Inference for the Mean

Ulrich K. Müller
Princeton University

November 2019
preliminary

Motivation

- Key building block of econometrics: t-statistic for inference about mean

  $\Rightarrow$ Inference for regression coefficients reduces to inference about mean of $x_i e_i$ for suitably defined e_i, similar in GMM

- Potential challenge: inaccurate approximations by CLT in numerator and LLN in denominator

  $\Rightarrow$ Induced by heavy-tailed population, especially asymmetry, in small samples

  $\Rightarrow$ Effective sample size often not very large due to clustering or nonparametrics

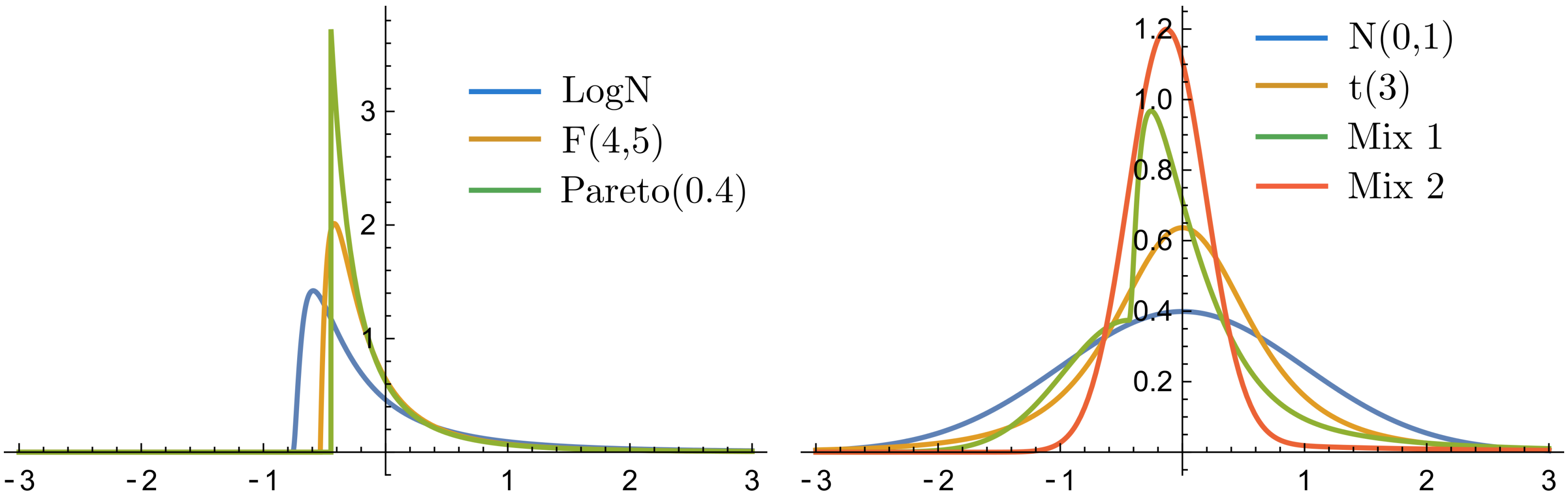
- Standard remedy: Bootstrap

  $\Rightarrow$ Provides refinement when at least three moments exist

Null Rejection Probabilities

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
			$n = 50$				
t-stat	5.1	10.0	12.7	4.5	13.8	7.8	18.7
sym-boot	5.1	7.9	10.1	4.0	10.7	7.4	18.1
asym-boot	5.2	6.8	8.3	7.3	8.4	8.4	17.7
			$n = 100$				
t-stat	4.9	8.3	11.0	4.6	11.8	7.0	15.9
sym-boot	4.9	6.7	9.0	4.2	9.4	6.4	14.5
asym-boot	5.0	6.3	7.7	6.8	7.8	7.1	13.6
			$n = 500$				
t-stat	4.9	6.0	7.8	4.8	8.3	5.7	9.3
sym-boot	5.0	5.4	6.8	4.7	7.2	5.3	7.7
asym-boot	5.0	5.7	6.7	5.9	7.0	6.0	7.3

Population Densities



Basic Idea

- W_i i.i.d. sample of size n with cdf F, $H_0 : E[W] = 0$, long right tail

- Divide and conquer: Largest k order statistics $\mathbf{W}^R = (W_1^R, \ldots, W_k^R)$, and remaining $n-k$ "small" observations W_i^s

- Conditional on $\mathbf{W}^R$, W_i^s i.i.d. with cdf $F(w)/F(W_k^R)$ for $w \leq W_k^R$

  $\Rightarrow$ Conditional mean under H_0:

  $(1 - P(W > w))E[W|W \leq w] + P(W > w)E[W|W > w] = 0$, so

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

Basic Idea, ctd.

- Asymptotic approximations:

  1. $\mathbf{W}^R$ has (joint) extreme value distribution

  2. Conditional on $\mathbf{W}^R$, $\sum_{i=1}^{n-k} W_i^s$ is approximately normal with mean $(n-k)(E[W] + m(W_k^R))$

  3. EVT assumptions imply parametric approximation for $m(\cdot)$

  $\Rightarrow$ Obtain approximate parametric model for $k+1$ observations $(\mathbf{W}^R, \sum_{i=1}^{n-k} W_i^s)$

- Determine 5% level test in approximate parametric model

  $\Rightarrow$ Numerically (very) challenging, but computations only need to be performed once, and application of test to new datasets computationally trivial

Contributions

- New asymptotic approximation for inference about mean

  $\Rightarrow$ Combines EVT and CLT

- Theory: Generates refinement for population with more than two but less than three moments, while bootstrap does not

- Practice: Implementation that only requires few "tail observations"

  Also applicable to inference about scalar parameter in (clustered) linear regression and GMM, but no theory to support potential improvements

- Bahadur and Savage (1956): Inference about mean impossible without further assumptions

  $\Rightarrow$ Assumption here: EVT provides good approximation

Related Literature

- Inference for mean under heavy tails (less than two moments)

  Romano and Wolf (2000), Peng (2001, 2004), Johansson (2003)

- Higher order approximation to distribution of t-statistic

  Bentkus and Götze (1996), Bentkus, Bloznelis and Götze (1996), Bloznelis and Putter (2003), Hall and Wang (2004)

- Fixed-k inference about tail properties

  Müller and Wang (2017)

- Nearly efficient tests and CIs in nonstandard problems

  Elliott, Müller and Watson (2015), Müller and Watson (2016, 2018), Müller and Norets (2016)

Companion Paper

- Use combination of CLT and extreme value distribution for refinement of CLT approximation

$$n^{-1/2}\sum_{i=1}^{n} W_i = n^{-1/2}\left(\sum_{i=1}^{n-k} W_i^s + \sum_{j=1}^{k} W_j^R\right)$$

$\Rightarrow$ Müller (2019) discusses rates of improvement

Outline of Talk

1. Introduction

2. Review of Extreme Value Theory

3. Review of distribution of t-statistic

4. Theory: Rates of errors in coverage probability

5. Implementation

6. Monte Carlo evidence

Review of Extreme Value Theory

- Sufficient for convergence of W_1^R to Fréchet extreme value distribution:

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

$\Rightarrow$ Convergence if upper tail is approximately Pareto

$\Rightarrow$ More-or-less necessary

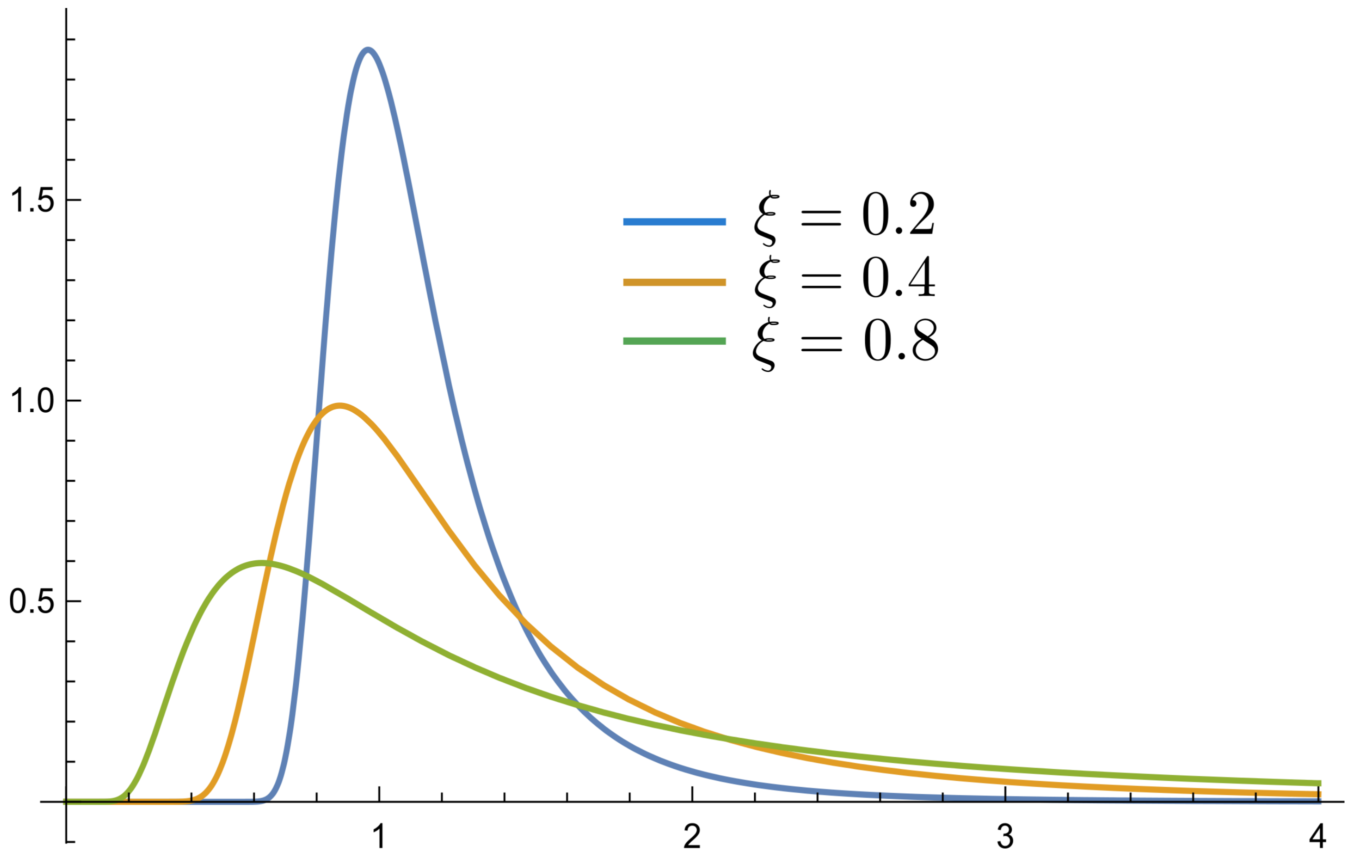
$\Rightarrow$ student-t with df degrees of freedom induces convergence with $\xi = 1/\,\mathrm{df}$, etc.

- Then

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

where for $x > 0$, $P(X_1 \leq x) = G(x) = \exp(-x^{-1/\xi})$

Fréchet Densities



Joint Convergence of Largest k Observations

- If $n^{-\xi} W_1^R \Rightarrow \sigma X_1$, then for any fixed k, also

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

  where joint pdf of $\mathbf{X}$ is given by

$$G(x_k) \prod_{i=1}^{k} g(x_i)/G(x_i)$$

  with $g(x) = dG(x)/dx$

- $\mathbf{X}$ can be generated via $X_1 \sim G$, $X_2|X_1 = x_1 \sim G(x)/G(x_1)$, $X_3|X_2 = x_2, X_1 = x_1 \sim G(x)/G(x_2)$, etc.

Review: Distribution of t-statistic

- If $E[W] = 0$ and $E[W^2] < \infty$, then $T_n \Rightarrow \mathcal{N}(0,1)$

- Let $T_n^*|\mathbf{W}$ be bootstrap draw of T_n conditional on $\mathbf{W} = \{W_i\}_{i=1}^n$

- Theorem (Bloznelis and Putter, 2003). If F is non-lattice and $E[|W|^3]$ exists, then

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

while, for $E[W^3] \neq 0$, $\liminf_{n\to\infty} n^{1/2} \sup_t |P(T_n < t) - \Phi(t)| > 0$, where $\Phi(t) = P(Z < t)$, $Z \sim \mathcal{N}(0,1)$.

Review: Distribution of t-statistic

- Theorem (Bentkus and Götze, 1996): For some $C > 0$, and $E[W^2] = 1$,

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

- Is sharp (Hall and Wang, 2004), holds uniformly in F (Bentkus, Bloznelis, Götze, 1996)

Theory Contributions

1. No bootstrap refinement if extreme value theory holds with $1/3 < \xi < 1/2$

2. Combining CLT for truncated sample with extreme value approximation for k largest observations yields refinement for $1/3 < \xi < 1/2$

Bootstrap under $1/3 < \xi < 1/2$

- Assume that for some $1/3 < \xi < 1/2$, $\lim_{w\to\infty} \frac{P(|W|>w)}{w^{-1/\xi}} > 0$, so that $|W|$ has Pareto tail with index ξ

- Theorem:

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

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

  $\Rightarrow$ Since $3(1/2-\xi) > 1/(2\xi) - 1$, also $\sup_t |P(T_n^* < t|\mathbf{W}) - P(T_n < t)| = O_p(n^{1-1/(2\xi)})$, so no refinement.

- Proof: (a) Follows from sharpness of Bentkus/Götze bound.

  (b) Apply Bentkus/Götze to bootstrap distribution: $n^{-1}\sum_{i=1}^n W_i^2 \mathbf{1}[|W_i| > \sqrt{n}] \xrightarrow{p} 0$ from $\max_i |W_i| = O_p(n^{\xi})$, and $\sum_{i=1}^n |W_i|^3 = O_p(n^{3\xi})$.

New Asymptotic Approximation

- Under approximate Pareto tail $\lim_{w\to\infty} \frac{1-F(w)}{(w/\sigma)^{-1/\xi}} = 1$,

$$m(w) = E[W|W \leq w] \approx -\sigma^{1/\xi}\frac{\xi}{1-\xi}w^{1-1/\xi}$$

- Let $s_n^2 = (n-k)^{-1}\sum_{i=1}^{n-k}(W_i^s - \bar{W}^s)^2$. Then $s_n^2 \xrightarrow{p} \text{Var}[W]$.
  By scale invariance of ultimate test, set $\text{Var}[W] = 1$ wlog.

- Under local alternatives $E[W] = n^{-1/2}\mu$, from $n^{-\xi}\mathbf{W}^R \overset{a}{\sim} \sigma\mathbf{X}$ and CLT

$$\left(\frac{\sum_{i=1}^{n-k} W_i^s}{\sqrt{(n-k)s_n^2}}, \frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}}\right) \overset{a}{\sim} \left(Z + \mu - \eta_n\frac{\xi}{1-\xi}X_k^{1-1/\xi}, \eta_n\mathbf{X}\right)$$

with $\eta_n = \sigma n^{-(1/2-\xi)}$ and $Z \sim \mathcal{N}(0,1)$ independent of $\mathbf{X}$

New Test

- Joint approximation

$$\mathbf{Y}_n := \left(\frac{\sum_{i=1}^{n-k} W_i^s}{\sqrt{(n-k)s_n^2}}, \frac{\mathbf{W}^R}{\sqrt{(n-k)s_n^2}} \right)$$

$$\overset{a}{\sim} \left(Z + \mu - \eta \frac{\xi}{1-\xi} X_k^{1-1/\xi}, \eta \mathbf{X} \right) := \mathbf{Y} = (Y_0, \mathbf{Y}^R)$$

- Construct test $\varphi : \mathbb{R}^{k+1} \mapsto [0,1]$ such that under $H_0 : \mu = 0$, for all $\eta = \eta_n > 0$ and $\xi < 1/2$, $E[\varphi(\mathbf{Y})] \leq \alpha$

  $\Rightarrow$ Many such φ

  $\Rightarrow$ Aim to maximize weighted average power, and apply numerical techniques of Elliott, Müller and Watson (2015)

Asymptotic Refinement

Theorem: Under some technical assumptions, for $k > 1$, $\frac{1+k}{1+3k} < \xi < 1/2$ *and all* $\epsilon > 0$*, under* H_0

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

where $r_k(\xi) = \frac{3(1+k)(1-2\xi)}{2(1+k+2\xi)} > 1/(2\xi) - 1$.

$\Rightarrow$ Recall $\liminf_{n\to\infty} n^{1/(2\xi)-1} \sup_t |P(T_n < t) - \Phi(t)| > 0$, so new approximation is refinement over usual t-test

$\Rightarrow$ Proof: Given Bentkus/Götze, only hard part is to deal with s_n^2 (but that is very involved)

Both Tails Potentially Heavy

- Same approach for two potentially fat tails, where now $\mathbf{W}^e = (\mathbf{W}^L, \mathbf{W}^R)$ and W_i^m are remaining $n - 2k$ middle observations

- Asymptotic approximations then become

$$\begin{pmatrix} \frac{\mathbf{W}^R}{\sqrt{(n-2k)s_n^2}} \\ \frac{-\mathbf{W}^L}{\sqrt{(n-2k)s_n^2}} \end{pmatrix} \overset{a}{\sim} \begin{pmatrix} n^{-(1/2-\xi^R)}\sigma^R \mathbf{X}^R \\ n^{-(1/2-\xi^L)}\sigma^L \mathbf{X}^L \end{pmatrix} = \begin{pmatrix} \mathbf{Y}^R \\ \mathbf{Y}^L \end{pmatrix}$$

and

$$\begin{aligned} \frac{\sum_{i=1}^{n-2k} W_i^m}{\sqrt{(n-2k)s_n^2}} | \mathbf{W}^e \quad &\overset{a}{\sim} \quad Z - \eta^R \frac{\xi^R}{1-\xi^R} (X_k^R)^{1-1/\xi^R} + \eta^L \frac{\xi^L}{1-\xi^L} (X_k^L)^{1-1/\xi^L} \\ &= \quad Y_0 | (\mathbf{Y}^R, \mathbf{Y}^L) \end{aligned}$$

Tail Location Parameters

- EVT holds regardless of (fixed) population shifts

  $\Rightarrow$ Poor small sample approximations

  $\Rightarrow$ Reflected in lower rates for EVT approximation

- Introduce location parameters κ^L and κ^R for tails

  $\Rightarrow$ Parametric problem now indexed by six dimensional nuisance parameter

$$\theta = (\kappa^L, \eta^L, \xi^L, \kappa^R, \eta^R, \xi^R)$$

Implementation

- $\xi^J \leq 1/2$, κ^J such that $E[X_{12}^J] \geq 0$
  Upper bound on η^J from assumption that tail model holds up to X_{25}^J

- Impose that φ never rejects if

$$\frac{|Y_0 + \sum_{j=1}^k Y_j^R - \sum_{j=1}^k Y_j^L|}{\sqrt{1 + \sum_{j=1}^k (Y_j^R)^2 + \sum_{j=1}^k (Y_j^L)^2}} < 2.0$$

- Seek to maximize power against alternative with tail parameters independent, $\xi^J \sim U(-1/2, 1/2)$ and improper density on (κ^J, σ^J) proportional to $1/\sigma^J$

- Determination of φ for $k = 8$ takes about one hour

  $\Rightarrow$ But evaluations of φ are essentially instantaneous

Null Rejection Probabilities

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
t-stat	5.1	10.0	12.7	4.5	13.8	7.8	18.7
sym-boot	5.1	7.9	10.1	4.0	10.7	7.4	18.1
asym-boot	5.2	6.8	8.3	7.3	8.4	8.4	17.7
new $k = 8$	3.8	3.4	4.6	3.1	5.8	3.2	11.8
				$n = 100$			
t-stat	4.9	8.3	11.0	4.6	11.8	7.0	15.9
sym-boot	4.9	6.7	9.0	4.2	9.4	6.4	14.5
asym-boot	5.0	6.3	7.7	6.8	7.8	7.1	13.6
new $k = 8$	4.7	2.9	3.6	3.8	3.7	3.3	8.6
				$n = 500$			
t-stat	4.9	6.0	7.8	4.8	8.3	5.7	9.3
sym-boot	5.0	5.4	6.8	4.7	7.2	5.3	7.7
asym-boot	5.0	5.7	6.7	5.9	7.0	6.0	7.3
new $k = 8$	4.6	4.2	4.4	4.2	4.5	4.2	2.8

Normalized Average Lengths

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
t-stat	**0.99**	0.75	0.67	**1.01**	0.63	0.87	0.59
sym-boot	**0.99**	1.06	1.33	**1.12**	1.45	1.09	1.42
asym-boot	**1.00**	0.96	1.07	1.07	1.15	1.00	1.07
new $k = 8$	**1.12**	**0.92**	**0.75**	**1.44**	**0.69**	**1.08**	0.61
				$n = 100$			
t-stat	**1.00**	0.84	0.73	**1.01**	0.71	0.91	0.60
sym-boot	**1.01**	1.05	1.19	**1.08**	1.25	1.06	1.18
asym-boot	**1.01**	0.98	1.01	1.05	1.04	1.00	0.95
new $k = 8$	**1.01**	**1.24**	**1.03**	**1.34**	**0.99**	**1.29**	0.73
				$n = 500$			
t-stat	**1.00**	0.95	0.87	**1.01**	0.85	**0.97**	0.79
sym-boot	**1.00**	**1.01**	1.13	**1.03**	1.18	**1.01**	1.06
asym-boot	**1.00**	**1.00**	1.03	**1.02**	1.04	**1.00**	0.97
new $k = 8$	**1.02**	**1.18**	**1.20**	**1.13**	**1.19**	**1.18**	**1.27**

Note: Normalized by average length of size corrected t-stat; bold indicates size$<$6%

Null Rejection Probabilities

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
			$n = 50$				
t-stat	4.9	10.3	12.6	4.7	13.5	7.7	19.2
new $k = 4$	3.5	3.0	3.9	3.4	4.2	2.9	10.6
new $k = 8$	3.8	3.4	4.6	3.1	5.8	3.2	11.8
new $k = 12$	3.3	6.2	8.3	3.0	9.9	3.1	11.5
			$n = 100$				
t-stat	5.2	8.2	10.8	4.6	11.5	7.1	15.4
new $k = 4$	4.8	2.8	3.5	3.5	3.6	2.9	4.8
new $k = 8$	4.7	2.9	3.6	3.8	3.7	3.3	8.6
new $k = 12$	4.2	2.7	3.3	3.7	3.5	3.0	7.8
			$n = 500$				
t-stat	5.2	5.8	7.6	5.1	7.9	5.8	9.2
new $k = 4$	4.9	3.7	3.7	4.3	3.7	3.8	2.7
new $k = 8$	4.6	4.2	4.4	4.2	4.5	4.2	2.8
new $k = 12$	4.9	3.4	3.6	3.6	3.7	3.5	2.7

Normalized Average Lengths

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
			$n = 50$				
t-stat	**1.00**	0.73	0.65	**1.01**	0.63	0.87	0.59
new $k = 4$	**1.22**	**0.98**	**0.80**	**1.48**	**0.76**	**1.17**	0.63
new $k = 8$	**1.12**	**0.92**	**0.75**	**1.44**	**0.69**	**1.08**	0.61
new $k = 12$	**1.17**	0.79	0.67	**1.34**	0.64	**1.01**	0.60
			$n = 100$				
t-stat	**0.98**	0.83	0.73	**1.01**	0.72	0.90	0.61
new $k = 4$	**1.03**	**1.24**	**1.02**	**1.54**	**0.98**	**1.32**	**0.73**
new $k = 8$	**1.01**	**1.24**	**1.03**	**1.34**	**0.99**	**1.29**	0.73
new $k = 12$	**1.05**	**1.24**	**1.05**	**1.29**	**1.01**	**1.29**	0.74
			$n = 500$				
t-stat	**0.99**	**0.96**	0.88	**1.00**	0.86	**0.96**	0.80
new $k = 4$	**1.01**	**1.34**	**1.30**	**1.35**	**1.29**	**1.35**	**1.21**
new $k = 8$	**1.02**	**1.18**	**1.20**	**1.13**	**1.19**	**1.18**	**1.27**
new $k = 12$	**1.01**	**1.20**	**1.17**	**1.19**	**1.17**	**1.21**	**1.28**

Note: Normalized by average length of size corrected t-stat; bold indicates size$<$6%

Empirical Monte Carlo

- Consider applicant's income in 2016 Home Mortgage Disclosure Act (HMDA) universe of ≈16m mortgage applications

- Create subpopulations by conditioning on gender, purpose of loan, US state, race

  ⇒ 330 subpopulations with more than 4000 individuals

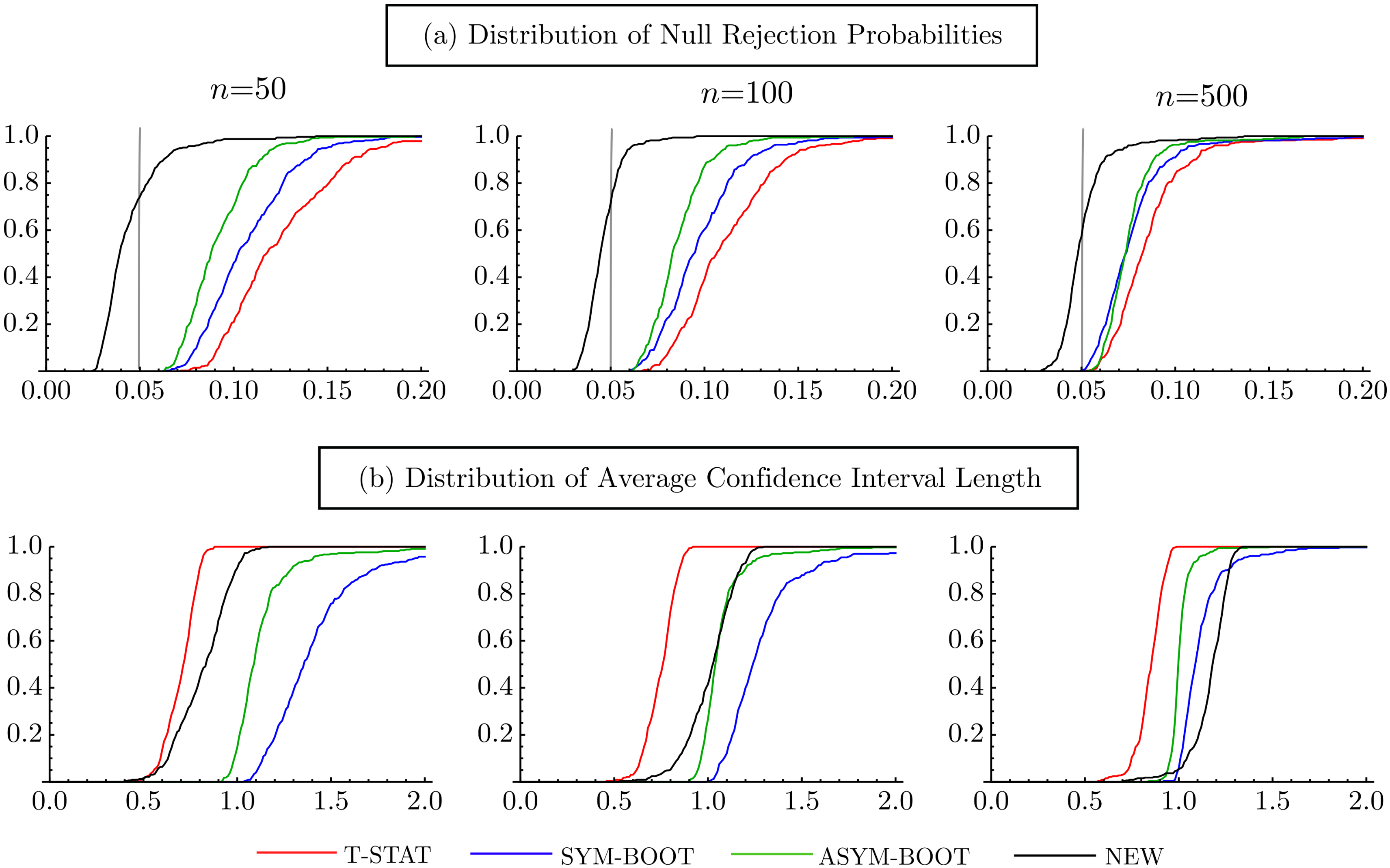
- For each subpopulation:

  – Compute mean of applicant's income

  – Repeat 20,000 times: Draw n data points at random, compute CIs and check whether it contains subpopulation mean

  ⇒ Obtain 330 null rejection probabilities, and average lengths

HMDA Results



Regression with Clustered Standard Errors

- Consider inference about β_0 in linear regression with clustered errors

$$\begin{aligned} Y_{it} &= \alpha + \beta R_{it} + X_{it}'\gamma + u_{it} \\ u_{it} &= \nu_i R_{it} + \varepsilon_{it} \end{aligned}$$

  for $i = 1, \ldots, n$, $t = 1, \ldots., T$ with R_{it}, $X_{it,j}$, $\varepsilon_{it} \sim iid\mathcal{N}(0,1)$, and ν_i i.i.d. mean-zero

- Let $\hat{R}_{it}$ be the residuals of a regression of R_{it} on X_{it} and a constant. By Frisch-Waugh

$$\hat{\beta} - \beta = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} \hat{R}_{it}^2 \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{R}_{it} u_{it}$$

  and STATA computes clustered standard error via

$$\hat{\sigma}_{\hat{\beta}}^2 = \frac{nT}{nT-k} \frac{n}{n-1} \left(\sum_{i=1}^{n} \sum_{t=1}^{T} \hat{R}_{it}^2 \right)^{-2} \sum_{i=1}^{n} \left(\sum_{t=1}^{T} \hat{R}_{it} \hat{u}_{it} \right)^2$$

Regression with Clustered Standard Errors

- Nearly equivalent to inference about mean of

$$W_i = \hat{\beta} + c^{-1} \sum_{t=1}^{T} \hat{u}_{it} \hat{R}_{it} \qquad c = n^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{R}_{it}^2$$

$\Rightarrow$ Apply new procedure to $\{W_i\}_{i=1}^n$

- Compare to alternative approaches
  - STATA
  - Cameron, Gelbach and Miller (2008): Wild Bootstrap with null hypothesis imposed
  - Imbens and Kolesár (2016): Degrees of freedom adjustment as function of design matrix and estimated intra-cluster random effect type correlation

Null Rejection Probabilities, $T = 10$

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
			$n = 50$				
STATA	5.5	10.5	13.4	4.6	12.5	7.8	17.8
Im-Ko	5.3	10.3	13.2	4.4	12.3	7.6	17.5
CGM	5.2	10.6	13.4	4.8	12.6	7.8	17.7
new $k = 8$	3.3	4.5	5.8	2.8	4.6	3.1	8.8
			$n = 100$				
STATA	4.9	9.0	11.2	4.9	11.0	7.1	15.9
Im-Ko	4.8	8.9	11.1	4.8	10.9	7.0	15.8
CGM	4.7	9.0	11.3	5.1	11.2	7.1	16.2
new $k = 8$	4.1	3.0	3.7	4.0	3.8	3.9	8.3
			$n = 500$				
STATA	5.6	6.0	8.1	4.6	8.1	6.0	10.2
Im-Ko	5.6	6.0	8.1	4.6	8.1	6.0	10.2
CGM	5.5	6.2	8.4	4.7	8.2	6.1	10.5
new $k = 8$	5.2	3.9	4.3	4.1	4.5	4.1	3.5

Normalized Average Lengths, $T = 10$

	N(0,1)	LogN	F(4,5)	t(3)	P(0.4)	Mix 1	Mix 2
				$n = 50$			
STATA	**0.98**	0.74	0.65	**1.02**	0.69	0.88	0.63
Im-Ko	**0.99**	0.74	0.66	**1.02**	0.70	0.88	0.64
CGM	**0.99**	0.71	0.61	**1.00**	0.65	0.85	0.57
new $k = 8$	**1.36**	**0.84**	**0.71**	**1.43**	**0.76**	**1.11**	0.66
				$n = 100$			
STATA	**1.00**	0.82	0.72	**1.01**	0.73	0.91	0.63
Im-Ko	**1.01**	0.82	0.73	**1.01**	0.74	0.91	0.64
CGM	**1.01**	0.80	0.68	**0.99**	0.69	0.89	0.57
new $k = 8$	**1.11**	**1.19**	**1.00**	**1.41**	**1.00**	**1.29**	0.76
				$n = 500$			
STATA	**0.98**	0.95	0.85	**1.02**	0.86	**0.96**	0.76
Im-Ko	**0.98**	**0.95**	0.86	**1.02**	0.86	**0.96**	0.76
CGM	**0.98**	0.94	0.82	**1.01**	0.82	0.95	0.72
new $k = 8$	**0.99**	**1.24**	**1.20**	**1.17**	**1.20**	**1.21**	**1.20**

Note: Normalized by average length of size corrected t-stat; bold indicates size<6%

Empirical Monte Carlo

- Treat 2018 CPS data of outgoing rotation as population, appr. $150,000$ observations

- Consider regressions of log wage on random 5 element subset of gender, race, age, education, union status, marriage status, education, etc.

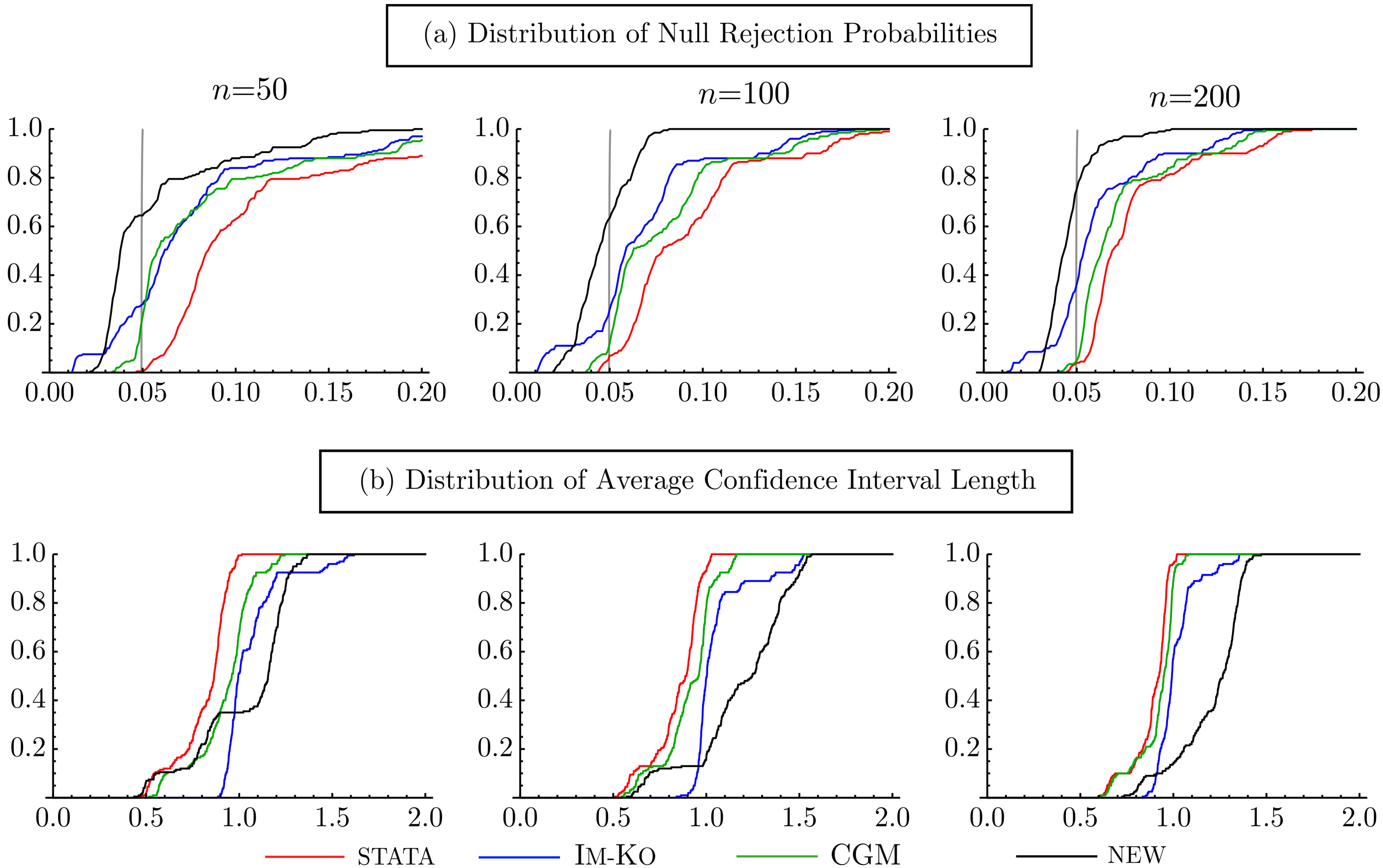
  $\Rightarrow$ obtain 200 population regression coefficients

- Interest in clustered inference by Metropolitan Statistical Areas (MSA), total of 308

- For each population regression, repeat 20,000 times:

  draw n clusters at random, compute clustered CI, and check whether first population coefficient is included

Log-Wage CPS Clustered by MSA



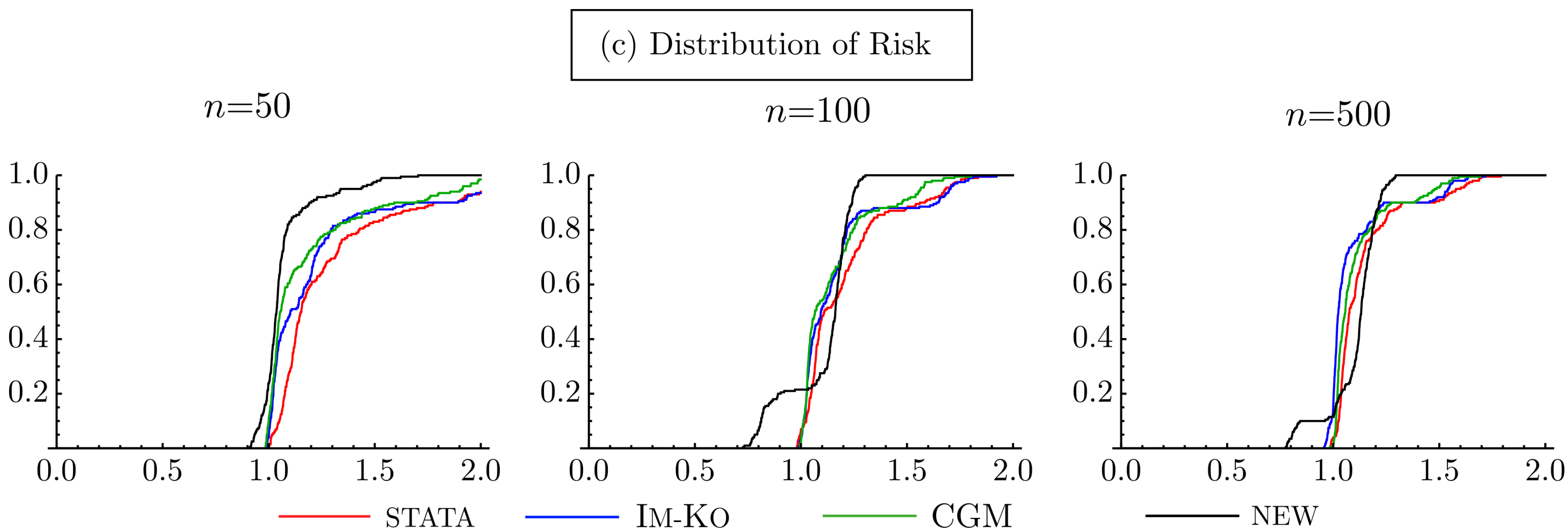
Log-Wage CPS Clustered by MSA

- Loss function of the form

$$\ell(\text{CI}) = \text{length}(\text{CI}) + c\mathbf{1}[\theta_0 \notin \text{CI}]$$

and for each population, determine c such that risk minimizing STATA cv yields 5% level test

For each population, normalize risk of optimal STATA to unity



Heavy Tails from Clustering

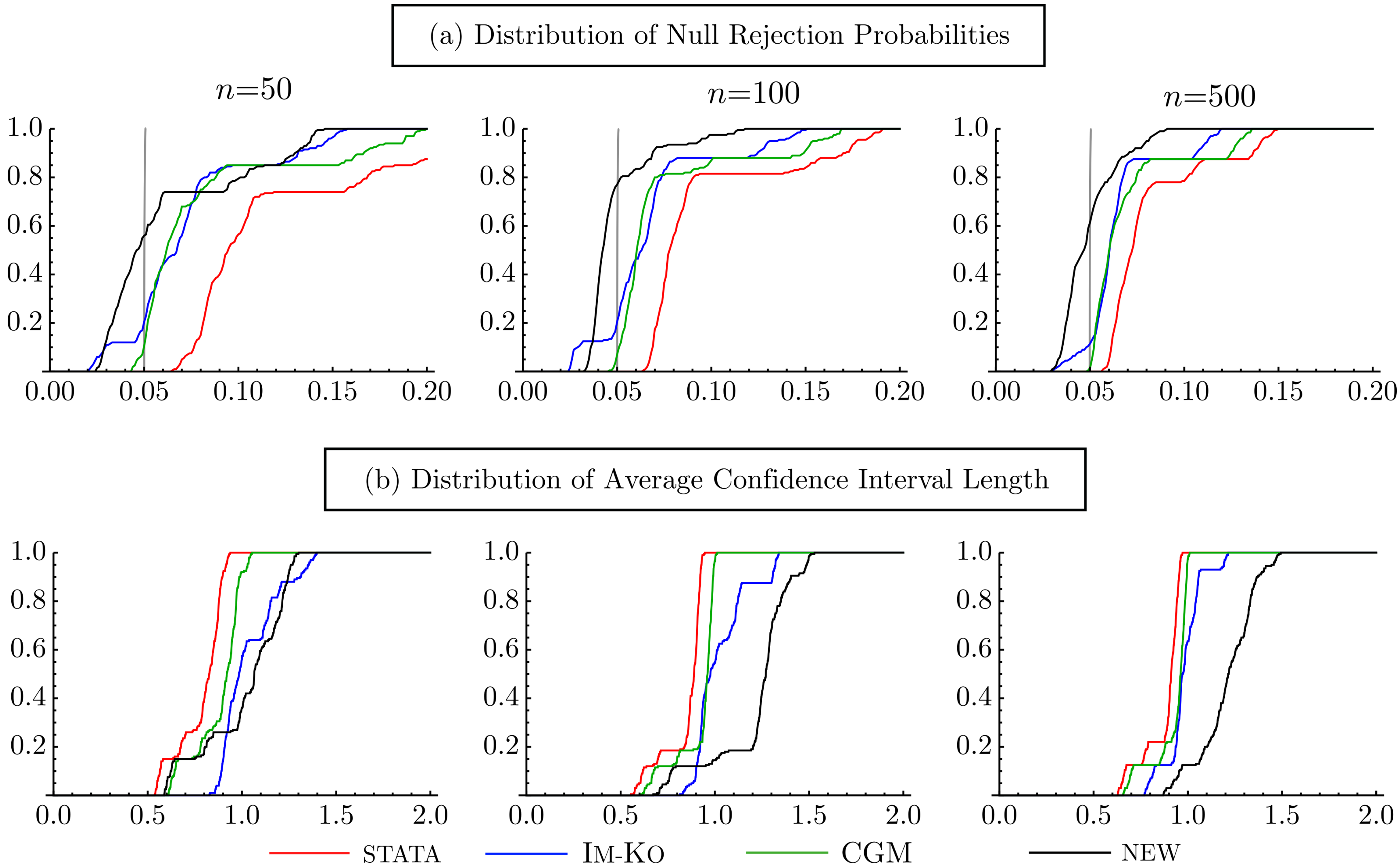
- Recall that under clustering, variability in $\hat{\beta}$ conditional on regressors is driven by variability of

$$\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\hat{R}_{it}u_{it}\right)=\sum_{i=1}^{n}W_i^0$$

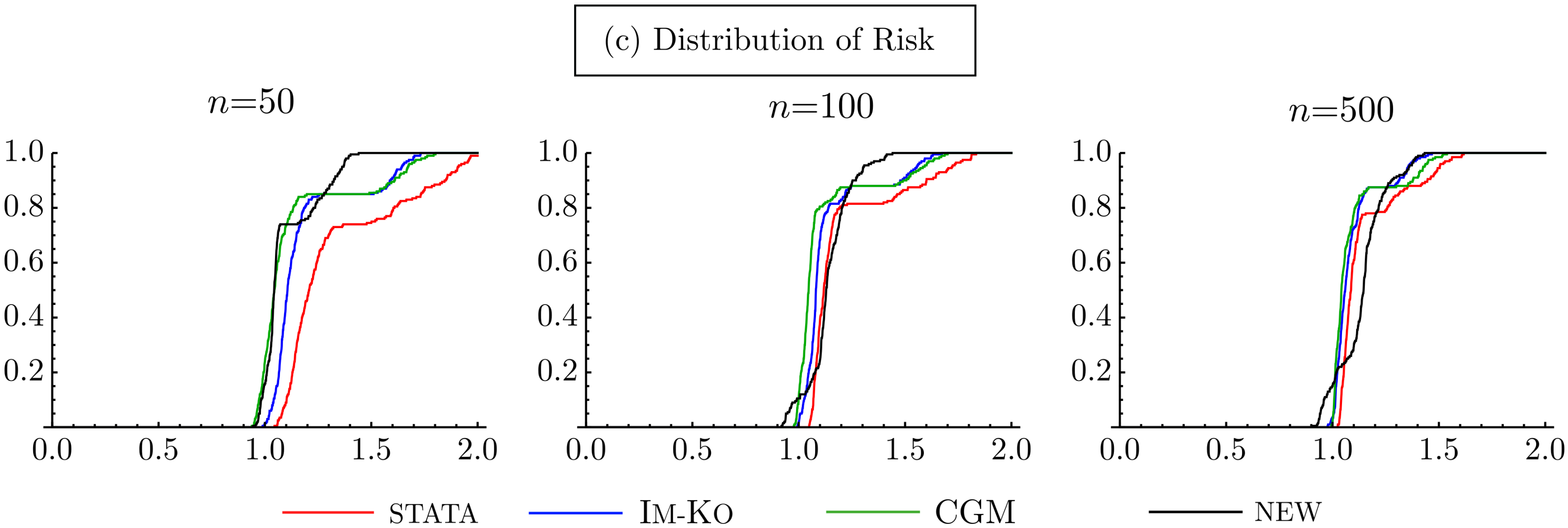
- W_i^0 can be heavy-tailed because
  - u_{it} has cluster-specific heavy-tailed component, $u_{it}=\varepsilon_{it}+\nu_i$
  - $\hat{R}_{it}$ is heterogeneous across clusters
  - T_i is heterogeneous across clusters

  or combinations thereof

Union Status Regression Clustered by MSA



Union Status Clustered by MSA



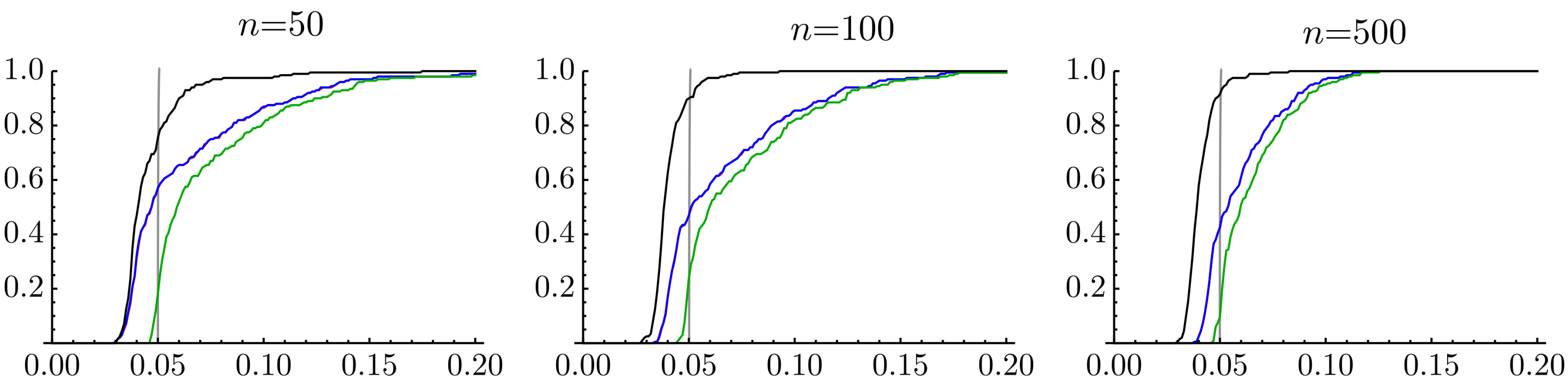
Two Sample t-statistic

- Inference about difference in population mean between two randomly chosen 330 subpopulations of the HMDA data set

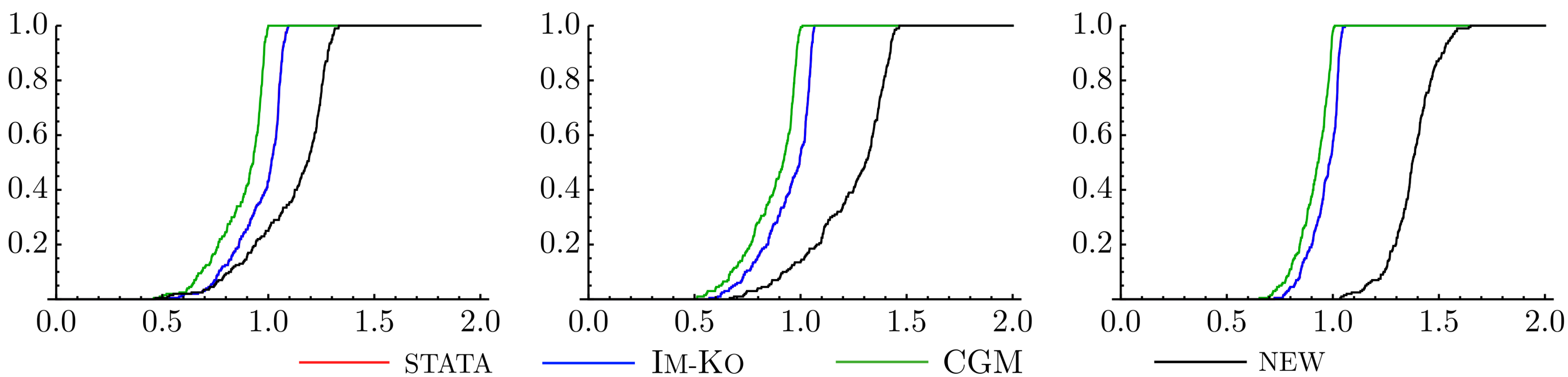
- Draw $n/2$ i.i.d. observations from each subpopulation and apply standard regression inference (= treat each observation as its own cluster)

HMDA Two Sample

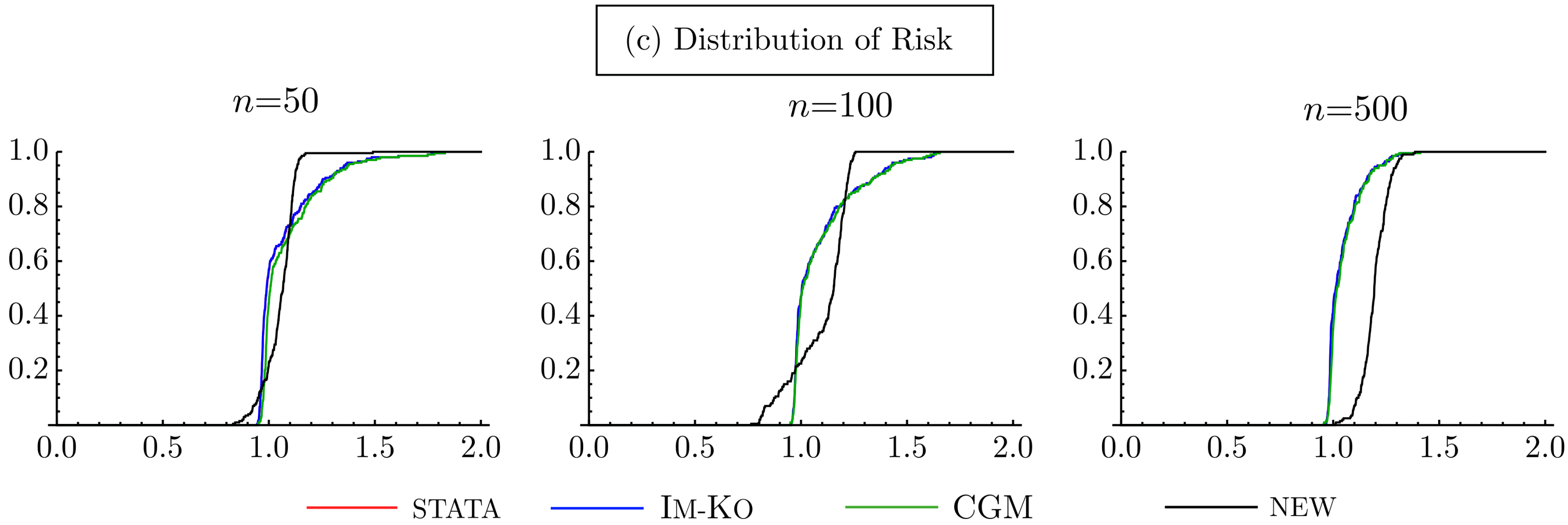
(a) Distribution of Null Rejection Probabilities



(b) Distribution of Average Confidence Interval Length



HMDA Two Sample



Conclusions

- New approach to inference for mean in presence of potentially fat tails that combines EVT and CLT

- Theory: Provides refinement under Pareto-like fat tails (more than two but less than three moments), while bootstrap does not

- Practice: Implementation that yields noticeably better size control in fat-tailed populations, also in clustered regression inference