k-1)/\xi} \\
=\ & Cn^{-r_k(\xi)-(1/2-\xi)/\xi}
\end{aligned}
$$

and the result follows.

(b) Let $D_n = \mathbf{1}[\Delta_n \le Cn^{-r_k(\xi)+1/2-\xi}]$. Using Lemmas 4 and 2 (a), we have for all $\epsilon > 0$

$$
\begin{aligned}
\mathbb{E}[A_n(1-D_n)] &= \mathbb{E}[\mathbb{E}[A_n\mathbf{1}[\tilde{\Delta}_n > Cn^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R]] \\
&\le Cn^{-r_k(\xi)+\epsilon}(1+\mathbb{E}[(n^{-\xi}W_k^R)^{k/\xi-\epsilon}]) \\
&\le Cn^{-r_k(\xi)+\epsilon}
\end{aligned}
$$

so it suffices to show the claim with A_n replaced by A_nD_n.

In the notation of the proof of part (a), we have

$$
\begin{aligned}
& \mathbb{E}[A_nD_n|\mathbf{1}[\mathbf{W}^R/\sqrt{(n-k)s_n^2} \in \mathcal{H}_j] - \mathbf{1}[n^{-1/2}\mathbf{W}^R/\sqrt{V(W_k^R)} \in \mathcal{H}_j]|] \\
\le\ & \mathbb{E}[A_nD_n\mathbf{1}[d(n^{-1/2}\mathbf{W}^R/\sqrt{V(W_k^R)}, \partial\mathcal{H}_j) \le Cn^{-1/2}\Delta_n||\mathbf{W}^R||]] \\
\le\ & \mathbb{E}[A_n\mathbf{1}[d(n^{-1/2}\mathbf{W}^R/\sqrt{V(W_k^R)}, \partial\mathcal{H}_j) \le Cn^{-r_k(\xi)-\xi}W_1^R]] \\
\le\ & \mathbb{E}[A_n^*\mathbf{1}[d(n^{-1/2+\xi}\sigma\mathbf{X}/\sqrt{V(n^\xi\sigma X_k)}, \partial\mathcal{H}_j) \le Cn^{-r_k(\xi)}\sigma X_1]] + n^{-\delta} \\
\le\ & \mathbb{E}[A_n^*\mathbf{1}[d(n^{-1/2+\xi}\sigma\mathbf{X}, \partial\mathcal{H}_j) \le Cn^{-r_k(\xi)}\sigma X_1]] + n^{-\delta} + Cn^{-r_k(\xi)}
\end{aligned}
$$

where the penultimate inequality follows from (5), and the last inequality applies the result form part (a). The desired inequality now follows from the same arguments as in the proof of part (a). ■

Proof of Theorem 2:

Let $B_n = \mathbf{1}[W_k^R > w_0]\mathbf{1}[n^{-1/2}\mathbf{W}^R/\sqrt{V(W_k^R)} \in \mathcal{H}_j]$ and $B_n^* = \mathbf{1}[n^\xi\sigma X_k > w_0]\mathbf{1}[n^{\xi-1/2}\sigma\mathbf{X} \in \mathcal{H}_j]$. Given the results in Lemmas 1 and 5 (a), it suffices to show that

$$
\left|\mathbb{E}B_n\mathbf{1}\left[\frac{\sum_{i=1}^{n-k}W_i^s}{\sqrt{(n-k)s_n^2}} \le b_j\left(\frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}}\right)\right]\right.
$$

$$-\mathbb{E}B_n^*\mathbf{1}\left[Z - n^{1/2}m^*(n^\xi\sigma X_k) \le b_j\left(n^{\xi-1/2}\sigma\mathbf{X}\right)\right]\Big| \le Cn^{-r_k(\xi)+\epsilon}$$

for all $j = 1, \ldots, m_\varphi$.

Notice that conditional on $\mathbf{W}^R$,

$$\varsigma_n = \frac{\sum_{i=1}^{n-k}(W_i^s + m(W_k^R))}{\sqrt{(n-k)s_n^2}}$$

has the same distribution as the t-statistic computed from the zero-mean i.i.d. sample $\tilde{W}_1, \tilde{W}_2, \ldots, \tilde{W}_n$ with $\tilde{W}_i \sim \tilde{W}$ and $\tilde{W}$ defined in Lemma 3 with $w = W_k^R$. Thus, by (6),

$$\begin{aligned} A_n \sup_s |\mathbb{P}(\varsigma_n \le s|\mathbf{W}^R) - \Phi(x)| &\le A_n CV(W_k^R)^{-1}\mathbb{E}[\tilde{W}^2\mathbf{1}[\tilde{W}^2 > V(W_k^R)n] \\ &\quad + A_n Cn^{-1/2}V(W_k^R)^{-3/2}\mathbb{E}[|\tilde{W}|^3\mathbf{1}[\tilde{W}^2 \le V(W_k^R)n]]. \end{aligned}$$

Using $V(W_k^R) \ge V(w_0) > 0$ if $A_n = 1$ and applying Lemma 3 (e) and (f), the right-hand side is bounded above by

$$\begin{aligned} & A_n C(n^{-\xi}W_k^R)^2 n^{2\xi-1/(2\xi)} + A_n C(n^{-\xi}W_k^R)^{(1/2-r_k(\xi))/\xi} n^{-r_k(\xi)} \\ & \qquad \le A_n Cn^{-r_k(\xi)}(1 + (n^{-\xi}W_k^R)^2) := L_{1,n}(W_k^R) \end{aligned} \tag{40}$$

since $2\xi - 1/(2\xi) \le -r_k(\xi)$ and $0 < (1/2 - r_k(\xi))/\xi \le 1$ for all $\xi \in [1/3, 1/2]$ and $k > 1$.

By the Lipschitz continuity of b_j

$$\left| b_j\left(\frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}}\right) - b_j\left(\frac{n^{-1/2}\mathbf{W}^R}{\sqrt{V(W_k^R)}}\right)\right| \le C\Delta_n n^{-1/2}||\mathbf{W}^R||,$$

and defining

$$\begin{aligned} \hat{M} &= \frac{(n-k)m(W_k^R)}{\sqrt{(n-k)s_n^2}}, \; \tilde{M} = \frac{(n-k)m(W_k^R)}{n^{1/2}\sqrt{V(W_k^R)}} \\ R_n &= b_j(\mathbf{W}^R/\sqrt{(n-k)s_n^2}) - b_j(n^{-1/2}\mathbf{W}^R/\sqrt{V(W_k^R)}) + \hat{M} - \tilde{M} \end{aligned}$$

we have

$$\begin{aligned} |A_n R_n| &\le A_n\Delta_n\left(n^{1/2}|m(W_k^R)| + Cn^{-1/2}||\mathbf{W}^R||\right) \\ &\le \Delta_n A_n Cn^{\xi-1/2}((n^{-\xi}W_k^R)^{1-1/\xi} + n^{-\xi}||\mathbf{W}^R||) := \Delta_n L_{2,n}(\mathbf{W}^R) \end{aligned}$$

where the second inequality invoked Lemma 3 (a). In this notation

$$\mathbf{1}\left[\frac{\sum_{i=1}^{n-k}W_i^s}{\sqrt{(n-k)s_n^2}} \le b_j\left(\frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}}\right)\right] = \mathbf{1}\left[\varsigma_n \le R_n + b_j\left(\frac{n^{-1/2}\mathbf{W}^R}{\sqrt{V(W_k^R)}}\right) + \tilde{M}\right].$$

From Lemma 4,

$$\mathbb{E}[A_n\mathbf{1}[\Delta_n > Cn^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R] \le CA_n n^{-r_k(\xi)+\epsilon}(1+(n^{-\xi}W_k^R)^{k/\xi-\epsilon}) := L_{3,n}(W_k^R).$$

Thus, uniformly in $s\in\mathbb{R}$,

$$\begin{aligned}
& \mathbb{E}[B_n\mathbf{1}[\varsigma_n \le s+R_n|\mathbf{W}^R]] \\
\le\ & \mathbb{E}[B_n\mathbf{1}[\varsigma_n \le s+R_n]\mathbf{1}[\Delta_n \le Cn^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R] + \mathbb{E}[A_n\mathbf{1}[\Delta_n > Cn^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R] \\
\le\ & \mathbb{E}[B_n\mathbf{1}[\varsigma_n \le s+Cn^{-r_k(\xi)+1/2-\xi}L_{2,n}(\mathbf{W}^R)|\mathbf{W}^R] + L_{3,n}(W_k^R) \\
\le\ & B_n\Phi(s+Cn^{-r_k(\xi)+1/2-\xi}L_{2,n}(\mathbf{W}^R)) + L_{1,n}(W_k^R) + L_{3,n}(W_k^R) \\
\le\ & B_n\Phi(s) + L_{1,n}(W_k^R) + Cn^{-r_k(\xi)+1/2-\xi}L_{2,n}(\mathbf{W}^R) + L_{3,n}(W_k^R)
\end{aligned}$$

where the third inequality follows from (40), and the fourth inequality follows from an exact first order Taylor expansion and the fact that the derivative of Φ is uniformly bounded. Thus, letting $s = b_j(\mathbf{W}^R/\sqrt{nV(W_k^R)}) + \tilde{M}$ and taking expectations, we obtain

$$\begin{aligned}
& \mathbb{E}\left[B_n\mathbf{1}\left[\frac{\sum_{i=1}^{n-k}W_i^s}{\sqrt{(n-k)s_n^2}} \le b_j\left(\frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}}\right)\right]\right] - \mathbb{E}\left[B_n\mathbf{1}\left[Z \le b_j\left(\frac{\mathbf{W}^R}{n^{1/2}\sqrt{V(W_k^R)}}\right)+\tilde{M}\right]\right] \\
\le\ & \mathbb{E}[L_{1,n}(W_k^R)] + Cn^{-r_k(\xi)+1/2-\xi}\mathbb{E}[L_{2,n}(\mathbf{W}^R)] + \mathbb{E}[L_{3,n}(W_k^R)]. \qquad (41)
\end{aligned}$$

Similarly, uniformly in $s\in\mathbb{R}$,

$$\begin{aligned}
& \mathbb{E}[B_n\mathbf{1}[\varsigma_n \le s+R_n|\mathbf{W}^R]] \\
\ge\ & \mathbb{E}[B_n\mathbf{1}[\varsigma_n \le s+R_n]\mathbf{1}[\Delta_n \le Cn^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R] - \mathbb{E}[A_n\mathbf{1}[\Delta_n > Cn^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R] \\
\ge\ & B_n\Phi(s) - L_{1,n}(W_k^R) - Cn^{-r_k(\xi)+1/2-\xi}L_{2,n}(\mathbf{W}^R) - L_{3,n}(W_k^R)
\end{aligned}$$

so that (41) holds with the left hand side replaced by its absolute value. By an application of Lemma 2 (a) and (b), the right hand side of (41) is $O(n^{-r_k(\xi)+\epsilon})$.

Furthermore, with $B_n^{**} = A_n^*\mathbf{1}[n^{\xi-1/2}\sigma\mathbf{X}/\sqrt{V(n^\xi\sigma X_k)} \in \mathcal{H}_j]$ and $\tilde{M}^* = n^{-1/2}(n-k)m(n^\xi\sigma X_k)/\sqrt{V(n^\xi\sigma X_k)}$, by (5),

$$\begin{aligned}
& \left|\mathbb{E}\left[B_n\mathbf{1}[Z \le b_j\left(\frac{\mathbf{W}^R}{n^{1/2}\sqrt{V(W_k^R)}}\right)+\tilde{M}]\right] - \mathbb{E}\left[B_n^{**}\mathbf{1}[Z \le b_j\left(\frac{n^\xi\sigma\mathbf{X}}{n^{1/2}\sqrt{V(n^\xi\sigma X_k)}}\right)+\tilde{M}^*\right]\right| \\
\le\ & Cn^{-\delta}.
\end{aligned}$$

By Lemma 5 (b), replacing B_n^{**} by B_n^* in this expression yields an additional approximation error of order at most $O(n^{-r_k(\xi)+\epsilon})$.

By a first order Taylor expansion $|1-V(w)^{-1/2}| \leq C|1-V(w)|$ uniformly in $w \geq w_0$. Thus, by the assumption about b_j, and using again the fact that the derivative of Φ is uniformly bounded,

$$\begin{aligned}
&\left|\mathbb{E}\left[B_n^*\Phi\left(b_j\left(\frac{n^\xi\sigma\mathbf{X}}{n^{1/2}\sqrt{V(n^\xi\sigma X_k)}}\right)+\tilde{M}^*\right)-B_n^*\Phi\left(b_j\left(n^{\xi-1/2}\sigma\mathbf{X}\right)+n^{1/2}m^*(n^\xi\sigma X_k)\right)\right]\right| \\
\leq\ & Cn^{-1/2}\mathbb{E}[B_n^*(||n^\xi\mathbf{X}||+(n-k)|m(n^\xi\sigma X_k)|)|1-V(n^\xi\sigma X_k)|] \\
&\qquad\qquad +Cn^{1/2}\mathbb{E}[B_n^*|m(n^\xi\sigma X_k)-m^*(n^\xi\sigma X_k)|].
\end{aligned}$$

By the Cauchy-Schwarz inequality and Lemma 3 (b),

$$\begin{aligned}
n^{-1/2}\mathbb{E}[B_n^*||n^\xi\mathbf{X}||\cdot|1-V(n^\xi\sigma X_k)|] &\leq n^{-1/2}\mathbb{E}[||n^\xi\mathbf{X}||^2]^{1/2}\mathbb{E}[B_n^*|1-V(n^\xi\sigma X_k)|^2]^{1/2} \\
&\leq Cn^{3(\xi-1/2)}\mathbb{E}[||\mathbf{X}||^2]^{1/2}\mathbb{E}[X_k^{4-2/\xi}]^{1/2} \qquad (42)
\end{aligned}$$

and by the Cauchy-Schwarz inequality and Lemma 3 (a) and (b),

$$\begin{aligned}
n^{1/2}\mathbb{E}[B_n^*|m(n^\xi\sigma X_k)|\cdot|1-V(n^\xi\sigma X_k)|] &\leq n^{1/2}\mathbb{E}[B_n^*|m(n^\xi\sigma X_k)|^2]^{1/2}\mathbb{E}[B_n^*|1-V(n^\xi\sigma X_k)|^2]^{1/2} \\
&\leq Cn^{3(\xi-1/2)}\mathbb{E}[X_k^{2-2/\xi}]^{1/2}\mathbb{E}[X_k^{4-2/\xi}]^{1/2}. \qquad (43)
\end{aligned}$$

Finally, by Lemma 3 (c),

$$n^{1/2}\mathbb{E}[B_n^*|m(n^\xi\sigma X_k)-m^*(n^\xi\sigma X_k)|] \leq Cn^{-1/2+\xi-\delta}\mathbb{E}[X_k^{1-(1+\delta)/\xi}]+Cn^{\xi-3/2}\mathbb{E}[X_k^{1-2/\xi}]. \qquad (44)$$

Note that $3(\xi-1/2)$, $\xi-3/2$ and $-1/2+\xi-\delta$ for $\delta \geq r_k(\xi)$ are weakly smaller than $-r_k(\xi)$ for all $\xi \in [1/3, 1/2]$, so the result follows from applying Lemma 2 (c) and (d) to (42)-(44).

A.2 Generalizing Theorem 2 to Two Potentially Heavy Tails

Condition 2 *Suppose for some $\xi^R, \sigma^R, \xi^L, \sigma^L, \delta, w_0 > 0$, F admits a density for $w > w_0$ of the form*

$$f^R(w) = (\xi^R\sigma^R)^{-1}(\frac{w}{\sigma^R})^{-1/\xi^R-1}(1+h^R(w))$$

and a density for $w < -w_0$ of the form

$$f^L(w) = (\xi^L\sigma^L)^{-1}(\frac{-w}{\sigma^L})^{-1/\xi^L-1}(1+h^L(-w))$$

with $|h^J(w)|$ uniformly bounded by $Cw^{-\delta/\xi^J}$ for $J \in \{L, R\}$ and some finite C.

Theorem 3 *Suppose Condition 2 holds, and for $k > 1$, $r_k(\xi) = \frac{3(1+k)(1-2\xi)}{2(1+k+2\xi)} \leq \delta$ where $\xi = \max(\xi^L, \xi^R)$. Let $\varphi : \mathbb{R}^{2k+1} \mapsto \{0,1\}$ be such that for some finite m_φ, $\varphi : \mathbb{R}^{2k+1} \mapsto \{0,1\}$ can be written as an affine function of $\{\varphi_j\}_{j=1}^{m_\varphi}$, where each φ_j is of the form*

$$\varphi_j(\mathbf{y}^R, \mathbf{y}^L, y_0) = \mathbf{1}[(\mathbf{y}^R, \mathbf{y}^L) \in \mathcal{H}_j]\mathbf{1}[y_0 \leq b_j(\mathbf{y}^R, \mathbf{y}^L)]$$

with $b_j : \mathbb{R}^{2k} \mapsto \mathbb{R}$ a Lipschitz continuous function and $\mathcal{H}_j$ a Borel measurable subset of $\mathbb{R}^{2k}$ with boundary $\partial\mathcal{H}_j$. For $\mathbf{u}^J = (1, u_2, \ldots, u_k)' \in \mathbb{R}^k$ with $1 \geq u_2 \geq u_3 \geq \ldots \geq u_k$, let $\mathcal{I}_j(\mathbf{u}^R, \mathbf{u}^L) = \{s^R, s^L > 0 : (s^R\mathbf{u}^R, s^L\mathbf{u}^L) \in \partial\mathcal{H}_j\}$. Assume further that for some $L > 0$, and Lebesgue almost all $\mathbf{u}^R, \mathbf{u}^L$, $\mathcal{I}_j(\mathbf{u}^R, \mathbf{u}^L)$ contains at most L elements in the interval $[L^{-1}, \infty)^2$.

Then under $H_0 : \mu = 0$, for $\frac{1+k}{1+3k} < \xi < 1/2$ and any $\epsilon > 0$

$$|\mathbb{E}[\varphi(\mathbf{Y}_n)] - \mathbb{E}[\varphi(\mathbf{Y}_n^*)]| \leq Cn^{-r_k(\xi)+\epsilon}$$

where $\mathbf{Y}_n$ and $\mathbf{Y}_n^*$ are the l.h.s. and r.h.s. of (16), respectively.

The proof of Theorem 3 follows from the same steps as Theorem 2 and is omitted for brevity.

Tests of the form described in Section 3 reject if all of the following four conditions hold

$$\begin{aligned}
h_1(\mathbf{y}^*) &= \sum_{i=1}^{M} \lambda_i \frac{f(\mathbf{y}^*|\theta_i, 0)}{f_a^S(\mathbf{y}^R) f_a^S(\mathbf{y}^L)} \leq 1 \\
h_2(\mathbf{y}^*) &= \exp[-5\chi(\mathbf{y}^L)] \sum_{i=1}^{M^S} \lambda_i^S \frac{f^S(\mathbf{y}^R, \mathbf{y}^L, y_0|\theta_i^S)}{f_a^S(\mathbf{y}^R)} \leq 1 \\
h_3(\mathbf{y}^*) &= \exp[-5\chi(\mathbf{y}^R)] \sum_{i=1}^{M^S} \lambda_i^S \frac{f^S(\mathbf{y}^L, \mathbf{y}^R, y_0|\theta_i^S)}{f_a^S(\mathbf{y}^L)} \leq 1 \\
h_4(\mathbf{y}^*) &= \frac{(y_0 + \sum_{i=1}^k y_i^R - \sum_{i=1}^k y_i^L)^2}{\mathrm{cv}_T(\mathbf{y}^*)^2(1 + \sum_{i=1}^k (y_i^R)^2 + \sum_{i=1}^k (y_i^L)^2)} \leq 1.
\end{aligned}$$

Inspection of the densities f in (19) and f^S in (25) shows that $h_j(\mathbf{y}^*)$, $j = 1, 2, 3$, viewed as function of y_0, are a linear combination of normal densities, with weights, means and variances a function of $(\mathbf{y}^R, \mathbf{y}^L)$. For any given $(\mathbf{y}^R, \mathbf{y}^L)$, the acceptance region of $\varphi(\mathbf{y}^*)$, $\{y_0 : \varphi(\mathbf{y}^*) = 0$ for $\mathbf{y}^* = (\mathbf{y}^R, \mathbf{y}^L, y_0)\}$, is thus a finite union of intervals, which an be written as an affine function of appropriately defined φ_j. By the inverse function theorem, the endpoints of these intervals are a function of $(\mathbf{y}^R, \mathbf{y}^L)$ with derivative equal to $-1/(\partial h_j(\mathbf{y}^*)/\partial y_0) = 0$ at $\mathbf{y}^*$ with $h_j(\mathbf{y}^*) = 1$. Define the sets

$$\mathcal{E}_j(\varepsilon) = \{(\mathbf{y}^R, \mathbf{y}^L) : \exists y_0 \in \mathbb{R} \text{ such that } h_j(\mathbf{y}^*) = 1 \text{ and } |\partial h_j(\mathbf{y}^*)/\partial y_0| < \varepsilon\} \in \mathbb{R}^{2k}$$

for $\varepsilon > 0$, $j = 1, \ldots, 4$, which have Lebesgue measure that can be made arbitrarily small for small enough ε. Furthermore, for any $\theta^S = (\mu, \sigma, \xi) \in \mathbb{R}^3$ and $c > 0$, the equation $f_T(s\mathbf{u}^J|\theta^S)/f_a^S(s\mathbf{u}^J) = c$ for a given $\mathbf{u}^J$ has at most two roots in $s > 0$. The modified test $\varphi_\varepsilon(\mathbf{y}^*) = \varphi(\mathbf{y}^*)\mathbf{1}[(\mathbf{y}^R, \mathbf{y}^L) \notin \bigcup_{j=1}^3 \mathcal{E}_j(\varepsilon)]$ is hence seen to satisfy the assumptions of Theorem 3, yet for small enough ε, φ_ε and φ have nearly indistinguishable rejection probabilities.

A.3 Results for $\alpha = 0.01$

Table 5: Small Sample Results in Inference for the Mean

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
T-STAT	0.9|**1.00**	5.1|0.62	6.8|0.58	0.7|**1.04**	7.8|0.53	2.7|0.82	10.8|0.60
SYM-BOOT	0.9|**1.02**	2.9|1.04	4.0|1.31	0.5|**1.19**	4.1|1.29	2.6|1.12	10.3|1.58
ASYM-BOOT	1.0|**1.02**	2.1|0.85	2.6|0.97	1.9|**1.10**	2.7|0.95	2.6|0.96	9.4|1.08
NEW DEFAULT	0.3|**1.39**	0.6|**0.79**	1.1|**0.69**	0.2|**1.57**	1.8|**0.63**	0.5|**1.10**	3.9|0.70
				$n = 100$			
T-STAT	1.1|**0.99**	3.6|0.74	5.3|0.65	0.8|**1.02**	5.9|0.62	2.6|0.82	9.3|0.57
SYM-BOOT	1.1|**0.99**	2.2|1.05	3.5|1.22	0.7|**1.09**	3.5|1.25	2.4|1.03	8.5|1.23
ASYM-BOOT	1.1|**0.99**	1.6|**0.90**	2.4|0.94	1.9|**1.04**	2.5|0.95	2.4|0.92	7.3|0.90
NEW DEFAULT	0.7|**1.08**	0.2|**1.22**	0.5|**0.97**	0.4|**1.67**	0.5|**0.92**	0.5|**1.31**	3.7|0.73
				$n = 500$			
T-STAT	1.0|**1.00**	1.9|**0.88**	3.0|0.80	0.8|**1.02**	3.5|0.78	1.5|**0.93**	4.2|0.69
SYM-BOOT	1.1|**1.00**	1.5|**0.99**	2.1|1.09	0.8|**1.04**	2.6|1.13	1.4|**1.01**	3.0|1.04
ASYM-BOOT	1.1|**1.00**	1.3|**0.93**	1.7|**0.94**	1.6|**1.02**	1.8|**0.95**	1.4|**0.97**	2.3|0.86
NEW DEFAULT	1.0|**1.01**	0.6|**1.33**	0.7|**1.34**	0.5|**1.32**	0.7|**1.30**	0.5|**1.37**	0.4|**1.14**

Notes: Entries are the null rejection probability in percent, and the average length of confidence intervals relative to average length of confidence intervals based on size corrected t-statistic (bold if null rejection probability is smaller than 2%) of nominal 1% level tests. Based on 20,000 replications.

Table 6: Small Sample Results of New Methods for Inference for the Mean

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 25$			
DEF: $k = 8, n_0 = 50$	0.6|**1.07**	6.7|0.57	8.9|0.50	0.4|**1.10**	10.3|0.46	2.1|0.84	6.6|0.68
$k = 4, n_0 = 50$	0.3|**1.34**	3.4|0.60	4.9|0.53	0.3|**1.22**	6.4|0.47	1.2|**0.88**	3.0|0.70
$k = 12, n_0 = 50$	0.0|**0.00**	0.0|**0.00**	0.0|**0.00**	0.0|**0.00**	0.0|**0.00**	0.0|**0.00**	0.0|**0.00**
$k = 4, n_0 = 25$	0.1|**1.71**	1.8|**0.65**	2.3|0.58	0.1|**1.54**	3.0|0.53	0.5|**1.04**	1.4|**0.77**
				$n = 50$			
DEF: $k = 8, n_0 = 50$	0.3|**1.39**	0.6|**0.79**	1.1|**0.69**	0.2|**1.57**	1.8|**0.63**	0.5|**1.10**	3.9|0.70
$k = 4, n_0 = 50$	0.3|**1.50**	0.5|**0.83**	0.9|**0.69**	0.2|**1.65**	1.3|**0.64**	0.5|**1.15**	3.4|0.70
$k = 12, n_0 = 50$	0.1|**1.50**	1.4|**0.75**	2.2|0.65	0.2|**1.54**	3.3|0.59	0.4|**1.03**	2.6|0.70
$k = 4, n_0 = 25$	0.4|**1.59**	0.3|**0.97**	0.5|**0.78**	0.2|**2.02**	0.6|**0.73**	0.5|**1.34**	3.5|0.79
				$n = 100$			
DEF: $k = 8, n_0 = 50$	0.7|**1.08**	0.2|**1.22**	0.5|**0.97**	0.4|**1.67**	0.5|**0.92**	0.5|**1.31**	3.7|0.73
$k = 4, n_0 = 50$	0.6|**1.22**	0.3|**1.06**	0.6|**0.85**	0.3|**1.85**	0.7|**0.81**	0.5|**1.27**	2.2|0.69
$k = 12, n_0 = 50$	0.6|**1.14**	0.3|**1.22**	0.4|**1.00**	0.5|**1.56**	0.5|**0.92**	0.3|**1.35**	3.3|0.75
$k = 4, n_0 = 25$	0.7|**1.20**	0.4|**1.21**	0.5|**0.98**	0.3|**2.13**	0.6|**0.95**	0.4|**1.47**	2.5|0.76
				$n = 500$			
DEF: $k = 8, n_0 = 50$	1.0|**1.01**	0.6|**1.33**	0.7|**1.34**	0.5|**1.32**	0.7|**1.30**	0.5|**1.37**	0.4|**1.14**
$k = 4, n_0 = 50$	1.1|**0.99**	0.6|**1.45**	0.7|**1.20**	0.5|**1.72**	0.8|**1.20**	0.6|**1.49**	0.5|**0.95**
$k = 12, n_0 = 50$	1.1|**0.99**	0.5|**1.29**	0.7|**1.20**	0.5|**1.30**	0.6|**1.22**	0.5|**1.31**	0.4|**1.23**
$k = 4, n_0 = 25$	0.9|**1.01**	0.6|**1.54**	0.7|**1.34**	0.5|**1.84**	0.7|**1.34**	0.6|**1.64**	0.5|**1.08**

Notes: See Table 5.

Figure 5: Small Sample Results for HMDA Populations

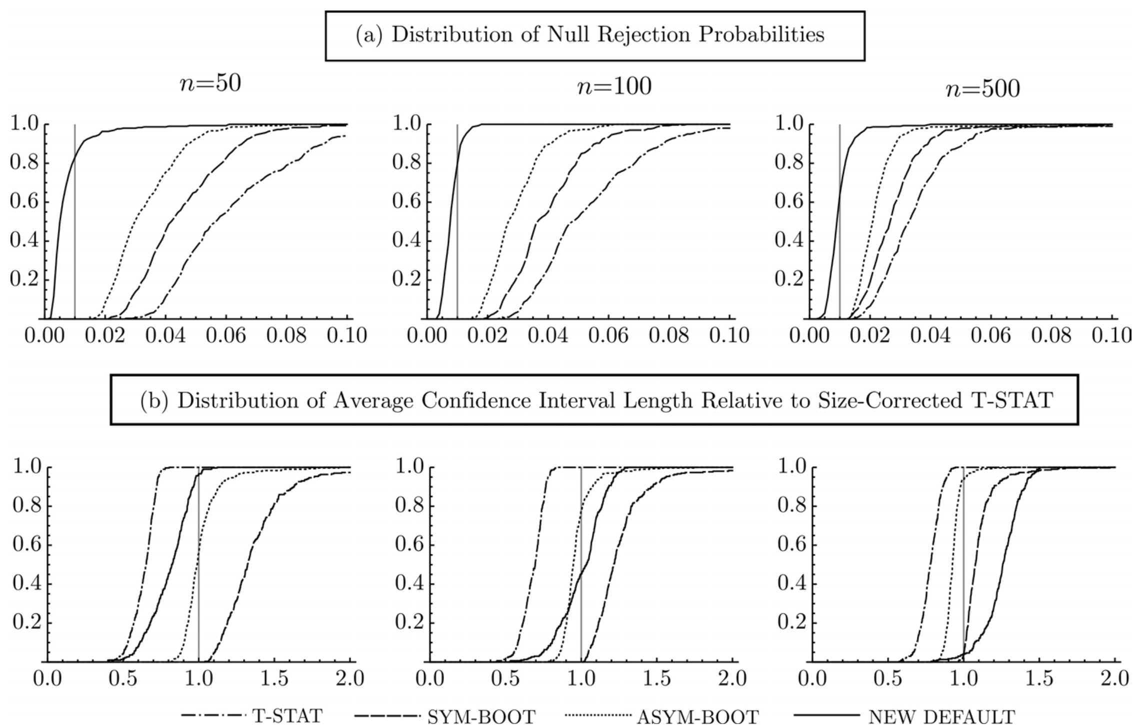


Table 7: Small Sample Results for Difference of Population Means

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
T-STAT	1.4|**0.94**	3.9|0.75	3.3|0.78	1.0|**0.99**	3.4|0.78	2.1|0.87	3.1|0.81
SYM-BOOT	1.4|**0.95**	3.6|0.97	3.1|1.02	0.9|**1.08**	3.3|1.05	2.1|1.02	3.1|1.12
ASYM-BOOT	1.6|**0.94**	3.2|0.83	3.1|0.87	2.3|1.01	3.2|0.89	2.7|0.92	3.2|0.93
NEW DEFAULT	0.1|**1.46**	0.7|**1.02**	0.7|**1.07**	0.1|**1.49**	0.7|**1.07**	0.5|**1.25**	0.9|**1.06**
				$n = 100$			
T-STAT	1.0|**0.99**	3.1|0.78	3.1|0.80	1.0|**1.00**	3.0|0.81	1.8|**0.89**	3.6|0.80
SYM-BOOT	1.1|**0.99**	2.8|0.99	3.0|1.08	0.9|**1.05**	2.9|1.10	1.8|**1.03**	3.6|1.15
ASYM-BOOT	1.1|**0.99**	2.5|0.87	2.6|0.92	2.0|1.01	2.5|0.93	2.1|0.94	3.6|0.94
NEW DEFAULT	0.3|**1.37**	0.5|**1.23**	0.9|**1.21**	0.4|**1.74**	0.9|**1.21**	0.5|**1.41**	1.9|**1.06**
				$n = 500$			
T-STAT	1.0|**1.00**	1.7|**0.92**	2.0|0.87	0.9|**1.01**	2.3|0.85	1.4|**0.94**	3.3|0.78
SYM-BOOT	1.1|**0.99**	1.5|**1.02**	1.9|**1.09**	0.9|**1.03**	2.2|1.08	1.3|**1.02**	3.1|1.06
ASYM-BOOT	1.1|**1.00**	1.5|**0.96**	1.9|**0.96**	1.5|**1.01**	1.9|**0.95**	1.5|**0.97**	2.9|0.90
NEW DEFAULT	0.9|**1.01**	0.5|**1.44**	0.7|**1.40**	0.7|**1.32**	0.7|**1.37**	0.6|**1.43**	1.2|**1.38**

Notes: See Table 5.

Figure 6: Small Sample Results for Two Samples from HDMA Populations

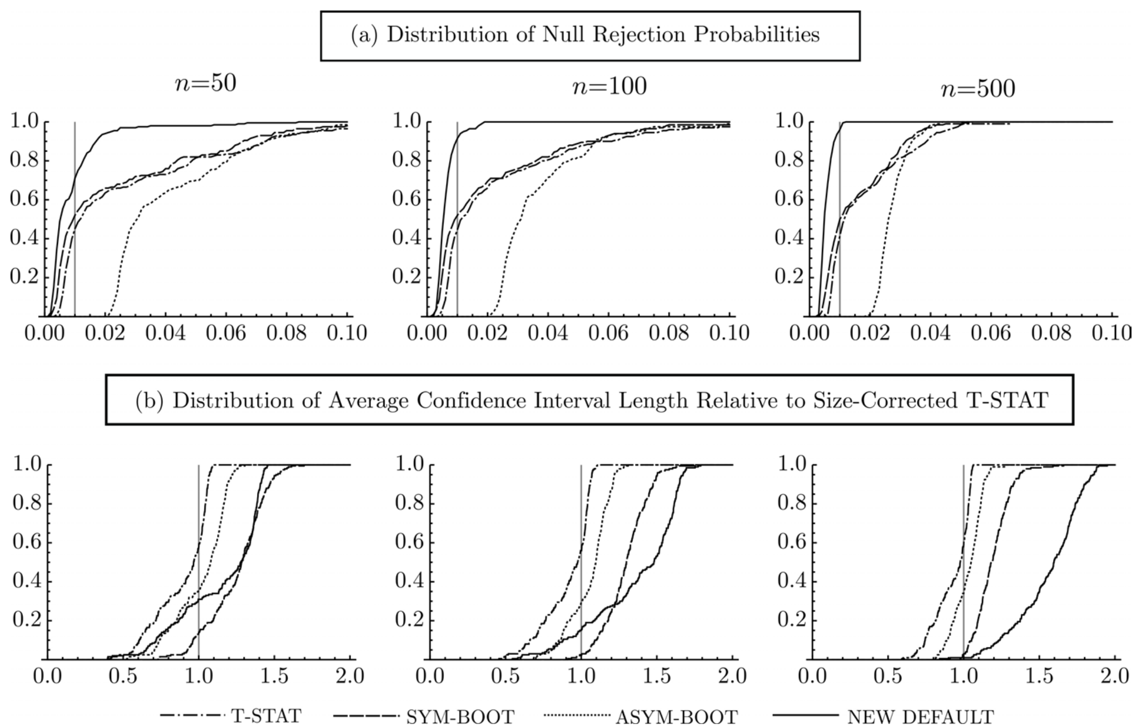
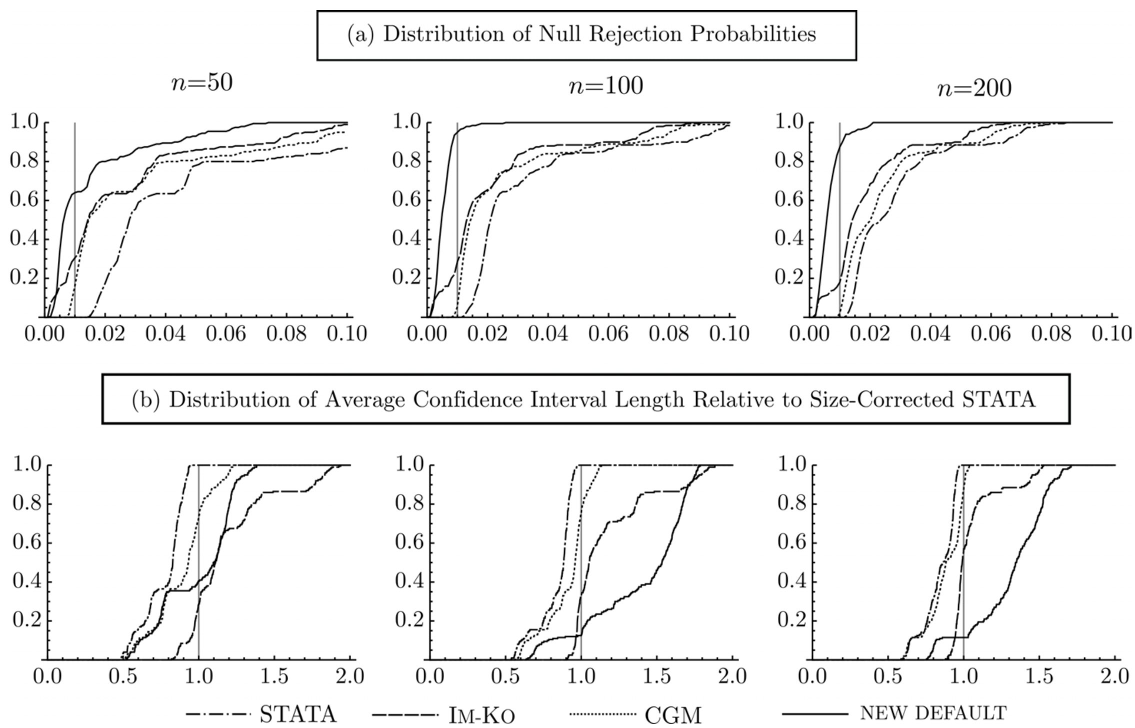


Table 8: Small Sample Results in Clustered Regression Design

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
STATA	1.0|**1.00**	3.8|0.72	4.3|0.72	0.9|**1.02**	4.9|0.71	2.1|0.86	4.8|0.74
IM-KO	1.0|**1.01**	3.7|0.73	4.1|0.73	0.8|**1.03**	4.7|0.72	2.0|**0.88**	4.6|0.75
CGM	0.9|**1.02**	3.7|0.68	4.1|0.66	1.0|**0.99**	4.7|0.65	2.0|0.83	4.6|0.65
NEW DEFAULT	0.2|**1.55**	0.5|**0.91**	0.7|**0.91**	0.3|**1.49**	0.7|**0.89**	0.5|**1.18**	1.4|**0.92**
				$n = 100$			
STATA	1.1|**0.99**	3.3|0.78	3.9|0.76	0.8|**1.03**	4.2|0.73	2.1|0.88	5.7|0.70
IM-KO	1.1|**0.99**	3.2|0.78	3.8|0.76	0.8|**1.03**	4.1|0.74	2.0|**0.88**	5.6|0.71
CGM	1.1|**1.00**	3.3|0.73	3.9|0.69	1.0|**1.00**	4.2|0.66	2.2|0.84	5.7|0.61
NEW DEFAULT	0.5|**1.33**	0.4|**1.22**	0.7|**1.13**	0.4|**1.74**	0.7|**1.09**	0.6|**1.39**	2.2|0.95
				$n = 500$			
STATA	1.1|**0.99**	1.9|**0.90**	2.7|0.82	0.9|**1.02**	2.7|0.82	1.6|**0.92**	4.0|0.74
IM-KO	1.1|**0.99**	1.9|**0.90**	2.7|0.82	0.9|**1.02**	2.6|0.82	1.6|**0.92**	4.0|0.74
CGM	1.1|**1.00**	2.0|**0.87**	2.8|0.77	0.9|**1.00**	2.8|0.77	1.6|**0.90**	4.2|0.68
NEW DEFAULT	1.0|**1.01**	0.6|**1.39**	0.7|**1.34**	0.6|**1.36**	0.7|**1.35**	0.6|**1.38**	0.7|**1.28**

Notes: Entries are the null rejection probability in percent, and the average length of confidence intervals relative to average length of confidence intervals based on size corrected STATA (bold if null rejection probability is smaller than 2%) of nominal 1% level tests.

Figure 7: Small Sample Results for CPS Clustered Regressions



References

Bahadur, R., and L. Savage (1956): "The Non-Existence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 25, 1115–1122.

Bakirov, N. K., and G. J. Székely (2005): "Student's T-Test for Gaussian Scale Mixtures," *Zapiski Nauchnyh Seminarov POMI*, 328, 5–19.

Balkema, A. A., and L. de Haan (1974): "Residual life time at great age," *The Annals of Probability*, 2, 792–804.

Bentkus, M., M. Bloznelis, and F. Götze (1996): "A Berry-Esseen Bound for Student's Statistic in the Non-I.I.D. Case," *Journal of Theoretical Probability*, 9, 765–796.

Bentkus, V., and F. Götze (1996): "The Berry-Esseen Bound for Student's Statistic," *The Annals of Probability*, 24, 491–503.

Bloznelis, M., and H. Putter (2003): "Second-order and bootstrap approximation to Student's t-statistic," *Theory of Probability and its Applications*, 47, 300–307.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008): "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics*, 90, 414–427.

Csörgö, S., E. Haeusler, and D. M. Mason (1988): "A probabilistic approach to the asymptotic distribution of sums of independent, identically distributed random variables," *Advances in Applied Mathematics*, 9(3), 259–333.

de Haan, L., and A. Ferreira (2007): *Extreme Value Theory: An Introduction*. Springer Science and Business Media, New York.

Elliott, G., U. K. Müller, and M. W. Watson (2015): "Nearly Optimal Tests When a Nuisance Parameter is Present Under the Null Hypothesis," *Econometrica*, 83, 771–811.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997): *Modelling extremal events for insurance and finance*. Springer, New York.

Falk, M., J. Hüsler, and R. Reiss (2004): *Laws of Small Numbers: Extremes and Rare Events*. Birkhäuser, Basel.

Falk, M., and F. Marohn (1993): "Von Mises conditions revisited," *The Annals of Probability*, 21(3), 1310–1328.

Galambos, J. (1978): *The Asymptotic Theory of Order Statistics*. Wiley, New York.

Gossner, O., and K. H. Schlag (2013): "Finite-Sample Exact Tests for Linear Regressions with Bounded Dependent Variables," *Journal of Econometrics*, 177, 75–94.

Hall, P., and Q. Wang (2004): "Exact Convergence Rate and Leading Term in Central Limit Theorem for Student's T Statistic," *The Annals of Probability*, 32, 1419–1437.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Huber, P. J. (1996): *Robust Statistical Procedures*. SIAM, Philadelphia, 2nd edn.

Imbens, G., and M. Kolesar (2016): "Robust Standard Errors in Small Samples: Some Practical Advice," *The Review of Economics and Statistics*, 98, 701–712.

Johansson, J. (2003): "Estimating the Mean of Heavy-Tailed Distributions," *Extremes*, 6, 91–109.

Kratz, M. (2014): "Normex, a new method for evaluating the distribution of aggregated heavy tailed risks," *Extremes*, 17, 661–691.

LePage, R., M. Woodroofe, and J. Zinn (1981): "Convergence to a stable distribution via order statistics," *The Annals of Probability*, 9(4), 624–632.

Müller, U. K. (2019): "Refining the central limit theorem approximation via extreme value theory," *Statistics & Probability Letters*, 155, 108564.

Müller, U. K., and Y. Wang (2017): "Fixed-k Asymptotic Inference about Tail Properties," *Journal of the American Statistical Association*, 112, 1334–1343.

Müller, U. K., and M. W. Watson (2018): "Long-Run Covariability," *Econometrica*, 86(3), 775–804.

——— (in preparation): "Low-Frequency Analysis of Economic Time Series," in *Handbook of Econometrics*. Elsevier.

Peng, L. (2001): "Estimating the mean of a heavy tailed distribution," *Statistics & Probability Letters*, 52, 255–264.

——— (2004): "Empirical-Likelihood-Based Confidence Interval for the Mean with a Heavy-Tailed Distribution," *The Annals of Statistics*, 32, 1192–1214.

Petrov, V. V. (1975): *Sums of Independent Random Variables*. Springer, New York.

Pickands, III, J. (1975): "Statistical inference using extreme order statistics," *Annals of Statistics*, 3(1), 119–131.

Reiss, R.-D. (1989): *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer Verlag, New York.

Romano, J. P. (2000): "Finite sample nonparametric inference and large sample efficiency," *Annals of Statistics*, 28, 756–778.

Rosenthal, H. P. (1970): "On the subspaces of L^p ($p > 2$) spanned by sequences of independent random variables," *Israel Journal of Mathematics*, 8, 273–303.

Schlag, K. H. (2007): "How to Attain Minimax Risk with Applications to Distribution-Free Nonparametric Estimation and Testing," *European University Institute Working Paper 2007/04*.

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–830.

Zaliapin, I. V., Y. Y. Kagan, and F. P. Schoenberg (2005): "Approximating the Distribution of Pareto Sums," *Pure and Applied Geophysics*, pp. 1187–1228.