A More Robust t-Test*

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

November 2022

Abstract

This paper combines extreme value theory for the smallest and largest k observations for some given $k > 1$ with a normal approximation for the average of the remaining observations to construct a more robust alternative to the usual t-test. The new test is found to control size much more successfully in small samples compared to existing methods in the presence of moderately heavy tails. This holds for the canonical inference for the mean problem based on an i.i.d. sample, but also when comparing two population means and when conducting inference about linear regression coefficients with clustered standard errors.

JEL: C12

*I am thankful for helpful comments and advice from two anonymous referees, the editor, Xiaoxia Shi, as well as 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 numerous direct applications, but also many other standard forms of inference reduce to the application of a t-test applied to a suitably defined sample. 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 and parameters estimated by GMM.

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, and more importantly, a central limit theorem applied to the numerator yields approximate normality. As is well known, a key condition for the central limit theorem is that each term contributes negligibly to the overall variation. But underlying populations with heavy tails are characterized by the presence of large terms. Even if the second moment exists, so that t-statistic based inference is asymptotically justified, large samples might be required before the normal approximation becomes accurate.

In the simple regression context above, there are many reasons why W_i might have long tail(s). The most straightforward one is simply that the disturbances ε_i are heavy tailed. Many economic variables of interest, such as income, health care costs, firm sizes or asset returns have long tails.

Even if the ε_i are not particularly heavy-tailed, W_i might have long left or right tails because the corresponding value of the regressor X_i is large—these are influential observations, which by definition contribute substantially to the sampling variation of $\hat{\beta}$. A particular version of this effect arises in clustered regressions: If the clusters are of fairly heterogeneous size (think of the 50 states of the U.S., say), then the large clusters typically make a non-negligible contribution to the sampling variation of $\hat{\beta}$, again threatening the validity of a normal approximation.

Of course, these effects are more pronounced in small samples. The effective sample size for the normal approximation of the t-statistic is often considerably smaller than the raw number of observations in a study, and not all that large. This may be 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 stems from a small fraction of observations, 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.

This paper develops an alternative to t-statistic based inference 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 exist, so that asymptotically, t-statistic based inference is valid. The aim is to devise an inference method that suffers from less size distortions 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 standard linearization arguments imply that inference about a scalar parameter estimated by GMM, with or without clustering, can be rewritten as a specific inference for the mean problem. The new robust t-test developed for the mean problem can hence, without any further modifications, be used to obtain more reliable inference for most problems of applied interest, such as for a coefficient in a linear regression.

To describe the key idea, consider testing $H_0 : \mathbb{E}[W_i] = 0$ 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) ex-

treme 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 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 of level α under the approximate parametric model. Once the test is applied to the original mean testing problem, it is no longer exactly valid 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. For this result, we consider a population for which extreme value theory provides accurate approximations, and that possesses a finite variance but no third moment. In the asymptotics, we treat k as 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. Thus, precisely under the conditions that lead to a relatively poor performance of analytical critical values, the bootstrap fails to generate an improvement.[1]

[1]On the flip side, the exact test in the parametric approximate model may well not provide

The approach 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 algorithm of Elliott, Müller, and Watson (2015) to numerically determine a powerful test in this parametric problem for selected values of k.

Our preferred default test uses $k = 8$ and is appropriate when the sample consists of at least 50 independent clusters or observations. (We also provide an alternative, even more robust test for $k = 4$ that is applicable to samples with as few as 25 independent clusters or observations.) We analyze this test with extensive Monte Carlo simulations, with data generated from "smooth" analytical distributions, and from draws with replacement from large economic data sets. We find that the new test 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. 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 empirical work, making the new approach particularly attractive in such settings.

The remainder of the paper is organized as follows. The next section describes the new test in a general GMM set-up. Section 3 reports its performance in some small sample simulations. Section 4 discusses the theoretical background for the new theoretical results. Section 5 contains the theoretical development in the inference for the mean problem. Section 6 provides details on how the new test of Section 2 was constructed. Section 7 concludes.

a refinement for underlying populations with more than three moments. It is arguably more important to improve the size control of the usual t-test under conditions where it performs poorly, however, that is, when the tails are not thin.

2 Suggested New Test

Suppose we are interested in testing $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ for a scalar parameter β that is part of a parameter vector $\vartheta = (\beta, \gamma')' \in \mathbb{R}^q$ estimated by Hansen's (1982) Generalized Method of Moments. In particular, assume that ϑ is identified from the $r \times 1$ moment condition $\mathbb{E}[g(\vartheta, z_j)] = 0$ imposed on the data z_j, $j = 1, \ldots, n_z$, and we use an $r \times r$ positive definite weighting function $\hat{\Psi}$ (which is irrelevant in the exactly identified $r = q$ case). 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

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

where $G_i = \sum_{j \in \mathcal{C}_i} g(\vartheta, z_j)$ are i.i.d., $\hat{\Gamma} = -n^{-1} \sum_{j=1}^{n_z} \partial g(\vartheta, z_j)/\partial \vartheta'|_{\vartheta=\hat{\vartheta}} \xrightarrow{p} \Gamma$ 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 a zero population mean applied to the the n observations

$$\hat{W}_i = \hat{\beta} - \beta_0 + \iota_1'(\hat{\Gamma}'\hat{\Psi}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{G}_i, \ i = 1, \ldots, n \tag{2}$$

where $\hat{G}_i = \sum_{j \in \mathcal{C}_i} g(\hat{\vartheta}, z_j)$ and ι_1 is the $q \times 1$ vector $(1, 0, \ldots, 0)'$. For instance, for a linear regression $Y_j = X_j'\vartheta + \varepsilon_j$ with OLS coefficient $\hat{\vartheta} = (\hat{\beta}, \hat{\gamma}')'$ and $\hat{\varepsilon}_j = Y_j - X_j'\hat{\vartheta}$, we obtain

$$\hat{W}_i = \hat{\beta} - \beta_0 + \iota_1' \left(n^{-1} \sum_{i=1}^{n} \sum_{j \in \mathcal{C}_i} X_j X_j' \right)^{-1} \sum_{j \in \mathcal{C}_i} X_j \hat{\varepsilon}_j. \tag{3}$$

In the inference for the mean case $H_0 : E[W_i] = 0$ for an i.i.d. sample W_i, $i = 1, \ldots, n$ with sample average $\hat{\beta} = n^{-1} \sum_{i=1}^{n} W_i$, (3) simply recovers $\hat{W}_i = \hat{\beta} + (W_i - \hat{\beta}) = W_i$.

The practical upshot of this paper is the suggestion to replace the standard GMM test of $H_0 : \beta = \beta_0$ with a more robust t-test applied to the observations $\{\hat{W}_i\}_{i=1}^n$. This new test has a somewhat complicated form due to numerical and other considerations described in Section 6 below, but the basic logic is the one described in the introduction. The test allows for both the left and right tail to be potentially (moderately) heavy.

To define the test, let $\hat{\mathbf{W}}^R = (\hat{W}_1^R, \hat{W}_2^R, \cdots, \hat{W}_k^R)'$ be the k largest order statistics of $\{\hat{W}_i\}_{i=1}^n$, with $\hat{W}_1^R$ the sample maximum, let $\hat{\mathbf{W}}^L = (\hat{W}_1^L, \hat{W}_2^L, \cdots, \hat{W}_k^L)'$ the k smallest

order statistics, with $\hat{W}_1^L$ the sample minimum, and let $s_n^2 = \frac{1}{n-2k}\sum_{i=1}^{n-2k}(\hat{W}_i^m - \overline{\hat{W}}_i^m)^2$ with $\overline{\hat{W}}_i^m = \frac{1}{n-2k}\sum_{i=1}^{n-2k}\hat{W}_i^m$ be the sample variance estimator of the remaining $n-2k$ "middle" observations. Define $\hat{Y}_0 = ((n-2k)s_n^2)^{-1/2}\sum_{i=1}^{n-2k}\hat{W}_i^m$, the two $k\times 1$ vectors $\hat{\mathbf{Y}}^L = -((n-2k)s_n^2)^{-1/2}\hat{\mathbf{W}}^L$ and $\hat{\mathbf{Y}}^R = ((n-2k)s_n^2)^{-1/2}\hat{\mathbf{W}}^R$, and the $2k+1$ vector $\hat{\mathbf{Y}} = (\hat{Y}_0, \hat{\mathbf{Y}}^{L\prime}, \hat{\mathbf{Y}}^{R\prime})'$. The suggested more robust t-test $\varphi^{\text{NEW}}(\hat{\mathbf{Y}})$ of level α rejects $H_0: \beta = \beta_0$, $\varphi^{\text{NEW}}(\hat{\mathbf{Y}}) = 1$, if and only if all of the following four conditions hold:

(i) $|T(\hat{\mathbf{Y}})| > \text{cv}_T(\hat{\mathbf{Y}})$ where

$$T(\mathbf{y}) = \frac{y_0 + \sum_{i=1}^k y_i^R - \sum_{i=1}^k y_i^L}{\sqrt{1 + \sum_{i=1}^k (y_i^R)^2 + \sum_{i=1}^k (y_i^L)^2}}, \tag{4}$$

$\mathbf{y} = (y_0, \mathbf{y}^{L\prime}, \mathbf{y}^{R\prime})' \in \mathbb{R}^{2k+1}$, $\mathbf{y}^L = (y_1^L, \ldots, y_k^L)'$, $\mathbf{y}^R = (y_1^R, \ldots, y_k^R)'$ and $\text{cv}_T(\mathbf{y}) = w_{\text{cv}}(\mathbf{y})\,\text{cv}_\alpha^Z + (1 - w_{\text{cv}}(\mathbf{y}))\,\text{cv}_\alpha^T$ with $w_{\text{cv}}(\mathbf{y}) = 1/(1 + \sum_{i=1}^k (y_i^R)^2 + \sum_{i=1}^k (y_i^L)^2)$ and $(\text{cv}_\alpha^Z, \text{cv}_\alpha^T)$ the $1-\alpha/2$ quantiles of a standard normal and student-t distribution with degrees of freedom equal to $80 + 10\log(\alpha)$, respectively.

(ii) $\varphi^S(\hat{\mathbf{Y}}) = 1$, where, for some integer M^S, $\theta_i^S \in \Theta_0^S \subset \mathbb{R}^3$and $\lambda_i^S > 0$,

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

and for some constants $\rho_1, \rho_r > 0$ and $s_\xi(u) = \frac{u^{-\xi}-1}{\xi}$, $\theta^S = (\kappa, \eta, \xi)$, $\xi_i = -0.49 + \frac{i-1}{9}$,

$$\begin{aligned}
\chi(\mathbf{y}^L) &= \max(0, 5\min(y_1 - \rho_1, \mathbf{1}[y_k^L > 0](y_1^L/y_k^L - 1 - \rho_r)) && (6)\\
f_a^S(\mathbf{y}^R) &= \frac{1}{10}\sum_{i=1}^{10}\int_0^1 \Gamma(k-\xi_i)u^{-\xi_i-1}s_{\xi_i}(u)^{k-1}(y_1^R - y_k^R)^{-k} && (7)\\
&\quad \times \exp\left[-(1+\xi_i^{-1})\sum_{i=1}^k \log\left(1 + \xi_i s_{\xi_i}(u)\frac{y_i^R - y_k^R}{y_1^R - y_k^R}\right)\right] du\\
f^S(\mathbf{y}|\theta^S) &= f_T(\mathbf{y}^R|\theta^S)\phi\left(\frac{y_0 - \sum_{i=1}^k y_i^L + M^*(\mathbf{y}^R, \theta^S)}{\sqrt{1 + \sum_{i=1}^k (y_i^L)^2}}\right) \Big/ \sqrt{1 + \sum_{i=1}^k (y_i^L)^2} && (8)\\
M^*(\mathbf{y}^R, \theta^S) &= \eta(z_k^R)^{-1/\xi}\left(\kappa + \frac{z_k^R}{\xi(1-\xi)} - \frac{1}{\xi}\right) && (9)\\
f_T(\mathbf{y}^R|\theta^S) &= \mathbf{1}[z_k^R > 0]\mathbf{1}[z_1^R > 0]\eta^{-k}\exp\left[-(z_k^R)^{-1/\xi} - (1+\xi^{-1})\sum_{i=1}^k \log(z_i^R)\right] && (10)
\end{aligned}$$

with $z_i^R = 1 + \xi(y_i^R/\eta - \kappa)$ and $\phi(\cdot)$ the standard normal p.d.f.

(iii) $\varphi^S((-\hat{Y}_0, \hat{\mathbf{Y}}^{R\prime}, \hat{\mathbf{Y}}^{L\prime})') = 1$.

(iv) $\varphi^*(\hat{\mathbf{Y}}) = 1$, where for some integer M, $\theta_i \in \Theta_0 \subset \mathbb{R}^6$ and $\lambda_i > 0$,

$$\varphi^*(\mathbf{y}) = \mathbf{1}\left[f_a(\mathbf{y}) > \sum_{i=1}^{M} \lambda_i f^{\text{av}}(\mathbf{y}|\theta_i)\right], \tag{11}$$

$f_a(\mathbf{y}) = f_a^S(\mathbf{y}^L) f_a^S(\mathbf{y}^R)$ and for $\theta = (\theta^L, \theta^R)$ with $\theta^L, \theta^R \in \mathbb{R}^3$,

$$f^{\text{av}}(\mathbf{y}|\theta) = \tfrac{1}{2}(f(\mathbf{y}|(\theta^L, \theta^R), 0) + f(\mathbf{y}|((\theta^R, \theta^L), 0)) \tag{12}$$

$$f(\mathbf{y}|\theta, \mu) = f_T(\mathbf{y}^R|\theta^R) f_T(\mathbf{y}^L|\theta^L) \phi(y_0 - \mu + M^*(\mathbf{y}^R, \theta^R) - M^*(\mathbf{y}^L, \theta^L)). \tag{13}$$

The numbers $\{\lambda_i^S, \theta_i^S\}_{i=1}^{M^S}$, $\{\lambda_i, \theta_i\}_{i=1}^{M}$, ρ_r and ρ_1 are specific to k, the level α of the test, and the nuisance parameter space $\Theta_0 \subset \mathbb{R}^6$ (and the derived space $\Theta_0^S \subset \mathbb{R}^3$; see Section 6.2), but do not depend on the data in any way. As described in more detail in the supplemental appendix, we have numerically determined these numbers for a wide range of significance levels α for $k = 8$ for our preferred default parameter space Θ_0 that is appropriate for all sample sizes larger than $n \geq n_0 = 50$. In addition, we also provide corresponding tables for $k = 4$ computed under a larger nuisance parameter space Θ_0 that is appropriate for samples with $25 \leq n < 50$. The p-value of $H_0 : \beta = \beta_0$ is given by the largest value of α such that the tests reject for all $\alpha' \leq \alpha$. Confidence intervals of level $1 - \alpha$ can be obtained via test inversion,[2] which in light of (2) simply amounts to evaluating φ^{NEW} at $\hat{\mathbf{Y}}$'s computed from location shifted $\hat{W}_i$'s.

While the *determination* of the numbers that fully determine the tests φ^{NEW} was computationally involved, we stress that the *evaluation* of the tests to obtain confidence intervals and p-values in applications does not pose any significant computational burden. A corresponding "post-estimation" STATA command that provides p-values and confidence intervals after running regressions and related commands is available under `http://github.com/ukmueller/robttest`.

3 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,

[2]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.

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 supplemental appendix.

3.1 Inference for the Mean

We initially compare the new test with $k = 8$ ("NEW") 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 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.

Table 1 reports null rejection probabilities, along with the average lengths of the resulting confidence intervals, 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 non-normal populations, the sample mean is not in general 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

Figure 1: Population Densities in Monte Carlo Experiments

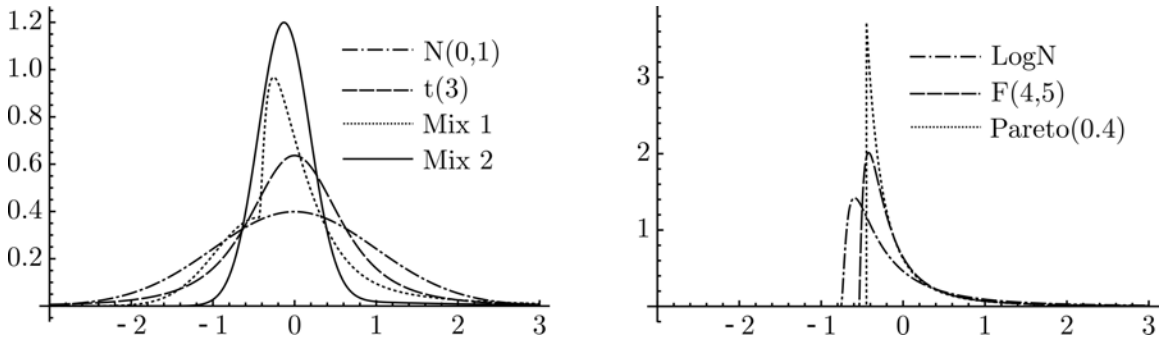


Table 1: Small Sample Results in Inference for the Mean

	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	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	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	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.

existing, its symmetry enables T-STAT and SYM-BOOT to control size at much less cost to average length compared to the new method.[3]

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, we consider a population that is equal to the (discrete) distribution from a large data set. We use the income data of 2016 mortgage applicants as reported by U.S. banks under the Home Mortgage Disclosure Act (HMDA). From this database of more than 16 million applications, we create subpopulations that condition 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).

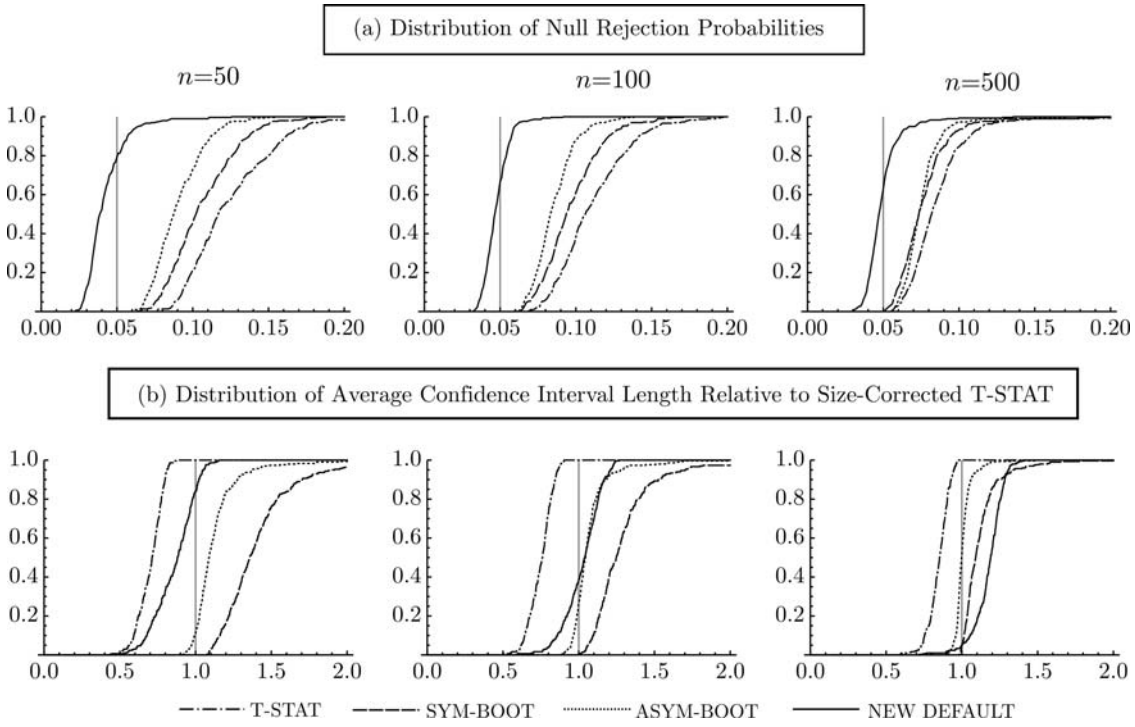
Panel (a) of Figure 2 plots the cumulative distribution function of the null rejection probabilities over the 300 subpopulations for each test considered in Table 1, estimated from 20,000 draws from each subpopulation. Nominally, all mass should be to the left of the 5% line, but the traditional tests don't come close. For instance, for $n = 100$, the usual t-test has a null rejection probability of less than 10% for only approximately 40% of the 300 subpopulations. 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 often 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 there is no robust way to transform a confidence interval for the population

[3]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.

Figure 2: Small Sample Results for HMDA Populations



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.

3.2 Difference between Two Population Means

Our second set of Monte Carlo experiment concerns inference about the difference of two population means, $\beta = \mathbb{E}[W^{\mathrm{I}}] - \mathbb{E}[W^{\mathrm{II}}]$, 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 (3) yields

$$\hat{W}_i = \begin{cases} \bar{W}^{\mathrm{I}} - \bar{W}^{\mathrm{II}} - \beta_0 + 2(W_i^{\mathrm{I}} - \bar{W}^{\mathrm{I}}) & \text{for } i \leq n/2 \\ \bar{W}^{\mathrm{I}} - \bar{W}^{\mathrm{II}} - \beta_0 - 2(W_{i-n/2}^{\mathrm{II}} - \bar{W}^{\mathrm{II}}) & \text{for } i > n/2 \end{cases}$$

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

$$W_i^{\mathrm{I}} = \nu_i + \varepsilon_i^{\mathrm{I}}, \quad W_i^{\mathrm{II}} = \varepsilon_i^{\mathrm{II}} \tag{14}$$

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 2 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 (14), W_i^{I} has much 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.

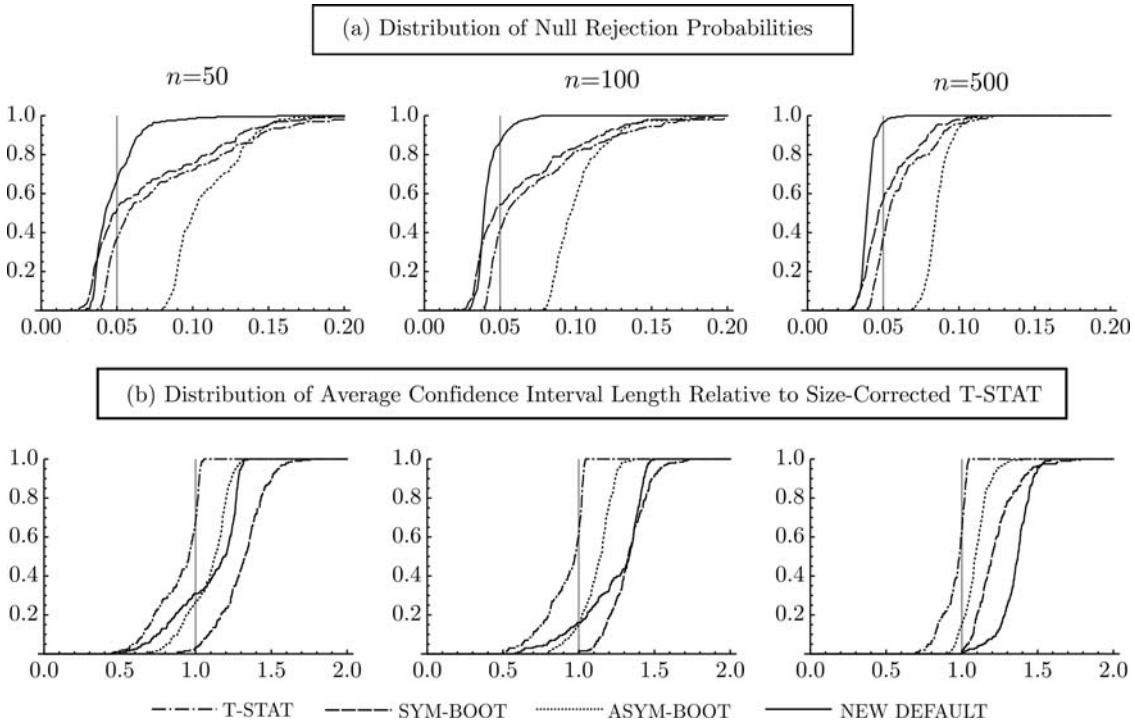
As a second exercise, we generate W_i^j as $n/2$ i.i.d. draws of two randomly selected subpopulations of the HMDA data set considered in the last section. Inference about $\mathbb{E}[W^{\mathrm{I}}] - \mathbb{E}[W^{\mathrm{II}}]$ then corresponds to inference about the average treatment effect if the treatment induces a

Table 2: 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	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	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	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



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 2 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.

3.3 Clustered Linear Regression

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

$$Y_{it} = \beta X_{it} + Z_{it}'\gamma + u_{it},\ t = 1, \ldots, T_i,\ i = 1, \ldots, n \tag{15}$$

with conditionally mean zero u_{it}, so that there are T_i observations in cluster i. By (3) and the Frisch-Waugh Theorem we obtain that

$$\hat{W}_i = \hat{\beta} - \beta_0 + \left(n^{-1} \sum_{j=1}^{n} \sum_{t=1}^{T_j} \hat{X}_{jt} \right)^{-1} \sum_{t=1}^{T_i} \hat{X}_{it} \hat{u}_{it}$$

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 test applied to $\hat{W}_i$ "NEW".

We initially consider data generated from model (15) where

$$u_{it} = \nu_i X_{it} + \varepsilon_{it}, \tag{16}$$

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 3: 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	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	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	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.

Table 3 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.

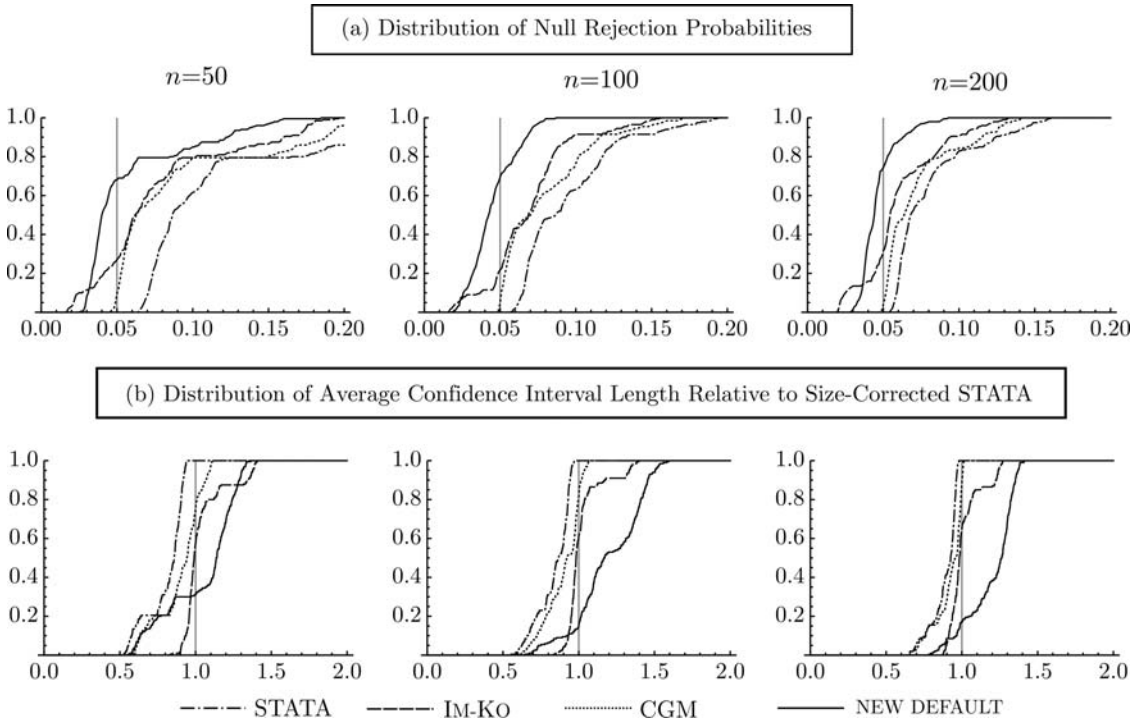
In the final Monte Carlo exercise we again consider a discrete population from a large economic data set. In particular, we consider a sample of all employed workers aged 18-65 from the 2018 merged outgoing rotation group sample of the Current Population Survey (CPS). We let the dependent variable Y_{it} 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 regressors 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).[4] That is, a sample of n clusters is generated by drawing n MSAs at random with replacement. The four different methods of Table 3 are then employed to conduct inference about β based on a sample consisting of all individuals that reside in the n randomly selected MSAs. By construction the clusters are thus i.i.d. and the population regression coefficient is equal to β.

Figure 4 depicts the results over 200 populations generated in this manner, where each population differs by the identity of the regressor of interest X_{it} and controls Z_{it}. For each population, we draw 20,000 samples of $n \in \{50, 100, 200\}$ clusters with replacement. (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 (1)), the variability of the OLS estimator $\hat{\beta}$ is driven by the average of the n i.i.d. random variables $G_i = \sum_{t=1}^{T_i} \hat{X}_{it} u_{it}$. The distribution of G_i may be heavy-tailed because (a) u_{it} has a heavy-tailed component, as in (16) above; (b) the joint distribution of $(\hat{X}_{it}, u_{it})$ is such that $\hat{X}_{it} u_{it}$ is heavy-tailed; (c) $\hat{X}_{it}$ is heavy-tailed; (d) T_i

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

Figure 4: Small Sample Results for CPS Clustered Regressions



is heterogeneous across i, so that clusters with large 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 (d) is thus clearly present, and the suggestion by Imbens and Kolesar (2016) is designed to accommodate effects (c) and (d). But as reported in Table 3, if G_i is heavy-tailed due to effect (a), 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.

4 Background on Theory

We now turn to the theoretical analysis that underlies the new test of Section 2. As mentioned in the introduction, the theory is developed exclusively for the canonical case of inference for the mean.

4.1 Relationship to Literature

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. One must put further restrictions on the underlying population for informative inference to become possible. 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 50 or 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. Our approach rules out such extremely heavy tails by assumption. While this reduces the degree of robustness of the suggested new method, our focus on populations with finite variance mirrors what is (at least, implicitly) assumed in the vast majority of applied work, and it allows for substantially more informative inference in small samples.

Another approach to overcome the Bahadur and Savage (1956) impossibility result is to assume bounded support, with known bounds; see, for instance, Romano (2000), Schlag (2007) and Gossner and Schlag (2013).

4.2 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 $\xi > 0$

$$\lim_{w \to \infty} \frac{1 - F(w)}{(w/\sigma)^{-1/\xi}} = 1 \tag{17}$$

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

$$n^{-\xi} W_1^R \Rightarrow \sigma X_1 \tag{18}$$

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)), (18) implies that extreme value theory also holds jointly for the first k order statistics

$$n^{-\xi}\mathbf{W}^R = n^{-\xi}\begin{pmatrix} W_1^R \\ \vdots \\ W_k^R \end{pmatrix} \Rightarrow \sigma\mathbf{X} = \sigma\begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix}. \tag{19}$$

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 (17) 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*

$$f(w) = (\xi\sigma)^{-1}(\frac{w}{\sigma})^{-1/\xi-1}(1 + h(w)) \tag{20}$$

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, (19) provides accurate approximations in the sense that

$$\sup_B |\mathbb{P}(n^{-\xi}\mathbf{W}^R \in B) - \mathbb{P}(\sigma\mathbf{X} \in B)| = O(n^{-\delta}) \tag{21}$$

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 (20) with $h(w) = 1/\log(1 + w)$, 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.

4.3 Approximations to the t-Statistic

Let $T_n = \sum_{i=1}^n W_i / \sqrt{\sum_{i=1}^n W_i^2 - (\sum_{i=1}^n W_i)^2/n}$ be the t-statistic computed from an i.i.d. sample $W_1, \ldots, W_n$, $W_i \sim W$.[5] 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 speed 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 $\mathbb{E}[W^2] = 1$,

$$\sup_t |\mathbb{P}(T_n < t) - \Phi(t)| \leq C\mathbb{E}[W^2 \mathbf{1}[|W| > n^{1/2}]] + Cn^{-1/2}\mathbb{E}[|W|^3 \mathbf{1}[|W| \leq n^{1/2}]] \tag{22}$$

where $\Phi(t) = \mathbb{P}(Z < t)$, $Z \sim \mathcal{N}(0,1)$. Subsequent research by Hall and Wang (2004) provides a sharp bound on the speed of convergence: If $\mathbb{E}[W^2] < \infty$, their results imply that

$$\frac{\sup_t |\mathbb{P}(T_n < t) - \Phi(t)|}{n\mathbb{P}(|W| > n^{-1/2}) + n^{1/2}|\mathbb{E}[\tilde{W}_n]| + n^{-1/2}\mathbb{E}[|\tilde{W}_n|^3] + n^{-1}\mathbb{E}[|\tilde{W}_n|^4]} \tag{23}$$

with $\tilde{W}_n = W\mathbf{1}[|W| \leq n^{1/2}]$ 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

$$\sup_t |\mathbb{P}(T_n^* < t|\mathbf{W}) - \mathbb{P}(T_n < t)| = o(n^{-1/2}) \text{ a.s.} \tag{24}$$

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-test.

5 New Theoretical Results

To ease exposition, we focus in this section on the case where the left tail of W is light. 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 the supplemental appendix.

[5]For notational simplicity, t-statistics in this paper do not use the degree of freedom correction in the variance estimator; this convention does not affect any of the following results.

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

Theorem 1 *Suppose (17) holds for* $1/3 < \xi < 1/2$, *and* $\int_{-\infty}^{0} |w|^3 dF(w) < \infty$. *Then under* $\mathbb{E}[W] = 0$

(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 (17). Part (a) now simply follows from evaluating the sharp bound in (23). Part (b) follows from applying the Bentkus and Götze (1996) bound (22) to the empirical distribution of $\bar{W}_i = W_i - n^{-1}\sum_{j=1}^{n} W_j$: By (19), $n^{-\xi} \max_i |W_i|$ 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 $n^{-3/2}\sum_{i=1}^{n} |\bar{W}_i|^3 = O_p(n^{-3(1/2-\xi)})$, 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 (17), $\mathbb{P}(W_1^R > \sqrt{n})$ is of order $1 - (1 - n^{-1/(2\xi)})^n \approx n^{1-1/(2\xi)}$. Thus, for $\xi > 1/3$, the largest observation is of order $\sqrt{n}$ with a probability that is an order of magnitude larger than $n^{-1/2}$. 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 (18), 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.

5.2 New Asymptotic Approximation

We first discuss the approximate parametric problem in more detail. Under the Pareto tail assumption (17), 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 variance estimator from the $n-k$ smallest observations.[6] With k fixed, s_n^2 still converges in probability to the unconditional variance of W, $s_n^2 \xrightarrow{p} \text{Var}[W]$. Since the tests we consider are scale invariant, it is without loss of generality to normalize $\text{Var}[W] = 1$. From the convergence to the joint extreme value distribution in (19), $n^{-\xi}\mathbf{W}^R \overset{a}{\sim} \sigma\mathbf{X}$, 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

$$\hat{\mathbf{Y}}_n = \begin{pmatrix} T_n^s \\ \mathbf{W}^R/\sqrt{(n-k)s_n^2} \end{pmatrix} \overset{a}{\sim} \begin{pmatrix} Z + \mu - \eta_n \frac{1}{1-\xi} X_k^{1-1/\xi} \\ \eta_n \mathbf{X} \end{pmatrix} = \mathbf{Y}_n \tag{25}$$

with $\eta_n = \sigma n^{-(1/2-\xi)}$ and $Z \sim \mathcal{N}(0,1)$ independent of $\mathbf{X}$. The last k elements of $\hat{\mathbf{Y}}_n$ are the largest k order statistics divided by the denominator of T_n^s, so that $\hat{\mathbf{Y}}_n$ 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 (25) 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 core idea of this paper is to use the parametric model $\mathbf{Y}_n$ to determine a level α test $\varphi : \mathbb{R}^{k+1} \mapsto \{0,1\}$ of $H_0 : \mu = 0$ that satisfies $\mathbb{E}[\varphi(\mathbf{Y}_n)] \leq \alpha$ 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 (25), $\varphi(\hat{\mathbf{Y}}_n)$, to test $H_0 : \mathbb{E}[W] = 0$ from the observations $W_1, \ldots, W_n$.

A natural one-sided hypothesis test $\varphi_b(\mathbf{Y}_n)$ with $\varphi_b : \mathbb{R}^{k+1} \mapsto \{0,1\}$ has the form $\varphi_b(\mathbf{y}) = \mathbf{1}[y_0 \leq b(\mathbf{y}^R)]$, where $\mathbf{y} = (y_0, \mathbf{y}^{R\prime})'$ and $b : \mathbb{R}^k \mapsto \mathbb{R}$ allows for nonlinear shifts and critical

[6] To avoid notational clutter, the notation in Sections 2, 5 and 6 isn't entirely consistent but adapted to whether only the right tail, or both tails are potentially heavy and whether the approximate parametric model contains a location parameter.

value adjustments as a function of the extreme value observations $\mathbf{y}^R$. Given the form of the Bentkus and Götze (1996) bound (22), this form of tests is convenient to analyze. Two-sided tests can be accommodated by considering linear combinations, as in $\mathbf{1}[|y_0| \geq b(\mathbf{y}^R)] = \mathbf{1}[y_0 \leq -b(\mathbf{y}^R)] + (1 - \mathbf{1}[y_0 \leq b(\mathbf{y}^R)])$ for $y_0 \neq b(\mathbf{y}^R)$. We further allow the form of the test to "switch" between different types depending on the realization of $\mathbf{Y}^R$, leading to the general form

$$\varphi(\mathbf{y}) = \sum_{j=1}^{m_\varphi} \varkappa_j \mathbf{1}[\mathbf{y}^R \in \mathcal{H}_j] \mathbf{1}[y_0 \leq b_j(\mathbf{y}^R)] \tag{26}$$

for some finite m_φ, constants $\varkappa_j \in \mathbb{R}$, Lipschitz continuous $b_j : \mathbb{R}^k \mapsto \mathbb{R}$ and Borel measurable subsets $\mathcal{H}_j$ 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 $\mathcal{I}_j(\mathbf{u}) = \{s > 0 : s\mathbf{u} \in \partial\mathcal{H}_j\}$, that is, $\mathcal{I}_j(\mathbf{u})$ contains the scales s for which $s\mathbf{u}$ falls on the boundary of $\mathcal{H}_j$. For technical reasons, we assume that for some $L > 0$ and Lebesgue-almost all $\mathbf{u}$, $\mathcal{I}_j(\mathbf{u})$ contains at most L elements in the interval $[L^{-1}, L]$, for all $j = 1, \ldots, m_\varphi$.

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 φ is of the form (26). Then under $H_0 : \mu = 0$, for $1/3 < \xi < 1/2$ and any $\epsilon > 0$*

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

Recall from Theorem 1 (a) above that the distribution of the t-statistic converges to the normal distribution at the rate $n^{-(1/(2\xi)-1)}$. A straightforward calculation shows that for $\frac{1+k}{1+3k} < \xi < 1/2$, $r_k(\xi) > 1/(2\xi) - 1$. Thus, for that range of values of ξ, the theorem shows that the difference in the rejection rate of φ in the parametric model $\mathbb{E}[\varphi(\mathbf{Y}_n)]$ and in the original inference for the mean problem $\mathbb{E}[\varphi(\hat{\mathbf{Y}}_n)]$ 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 (22) 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 distribution whose heavy tail has been truncated. Furthermore, under Condition 1, it follows from (21) 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 $\hat{\mathbf{Y}}_n$. While it is easy to show that $s_n^2 \xrightarrow{p} \text{Var}[W] = 1$, 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).

A tedious but straightforward calculation shows that the full sample t-statistic of H_0 : $\mathbb{E}[W] = 0$, T_n, can be written in terms of $\hat{\mathbf{Y}}_n = (T_n^s, \hat{Y}_1^R, \ldots, \hat{Y}_k^R)'$ as

$$T_n = \frac{T_n^s + \sum_{i=1}^k \hat{Y}_i^R}{\sqrt{1 + \sum_{i=1}^k (\hat{Y}_i^R)^2 + R_n}} \tag{27}$$

with $R_n = k(T_n^s)^2/(n(n-k)) - 2T_n^s \sum_{i=1}^k \hat{Y}_i^R/n - (\sum_{i=1}^k \hat{Y}_i^R)^2/n = o_p(n^{-1})$. Under (25), $\hat{\mathbf{Y}}_n \overset{a}{\sim} \mathbf{Y}_n = (Y_0, Y_1^R, \ldots, Y_k^R)'$, so the distribution of T_n is approximated by

$$T(\mathbf{Y}_n) = \frac{Y_0 + \sum_{i=1}^k Y_i^R}{\sqrt{1 + \sum_{i=1}^k (Y_i^R)^2}}. \tag{28}$$

Application of Theorem 2 with $\varphi(\mathbf{y}) = \mathbf{1}[T(\mathbf{y}) > t] = \mathbf{1}[y_0 > -\sum_{i=1}^k y_i^R + t\sqrt{1 + \sum_{i=1}^k (y_i^R)^2}]$ shows that this approximation provides a refinement over the usual standard normal approximation. Müller (2019) shows 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 $\xi > 1/3$. One implication of Theorem 2 is thus a corresponding result for the self-normalized sums (27) and (28).

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.[7]

Proceeding as in the proof of Theorem 2, it can be shown that under Condition 1, $|\mathbb{E}[\varphi(\hat{\mathbf{Y}}_n)] - \mathbb{E}[\varphi(\mathbf{Y}_n)]| \leq Cn^{-\min(r_k(\xi)-\epsilon, 1/2)}$ for $0 < \xi \leq 1/3$. Thus, the approximation (25) also holds for thin-tailed populations. However, since for populations with finite third moment, the rate of convergence of the distribution of the t-statistic to the normal approximation is $n^{-1/2}$, this does not constitute a refinement, irrespective of the choice of k.

[7] A 5% level two-sided test based on $T(\mathbf{Y}_n)$ for $k = 8$ would need to employ a critical value of about 4.2, rather than the usual 1.96, to be valid for $\xi \leq 1/2$ in the $n_0 = 50$ parameter space discussed in Section 6.2 below.

6 Construction of New Test

6.1 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

$$F(w) \approx 1 - (1 + \xi(w/\sigma - \nu))^{-1/\xi}, \quad \xi \in (-\infty, 1/2] \tag{29}$$

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 (17) of Section 4.2 is recovered as a special case with $\xi > 0$, $\nu = 1/\xi$ and σ rescaled by ξ.

Assumption (29) 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), it follows that under an appropriate formalization of (29), there exist real sequences a_n and κ_n such that

$$\frac{\mathbf{W}^R}{a_n} - \kappa_n \Rightarrow \mathbf{X} = (X_1, \ldots, X_k)' \tag{30}$$

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 of (29), then Corollary 1.6.9 of Reiss (1989) implies

$$\left\{ \left(\xi \left(\frac{W_j^R}{a_n} - \kappa_n \right) + 1 \right)^{-1/\xi} \right\}_{j=1}^k \sim \left\{ \left(\frac{n}{\sum_{l=1}^{n+1} E_l} \right) \sum_{l=1}^j E_l \right\}_{j=1}^k \tag{31}$$

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 (30).

Under (29) and (30), from the same logic that led to (25), we obtain the approximate model

$$\hat{\mathbf{Y}}_n = \begin{pmatrix} T_n^s \\ \mathbf{W}^R / \sqrt{(n-k)s_n^2} \end{pmatrix} \overset{a}{\sim} \begin{pmatrix} Z + \mu - \eta_n m^*(\mathbf{X}, \kappa_n, \xi) \\ \eta_n(\mathbf{X} + \kappa_n \mathbf{e}) \end{pmatrix} = \mathbf{Y}_n \tag{32}$$

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

$$m^*(\mathbf{X}, \kappa_n, \xi) = (1 + \xi X_k)^{-1/\xi} \left(\kappa_n + \frac{1 + \xi X_k}{\xi(1-\xi)} - \frac{1}{\xi} \right). \tag{33}$$

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 (32), and this is implicitly applied in standard extreme value theory as reviewed in Section 4.2. However, unless n is very large, this leads to a considerably deterioration of the approximation in (30), and hence (32), 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 (29): 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 (32),

$$\hat{\mathbf{Y}}_n = \begin{pmatrix} \frac{\sum_{i=1}^{n-2k} W_i^m}{\sqrt{(n-2k)s_n^2}} \\ -\frac{\mathbf{W}^L}{\sqrt{(n-2k)s_n^2}} \\ \frac{\mathbf{W}^R}{\sqrt{(n-2k)s_n^2}} \end{pmatrix} \overset{a}{\sim} \begin{pmatrix} Z + \mu - \eta_n^R m^*(\mathbf{X}^R, \kappa_n^R, \xi^R) + \eta_n^L m^*(\mathbf{X}^L, \kappa_n^L, \xi^L) \\ \eta_n^L(\mathbf{X}^L + \kappa_n^L \mathbf{e}) \\ \eta_n^R(\mathbf{X}^R + \kappa_n^R \mathbf{e}) \end{pmatrix} = \mathbf{Y}_n \quad (34)$$

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: For instance, in (31), $\eta_n = n^{-1/2} a_n = \sigma n^{\xi - 1/2}$. 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 (34). 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

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\varphi(\mathbf{Y})] \leq \alpha, \quad (35)$$

where $\mathbf{Y} = (Y_0, \mathbf{Y}^{L\prime}, \mathbf{Y}^{R\prime})'$ and $\mathbf{Y}^J = (Y_1^J, \ldots, Y_k^J)' \in \mathbb{R}^k$ 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 below (30) and (33), it follows that the density of $\mathbf{Y}$ is given by $f(\mathbf{y}|\theta, \mu)$ in (13).

6.2 Nuisance Parameter Space

Allowing for arbitrary values of the location and scale parameters in the testing problem (35) 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 a demeaned 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 (31), 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) $\sum_{i=1}^{n_0} \mathbb{E}[Y_i^J] \geq 0$ and (d) $\sum_{i=k+1}^{n_0-k} \mathbb{E}[(Y_i^J)^2] \leq 2$. Let $\Theta_0^S \subset \mathbb{R}^3$ be the corresponding parameter set on the "single tail" parameter $\theta^S = (\kappa, \eta, \xi) \in \mathbb{R}^3$. Restriction (a) imposes that the tails are such that at least two moments exists. Restriction (b) says that any potential tail shift is inward relative to the non-demeaned Pareto default. Very 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, so Θ_0 is smaller than $\Theta_0^S \times \Theta_0^S$: (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 with a Pareto right tail and a uniform left tail, with overall continuous density. Since the uniform distribution is relatively more spread out compared to the left-tail of a demeaned Pareto distribution, the right tail is shifted outward compared to a demeaned Pareto distribution (but it is still shifted inward relative to a non-demeaned Pareto distribution, so there is no contradiction to requirement (b)). Thus, restriction (c) is satisfied for this population, and as long as k is smaller than $n_0/2 = 25$, the approximate parametric model (34) 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 demeaned 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 demeaned 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.

6.3 Choice of k

As noted in Section 4.2, whenever extreme value theory applies to the sample maximum, then it also applies to the first k order statistics, for any k. This suggests that one may choose k very large, at least for large n, to improve the quality of inference about the tail features. At the same time, as just noted, for any n, there exist populations for which extreme value theory only provides good approximations for a small k (or, indeed, not at all). So there cannot exist a mechanical rule that chooses k as a function of n that is guaranteed to work well. Similarly, rules that are a function of the observed data and that (appropriately) choose $k = k_n$ as diverging whenever the data comes from an exact Pareto distribution also behave poorly for some sequence of populations for which extreme value theory would have provided accurate approximations for finite k (cf. Theorem 5.1 of Müller and Wang (2017)). Ultimately, inference about the mean requires a substantial assumption about the tails, and the stronger the assumptions, the more powerful the potential inference. The assumptions here are embodied in the choice of k and the nuisance parameter space Θ_0.

With that said, we suggest a default value of $k = 8$. On the one hand, as demonstrated in Section 3, $k = 8$ tail observations are already sufficiently informative to allow for tests that are nearly as efficient as the standard t-test when the tails are thin. At the same time, assuming that extreme value theory provides useful approximations for the 8 most extreme observations seems defensible even for relatively small samples, such as for $n = 100$, say. For even smaller samples, say $n < 50$, we suggest using $k = 4$. Finally, relatively small k also facilitate the numerical determination of powerful tests as described below, so these choices are pragmatic also from that perspective.

6.4 Numerical Determination of Powerful Tests

The hypothesis testing problem (35) is a parametric non-standard problem with a one-dimensional parameter of interest $\mu \in \mathbb{R}$ and a six dimensional nuisance parameter $\theta \in \Theta_0 \subset \mathbb{R}^6$.

We seek to construct a test $\varphi : \mathbb{R}^{2k+1} \mapsto [0, 1]$ that comes close to maximizing *weighted average power*. This criterion trades off power against specific alternatives $\mu \neq 0$, $\theta \in \Theta_0$, $\mathbb{E}_{\mu,\theta}[\varphi(\mathbf{Y})]$, by maximizing the weighted average $\text{WAP}(\varphi) = \int \mathbb{E}_{\mu,\theta}[\varphi(\mathbf{Y})]dF_a(\theta, \mu)$ relative to some given weighting function $F_a(\theta, \mu)$. Maximizing weighted average power amounts to maximizing power against the single alternative $f_a(\mathbf{y}) = \int f(\mathbf{y}|\theta, \mu)dF_a(\theta, \mu)$, since $\text{WAP}(\varphi) = \int \varphi(\mathbf{y})f_a(\mathbf{y})d\mathbf{y}$ by a change of the order of integration. We discuss the choice of F_a for the test of Section 2 in the supplemental appendix.

The WAP maximizing test is characterized by the *least favorable distribution* Λ, a probability distribution with support in the null parameter space Θ_0. The problem of identifying the optimal test may be viewed as an adversarial game between the econometrician and nature. Nature chooses $\theta \in \Theta_0$ from the null parameter space, and the econometrician chooses test functions φ. If nature plays a deterministic strategy, that is chooses a specific θ_0, then by the Neyman-Pearson Lemma, the best response by the econometrician is simply the likelihood ratio test that rejects for large values of $f_a(\mathbf{y})/f(\mathbf{y}|\theta_0, 0)$. In general, nature's optimal strategy will be randomized, with θ drawn from the probability distribution Λ over Θ_0. The econometrician's best response is then to reject for large values of $f_a(\mathbf{y})/\int f(\mathbf{y}|\theta, 0)d\Lambda(\theta)$.

The non-standard nature of the testing problem (35) precludes analytic determination of Λ. Instead, we follow Elliott, Müller, and Watson (2015) (abbreviated EMW in the following) and numerically determine an approximate least favorable distribution with finite support.

Their algorithm has a straightforward logic: Start with an arbitrary parameter $\theta_1 \in \Theta_0$ and determine the corresponding optimal test $\mathbf{1}[f_a(\mathbf{y})/f(\mathbf{y}|\theta_1,0) > \mathrm{cv}_1]$ with critical value cv_1 chosen so that it is of level α under $\theta = \theta_1$. If this test controls size for all $\theta \in \Theta_0$, then we have found the overall optimal test. If it fails to control size at, say, $\theta_2 \in \Theta_0$, then determine the two-point distribution Λ_2 with support equal to $\{\theta_1, \theta_2\}$ and critical value cv_2 such that econometrician's best response $\mathbf{1}[f_a(\mathbf{y})/\int f(\mathbf{y}|\theta,0)d\Lambda_2 > \mathrm{cv}_2]$ is of level α under Λ_2 and controls size under $\theta \in \{\theta_1, \theta_2\}$. Now check again if this new test controls size for all $\theta \in \Theta_0$. If it does, we are done. If not, there is a parameter $\theta_3 \in \Theta_0$ for which the test overrejects. Thus, determine the three-point distribution Λ_3 with support $\{\theta_1, \theta_2, \theta_3\}$ and critical value cv_3 such that econometrician's best response $\mathbf{1}[f_a(\mathbf{y})/\int f(\mathbf{y}|\theta,0)d\Lambda_3 > \mathrm{cv}_3]$ is of level α under Λ_3 and controls size under $\theta \in \{\theta_1, \theta_2, \theta_3\}$. Etc.

In practice, it is necessary to introduce some numerical tolerances for this to work well; see EMW and Müller and Watson (2020) for details. Regardless, the approach yields a likelihood ratio-type test of the form (11) where $\theta_i \in \Theta_0$ and λ_i are associated positive weights (which incorporate the critical value, so they don't necessarily sum to unity). The averaging in f^{av} in (11) imposes symmetry in the least favorable distribution Λ in the sense that switching the role of the left and right tail parameter yields the same distribution.

A key ingredient in EMW's numerical approach 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 φ^c under $\theta \in \Theta_0$,

$$\widehat{\mathrm{RP}}(\theta) = N^{-1}\sum_{l=1}^{N} \varphi^c(\mathbf{Y}_{(l)}) \frac{f(\mathbf{Y}_{(l)}|\theta,0)}{\bar{f}(\mathbf{Y}_{(l)})} \tag{36}$$

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). 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, $J \in \{L, R\}$. It is hence 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 by EMW 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}. In the specific problem under study here, it makes sense to switch to simpler tests when one or both tails appear to be sufficiently "thin" so that aggregating the corresponding tail observations with the Gaussian "middle" observation Y_0 still yields an accurate normal approximation. For instance, suppose the left tail seems thin in this sense. Under approximation (32), 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 $\tilde{\varphi}^S : \mathbb{R}^{2k+1} \mapsto \{0,1\}$ 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 $m^*((\mathbf{Y}^R - \kappa^R \mathbf{e})/\eta^R, \kappa^R, \xi^R) = -M^*(\mathbf{Y}^R, \theta^R)$ and variance $\tilde{V}^L$. In analogy to (11), such a test is of the form

$$\tilde{\varphi}^S(\mathbf{y}) = \mathbf{1}\left[f_a^S(\mathbf{y}^R) > \sum_{i=1}^{M^S} \lambda_i^S f^S(\mathbf{y}|\theta_i^S) \right] \tag{37}$$

with $f^S(\mathbf{y}|\theta^S)$ defined in (8).

If both tails seem thin, then one would expect that the distribution of the analogue to the full-sample t-statistic (4) (cf. (28) from Section 5.2) 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. We allow for some adjustment in the usual Gaussian critical value, though, so the resulting test based on $T(\mathbf{Y})$ rejects if $|T(\mathbf{Y})| > \mathrm{cv}_T(\mathbf{Y})$ for a function $\mathrm{cv}_T(\mathbf{Y})$ that is slightly larger than the $1 - \alpha/2$ quantiles of a standard normal and depends on $1 + \sum_{i=1}^k (Y_i^R)^2 + \sum_{i=1}^k (Y_i^L)^2$ and the significance level α. For $\alpha = 5\%$ and $w_{\mathrm{cv}}(\mathbf{Y}) = 0$ (that is, when both tail observations Y^L and Y^R are vanishingly small), $\mathrm{cv}_T(\mathbf{Y}) = 2.009$, which is only slightly larger than the usual critical value of 1.96.

To make this operational, one must take a stand on what constitutes a sufficiently "thin" tail for these normal approximations to be reasonably good. Intuitively, a tail J is thin if either Y_1^J / Y_k^J is close to one, indicating that density of W in tail J drops rapidly; or Y_1^J is small, so that the contribution of the tail relative to Y_0 is nearly negligible, regardless of the shape of the tail density of W. The function $\chi : \mathbb{R}^k \mapsto [0, \infty)$ in (6) is our choice of corresponding "thickness index": $\chi(\mathbf{Y}^J) = 0$ indicates that the tail $J \in \{L, R\}$ is sufficiently thin for the normal approximation to be sensible. The additional term $\exp[\chi(\mathbf{y}^L)]$ in (5) compared to (37) now ensures that the "single right tail" test in condition (ii) of Section 2 is 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 $\varphi^S(\mathbf{y}) = \mathbf{1}[\chi(\mathbf{y}^L) = 0]\tilde{\varphi}^S(\mathbf{y})$. Condition (iii) simply imposes the same rejection rule with the role of the two tails reversed.

Condition (i) implies that φ^{NEW} never rejects if the analogue $T(\mathbf{Y})$ of the full sample t-statistic does not reject; in that sense, φ^{NEW} "robustifies" the usual t-test to obtain better size control. Consequently φ^{NEW} is asymptotically valid whenever the underlying population has two moments, whether or not the tails are approximately Pareto. Condition (i) has the additional appeal that sums of the form (36) then effectively only involve $\mathbf{Y}_{(l)}$ for which $|T(\mathbf{Y}_{(l)})| \geq \text{cv}_T(\mathbf{Y}_{(l)})$, with an associated reduction in computational complexity.

We emphasize that the definition of a "thin tail" in the switching index, the approximate normality of (4) 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 using the density $f(\mathbf{y}|\theta, 0)$ of $\mathbf{Y}$ in (36). The heuristics merely motivate the particular form of φ^{NEW} of Section 2. As discussed, it is not possible to numerically check that $\mathbb{E}_\theta[\varphi(\mathbf{Y})] \leq \alpha$ for *all* $\theta \in \Theta_0$, $\mu = 0$. So technically, we cannot give theoretical guarantees about the size control of φ^{NEW} in the hypothesis testing problem (35). Still, given that the importance sampling approximation of null rejection probabilities (36) is differentiable in θ, one can use fast derivative based hill-climbers (repeatedly, with random starting values), to perform fairly exhaustive checks over a large subset of Θ_0, 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^{NEW} is constrained to also in the remainder of the parameter space makes it plausible that $\mathbb{E}_\theta[\varphi^{\text{NEW}}(\mathbf{Y})] \leq \alpha$ for all $\theta \in \Theta_0$, or at the least, very nearly so.

For $k = 8$ and a given level α, the computations take about one hour on a modern workstation in a Fortran implementation, and about 3 hours for $k = 12$. As discussed in greater detail in the supplemental appendix, we have determined φ^{NEW} for a wide range of significance levels α for $k = 8$ in the default parameter space with $n_0 = 50$, and also for $k = 4$ in the larger nuisance parameter space with $n_0 = 25$. For comparison purposes, we also generated φ^{NEW} with $k \in \{4, 12\}$ for $\alpha \in \{0.01, 0.05\}$ in the default parameter space; see the supplemental appendix for a small sample comparison. After technical modifications that decrease their rejection probability by an arbitrarily small amount, φ^{NEW} satisfies the condition of the two-tailed analogue of Theorem 2.

7 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, 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 the small sample simulations suggest that they help generate more reliable inference also when the underlying population is more loosely characterized by a moderately heavy tail.

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, 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.

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.

——— (2020): "Low-Frequency Analysis of Economic Time Series," *Chapter prepared for the Handbook of Econometrics*.

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.

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.

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.

Supplemental Appendix

"A More Robust t-Test"

Ulrich K. Müller

A 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 $w \geq w_0$,

$$C^{-1}w^{-1/\xi} \leq 1 - F(w) \leq Cw^{-1/\xi} \tag{38}$$

$$f(w) \leq Cw^{-1/\xi-1}. \tag{39}$$

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

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

$V(w) = \text{Var}[W|W \leq w]$ 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 $\mathbb{P}(W_k^R \leq w_0) = \sum_{i=0}^{k-1} \binom{n}{n-i} F(w_0)^{n-i}(1-F(w_0))^i \leq n^k F(w_0)^{n-k}$ 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 (38) and (39), 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 (38), 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 $\beta_0 > 1/\xi$ 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 $m(w) = \mathbb{E}[W\mathbf{1}[W > w]]/F(w)$, (39) 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 (38), and $m(w) \leq Cw^{-1/\xi+1}$ from part (a). Furthermore, using (39)

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

so the result follows.

(c) For $w > w_0$

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

and

$$\begin{aligned} \left| \int_w^\infty sf(s)ds - \int_w^\infty s(\xi\sigma)^{-1}(\frac{s}{\sigma})^{-1/\xi-1} ds \right| &\leq C\int_w^\infty s^{-1/\xi}|h(s)|ds \\ &\leq C\int_w^\infty s^{-(\delta+1)/\xi} ds \\ &\leq Cw^{1-(1+\delta)/\xi} \end{aligned}$$

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)

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

$$\leq \quad Cw^{\beta_0-1/\xi} + Cw^{-\beta/\xi+\beta}$$

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

(e) We have $\mathbb{E}[\tilde{W}^2\mathbf{1}[\tilde{W} < -\sqrt{V(w)n}]] \leq \mathbb{E}[\tilde{W}^2\mathbf{1}[\tilde{W} < 0]] < \infty$. Further, 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

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

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 $\mathbf{1}[m(w) > \frac{1}{2}\sqrt{V(w_0)n}] = 0$ for all large enough n, since $m(w) \leq C$ uniformly in w from part (a). Finally, $\mathbb{P}(W > \frac{1}{2}\sqrt{V(w_0)n}) \leq Cn^{-1/(2\xi)}$ from (38).

(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

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

■

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

$$\mathbb{E}[A_n\mathbf{1}[|s_n^2 - V(W_k^R)| > n^{-r_k(\xi)+1/2-\xi}]|\mathbf{W}^R] \leq Cn^{-r_k(\xi)+\epsilon}(1 + (W_k^R/n^\xi)^{k/\xi-\epsilon}). \tag{40}$$

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

$$\begin{aligned}
\mathbb{E}[|\tilde{W}^2 - V(w)|^\beta] &\leq C\mathbb{E}[|\tilde{W}|^{2\beta}] + CV(w)^\beta \\
&\leq C\mathbb{E}[|\tilde{W}|^{2\beta}] \qquad (41)
\end{aligned}$$

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 (41) 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

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{i=1}^{n-k} Q_i\right| > \tfrac{1}{2}n^{-r_k(\xi)+1/2-\xi}\right) &\leq Cn^{-r_k(\xi)+\epsilon(3/2-2\xi-r_k(\xi))}(1 + (w/n^\xi)^{k/\xi-\epsilon}) \qquad (42) \\
&\leq Cn^{-r_k(\xi)+\epsilon}(1 + (w/n^\xi)^{k/\xi-\epsilon})
\end{aligned}$$

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

$$\mathbb{P}\left(\left|(n-k)^{-1}\sum_{i=1}^{n-k} \tilde{W}_i\right|^2 > \tfrac{1}{2}n^\alpha\right) \leq 2\frac{(n-k)^{-1}V(w)}{n^\alpha} \qquad (43)$$

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

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 (40) follows from (42) and (43).

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 (40) 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

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

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

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

where the second equality follows because the set $\mathcal{I}_j(\mathbf{u})$ is bounded above by L for Lebesgue almost all $\mathbf{u}$, and the last inequality follows because the density of X_1 is bounded above by $Cx_1^{-1/\xi-1}$.

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

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

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

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

where the penultimate inequality follows from (21), and the last inequality applies the result form part (a). The desired inequality now follows from the same reasoning 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

$$\left|\mathbb{E}B_n\mathbf{1}\left[\frac{\sum_{i=1}^{n-k} W_i^s}{\sqrt{(n-k)s_n^2}} \leq b_j\left(\frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}}\right)\right]\right.$$
$$\left. -\mathbb{E}B_n^*\mathbf{1}\left[Z - n^{1/2}m^*(n^{\xi}\sigma X_k) \leq b_j\left(n^{\xi-1/2}\sigma\mathbf{X}\right)\right]\right| \leq Cn^{-r_k(\xi)+\epsilon}$$

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 (22),

$$A_n \sup_s |\mathbb{P}(\varsigma_n \leq s|\mathbf{W}^R) - \Phi(x)| \leq A_n C V(W_k^R)^{-1}\mathbb{E}[\tilde{W}^2\mathbf{1}[\tilde{W}^2 > V(W_k^R)n]$$
$$+ A_n C n^{-1/2}V(W_k^R)^{-3/2}\mathbb{E}[|\tilde{W}|^3\mathbf{1}[\tilde{W}^2 \leq V(W_k^R)n]].$$

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

$$A_n C(n^{-\xi}W_k^R)^2 n^{2\xi-1/(2\xi)} + A_n C(n^{-\xi}W_k^R)^{(1/2-r_k(\xi))/\xi}n^{-r_k(\xi)} \tag{44}$$
$$\leq A_n C n^{-r_k(\xi)}(1 + (n^{-\xi}W_k^R)^2) := L_{1,n}(W_k^R)$$

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

By the Lipschitz continuity of b_j

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

and defining

$$\hat{M} = \frac{(n-k)m(W_k^R)}{\sqrt{(n-k)s_n^2}}, \tilde{M} = \frac{(n-k)m(W_k^R)}{n^{1/2}\sqrt{V(W_k^R)}}$$

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

we have

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

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

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

From Lemma 4,

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

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

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

where the third inequality follows from (44), 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

$$\begin{aligned}&\mathbb{E}\left[B_n\mathbf{1}\left[\frac{\sum_{i=1}^{n-k}W_i^s}{\sqrt{(n-k)s_n^2}} \leq b_j\left(\frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}}\right)\right]\right] - \mathbb{E}\left[B_n\mathbf{1}\left[Z \leq b_j\left(\frac{\mathbf{W}^R}{n^{1/2}\sqrt{V(W_k^R)}}\right) + \tilde{M}\right]\right]\\ \leq\;& \mathbb{E}[L_{1,n}(W_k^R)] + Cn^{-r_k(\xi)+1/2-\xi}\mathbb{E}[L_{2,n}(\mathbf{W}^R)] + \mathbb{E}[L_{3,n}(W_k^R)]. \qquad (45)\end{aligned}$$

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

$$\mathbb{E}[B_n\mathbf{1}[\varsigma_n \leq s + R_n|\mathbf{W}^R]]$$

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

so that (45) 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 (45) 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 (21),

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

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,

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

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

$$n^{-1/2}\mathbb{E}[B_n^* ||n^\xi \mathbf{X}|| \cdot |1 - V(n^\xi \sigma X_k)|] \leq n^{-1/2}\mathbb{E}[||n^\xi \mathbf{X}||^2]^{1/2}\mathbb{E}[B_n^* |1 - V(n^\xi \sigma X_k)|^2]^{1/2}$$
$$\leq Cn^{3(\xi-1/2)}\mathbb{E}[||\mathbf{X}||^2]^{1/2}\mathbb{E}[X_k^{4-2/\xi}]^{1/2} \quad (46)$$

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

$$n^{1/2}\mathbb{E}[B_n^* |m(n^\xi \sigma X_k)| \cdot |1 - V(n^\xi \sigma X_k)|] \leq n^{1/2}\mathbb{E}[B_n^* |m(n^\xi \sigma X_k)|^2]^{1/2}\mathbb{E}[B_n^* |1 - V(n^\xi \sigma X_k)|^2]^{1/2}$$
$$\leq Cn^{3(\xi-1/2)}\mathbb{E}[X_k^{2-2/\xi}]^{1/2}\mathbb{E}[X_k^{4-2/\xi}]^{1/2}. \quad (47)$$

Finally, by Lemma 3 (c),

$$n^{1/2}\mathbb{E}[B_n^* |m(n^\xi \sigma X_k) - m^*(n^\xi \sigma X_k)|] \leq Cn^{-1/2+\xi-\delta}\mathbb{E}[X_k^{1-(1+\delta)/\xi}] + Cn^{\xi-3/2}\mathbb{E}[X_k^{1-2/\xi}]. \quad (48)$$

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 (46)-(48). □

B 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*

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

with $\mathbf{y} = (y_0, \mathbf{y}^{L\prime}, \mathbf{y}^{R\prime})$, $b_j : \mathbb{R}^{2k} \mapsto \mathbb{R}$ Lipschitz continuous functions and $\mathcal{H}_j$ Borel measurable subsets of $\mathbb{R}^{2k}$ with boundary $\partial\mathcal{H}_j$. For $\mathbf{u}^J = (1, u_2^J, \ldots, u_k^J)' \in \mathbb{R}^k$ with $1 \geq u_2^J \geq u_3^J \geq \ldots \geq u_k^J$, let $\mathcal{I}_j(\mathbf{u}^L, \mathbf{u}^R) = \{s^L, s^R > 0 : (s^L \mathbf{u}^L, s^R \mathbf{u}^R) \in \partial\mathcal{H}_j\}$. Assume further that for some $L > 0$, $\mathcal{I}_j(\mathbf{u}^L, \mathbf{u}^R)$ contains at most L elements in the set $[L^{-1}, L]^2$, for Lebesgue almost all $(\mathbf{u}^L, \mathbf{u}^R)$ and $j = 1, \ldots, m_\varphi$.

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

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

where $\hat{\mathbf{Y}}_n$ and $\mathbf{Y}_n$ are the l.h.s. and r.h.s. of (34), respectively.

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

C Implementation Details

C.1 Specification of Weighting Function

We choose $F_a(\theta, \mu)$ to be an improper[8] weighting function with density that is proportional to

$$\mathbf{1}[-1/2 \leq \xi^L \leq 1/2]\mathbf{1}[-1/2 \leq \xi^R \leq 1/2]/(\eta^L \eta^R) \tag{49}$$

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 in practice we approximate f_a^S by an average of those over a grid of values for ξ (cf. equation (7), where the shift by 0.01 avoids evaluation at $\xi_i = 0$).

The lower bound of $-1/2$ on (ξ^L, ξ^R) in (49) plays no important role, since for values of ξ^J that imply an even thinner tail, φ^{NEW} does not overreject due to conditions (i)-(iii) of Section 2.

C.2 Importance Sampling

We use the algorithm in Müller and Watson (2018) to determine an appropriate proposal density $\bar{f}$ for the importance sampling approximation $\widehat{\text{RP}}(\theta)$.

Even though the switching rule 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 $\widehat{\text{RP}}(\theta)$ in (36) 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).

[8]Technically, weighted average power is not properly defined for an improper weighting function. But one can approximate the improper weighting function arbitrarily well by a vague but integrable function, which leads to numerically nearly identical tests.

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

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

where $\mathbf{X}$ is distributed as as in (30), 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 $\theta^S_1 = \theta^L$ and $\theta^S_2 = \theta^R$, respectively, note that their combination into the "both tails" observation $(Y^e_{0,(1)} - Y^e_{0,(2)}, \mathbf{Y}^{S\prime}_{(1)}, \mathbf{Y}^{S\prime}_{(2)})' \in \mathbb{R}^{2k+1}$ has the same distribution as $\mathbf{Y}$ in (34), 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

$$\widetilde{\mathrm{RP}}(\theta) = (KN)^{-1} \sum_{k=1}^{K} \sum_{l=1}^{N} \varphi^c((Y^e_{0,(l)} - Y^e_{0,(l+k)}, \mathbf{Y}^{S\prime}_{(l)}, \mathbf{Y}^{S\prime}_{(l+k)})') \frac{f^e(\mathbf{Y}^e_{(l)}|\theta^L) f^e(\mathbf{Y}^e_{(l+k)}|\theta^R)}{\bar{f}^e(\mathbf{Y}^e_{(l)}) \bar{f}^e(\mathbf{Y}^e_{(l+k)})} \tag{50}$$

that recombines each extended single tail with K different other extended single tails, for a total of KN importance draws. Yet evaluation of (50) 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.

C.3 Computation

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 6.2 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%. Note that Θ^S_s and Θ^S_{ss} depend on (ρ_r, ρ_1) in (6).

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

2. Use the algorithm of EMW to numerically determine φ^S via $\{\lambda^S_i\}_{i=1}^{M_S}$ and $\{\theta^S_i\}_{i=1}^{M_S}$ in (5)

so that

$$\mathbb{E}_\theta[\mathbf{1}[|T(\mathbf{Y})| > \mathrm{cv}_T(\mathbf{Y})]\varphi^S(\mathbf{Y})] \leq \alpha$$

for all $\theta = (\theta^{L\prime}, \theta^{R\prime})' \in \Theta_0$ with $\theta^L \in \Theta_{ss}^S$ and $\theta^R \in \Theta_0^S \backslash \Theta_s^S$.

3. Use the algorithm of EMW to determine φ^* in (11) via $\{\lambda_i\}_{i=1}^M$ and $\{\theta_i\}_{i=1}^M$ so that the overall test φ^{NEW} of Section 2 satisfies $\mathbb{E}_\theta[\varphi^{\mathrm{NEW}}(\mathbf{Y})] \leq \alpha$ under all $\theta = (\theta^{L\prime}, \theta^{R\prime})' \in \Theta_0$ for $\theta^L, \theta^R \in \Theta_0^S \backslash \Theta_s^S$.

4. Spot-check that φ^{NEW} indeed satisfies $\mathbb{E}_\theta[\varphi^{\mathrm{NEW}}(\mathbf{Y})] \leq \alpha$ 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, which ensures that any potential remaining overrejections of a stage can be corrected by a subsequent stage, which increases the numerical stability of the algorithm. Null rejection probabilities are estimated throughout with the importance sampling estimator (50). This estimator has an importance sampling standard error (appropriately adjusted for the dependence) of no more than 0.05%, 0.15% and 0.2% for $\alpha = 1\%$, 5%, 10%, respectively. In Steps 2 and 3, we search for size violations by running a gradient-based hill-climber with up to 200 randomly chosen starting values (with analytically determined derivatives obtained from f^e).

We apply this approach to the default nuisance parameter space with $n_0 = 50$ of Section 6.2 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 $n_0 = 25$ in the constraints of Section 6.2. We ensure that the 95% and 99% level confidence intervals obtained via test inversion 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 by adding the obvious additional constraints to the form of the tests for $\alpha \neq 0.05$. For $k = 8$, a suitable value for ρ_r in (6) is $\rho_r = 0.8$; $\rho_r = 0.5$ for $k = 4$; and $\rho_1 = 0.5$ for both values of k. The replication files provide corresponding tables of $\{\lambda_i^S, \theta_i^S\}_{i=1}^{M_S}$ and $\{\lambda_i, \theta_i\}_{i=1}^M$ for each significance level and value of $k \in \{4, 8\}$. For instance, for $k = 8$ and $\alpha = 5\%$, $M_S = 6$ and $M = 84$.

D Monte Carlo Comparison of New Tests for Various k

Table 4 compares different versions of the new method across the same set of seven populations of Table 1. 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 6.2 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.

E Small Sample Results for $\alpha = 0.01$

Table 4: 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.

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	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	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	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

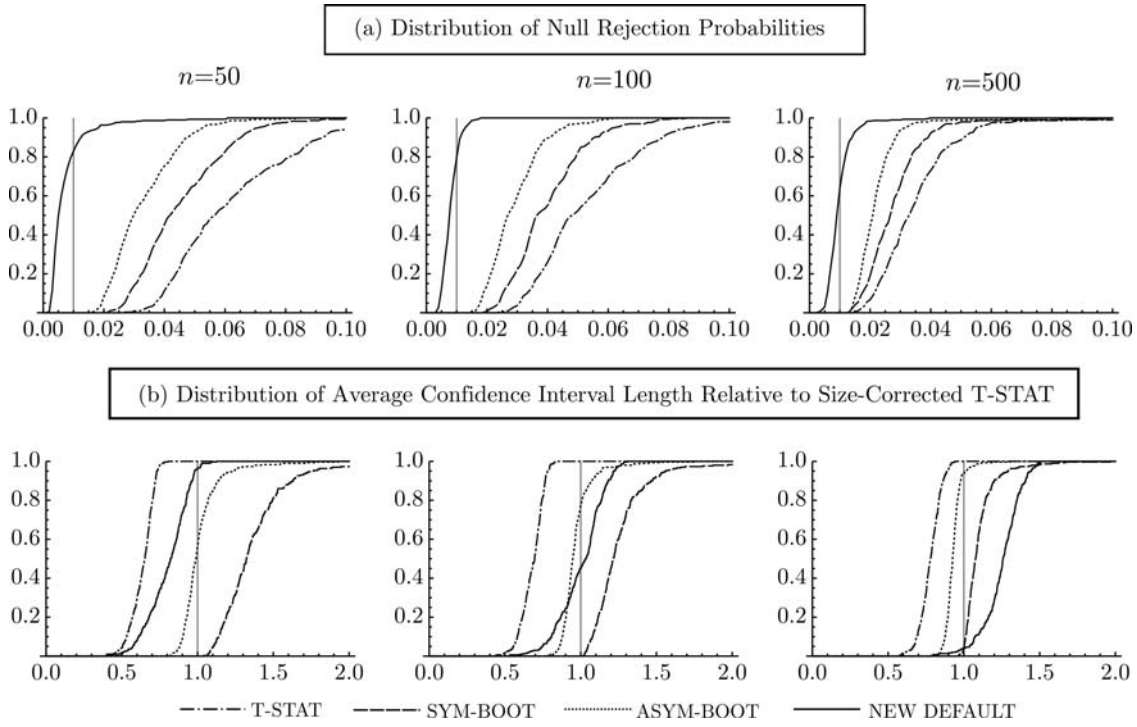


Table 7: Small Sample Results for Difference of Population Means

	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	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	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	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

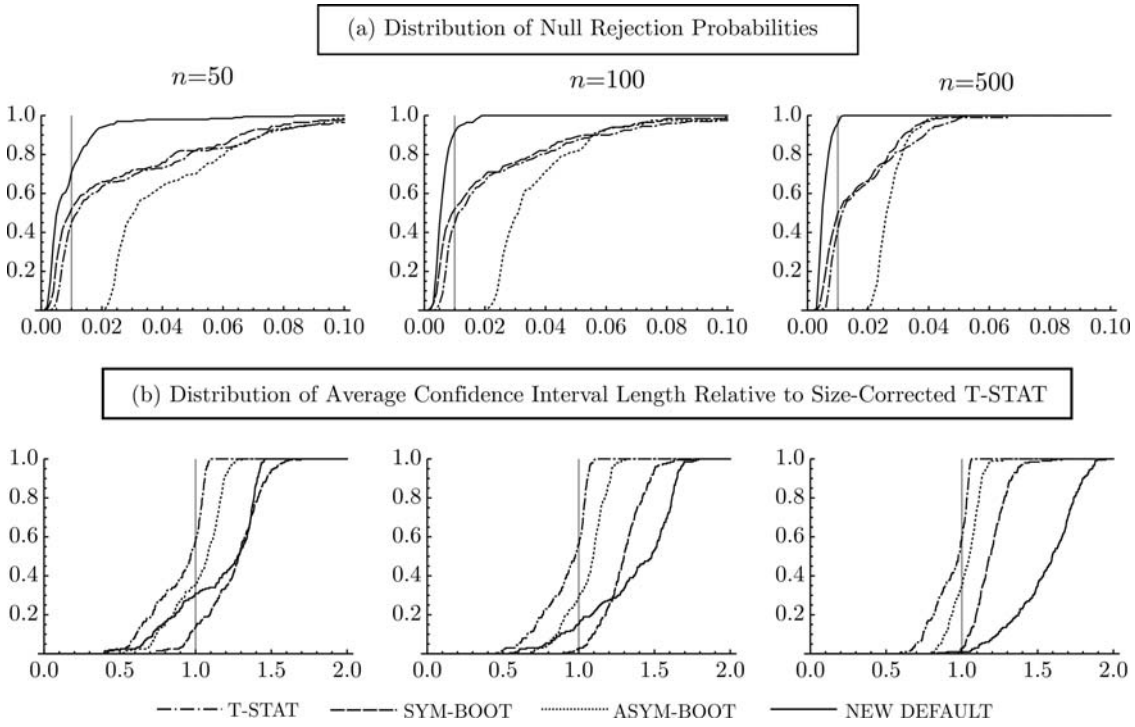
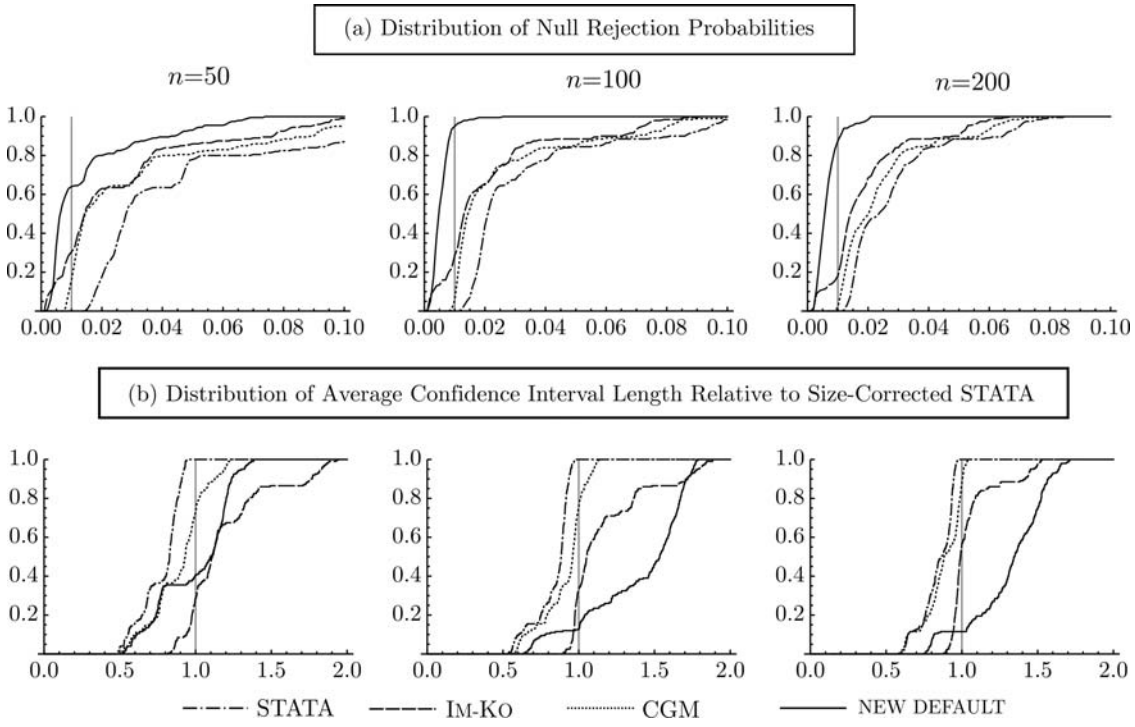


Table 8: Small Sample Results in Clustered Regression Design

	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	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	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	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



Additional References

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