Discussion of Abrams, Bertrand and Mullainathan
"Do Judges Vary in their Treatment of Race?"

Conference on Empirical Legal Studies
November 9, 2007

Justin Wolfers
Wharton School, University of Pennsylvania
CEPR, CESifo, IZA and NBER

Inter-Judge Disparity

- Do different judges yield different decisions?
  - Exploit random assignment of judges to cases

"Since the rule is that there is no selection of the cases which the judge is to sentence but that the sentencing of a particular prisoner by a particular judge is a matter of chance (the judges rotate), it is obvious that, by chance, each judge should get an equal number of cases whose sentences would normally be long or short.

In other words, given a sufficient number of cases, one could expect that two judges would give sentences whose average severity would be about equal (providing that the judges were influenced only by the circumstances of the crime and those of the prisoner).

Conversely, given a sufficiently large number of cases, if one finds that the average severity of the sentences of two judges is appreciably different, one is justified in saying that the factors which determine this difference in the sentencing tendencies are to be found outside of the circumstances of the crime and those of the prisoner, and hence probably in the judge since he is the other factor which is always present.

Inter-Judge Disparity

- Do different judges yield different decisions?
  - Exploit random assignment of judges to cases

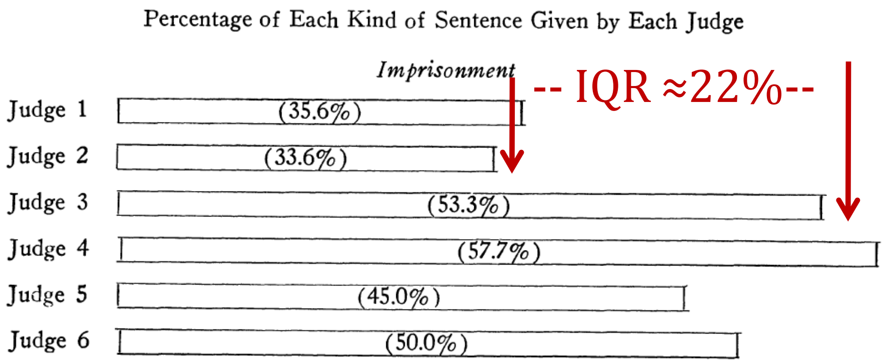
Gaudet et al (1933)

"Individual Differences in the Sentencing Tendencies of Judges"

-Criminal cases from a NJ county

- ≈1000 cases per judge

-Finds large variation in incarceration rates



-- IQR ≈22%--

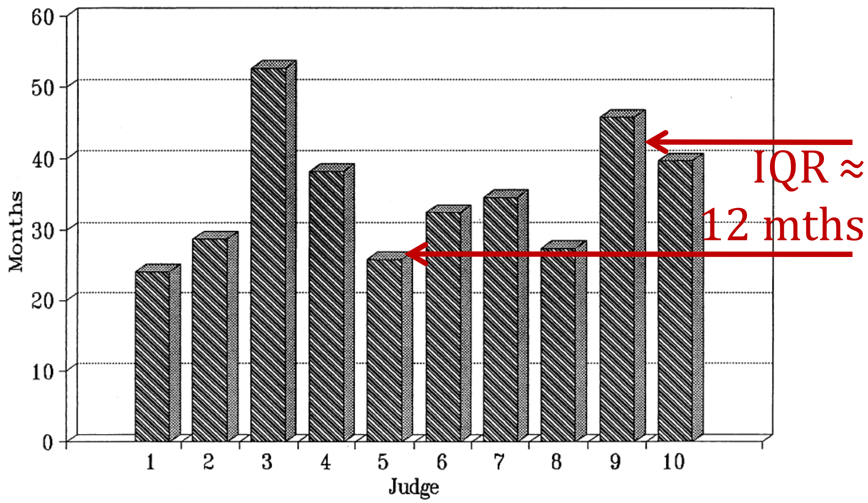
Waldfogel (1998)

"Does Inter-Judge Disparity Justify Empirically Based Sentencing Guidelines"

-Federal criminal cases in San Francisco

- ≈100 cases per judge

-Finds large variation in sentence lengths

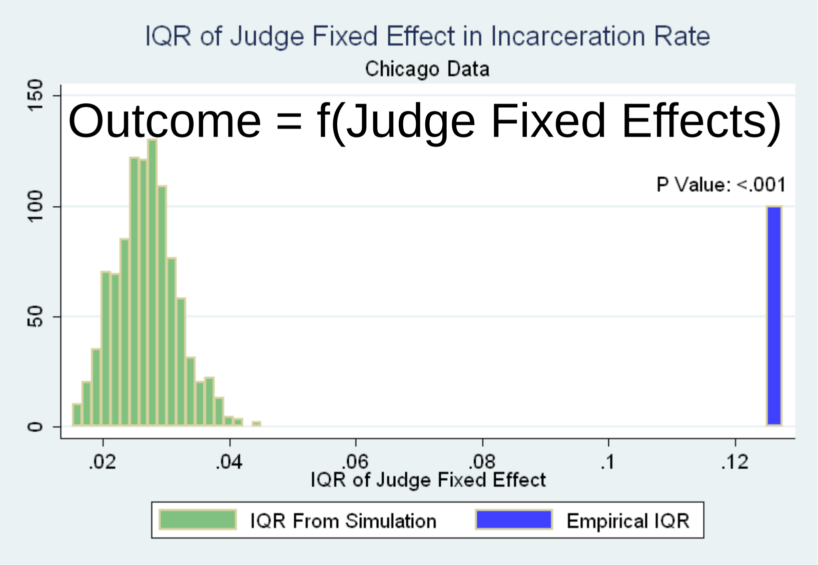


IQR ≈ 12 mths

FIG. 1. Average prison terms by judge for the Northern District of California from 1984 to 1987.

Contributions of this paper

1. Methodological
   Use of resampling methods to assess whether inter-judge disparity is larger than random
   - Why aren't asymptotic results useful here?



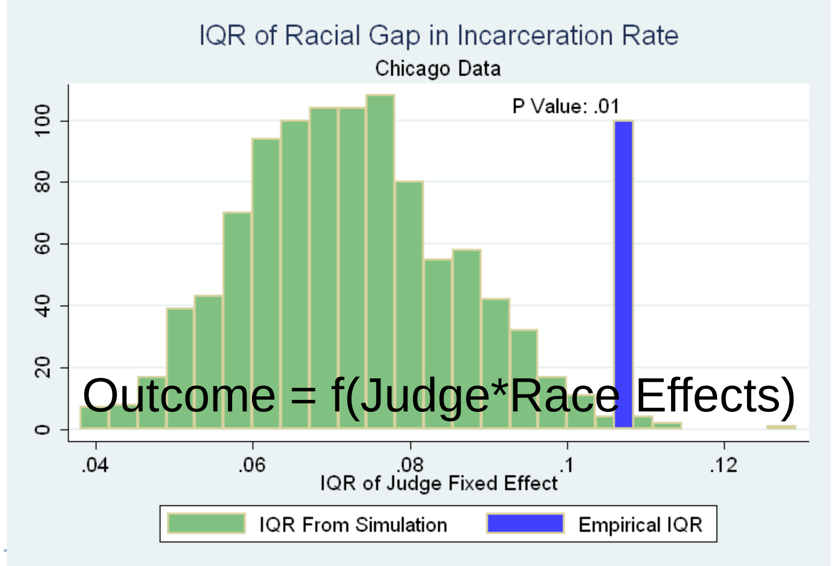
IQR of Judge Fixed Effect in Incarceration Rate

Chicago Data

Outcome = f(Judge Fixed Effects)

P Value: <.001

2. Substantive
   Test whether:
   - Judges vary in treatment of race
   - Alternatively phrased:
     Inter-judge disparity differs in a sample of black v. white defendants



IQR of Racial Gap in Incarceration Rate

Chicago Data

Outcome = f(Judge*Race Effects)

P Value: .01

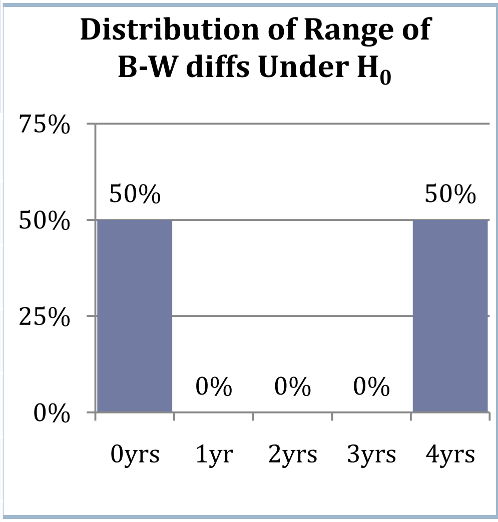
A Menu of Methods

- Parametric hypothesis testing: Analytic solutions
  - Hypothesis testing: Solve for the distribution under the null
    - Most common : Asymptotic results as $n \to \infty$ (eg F-test)
    - Exact tests: Yield exact sampling distribution (eg Permutation test)
      - Analytically difficult
- Resampling methods
  - Bootstrap: Estimate the sampling distribution of an estimator
    - By drawing randomly *with replacement* from data
    - Creates alternative "samples" that might have arisen
  - Randomization test: Hypothesis testing
    Estimate the sampling distribution of a test statistic *under the null of exchangability*
    - Exchangability: Changing the labels on observations has no effects
      - This paper: H_0 Changing the names of judges assigned to cases has no effect
      - Changing labels = Draw alternative assignments *without replacement*

An example of a randomization test

- Consider sentences of cases assigned to two judges
  - Judge Judy: Billy gets 8 years; Willy gets 8 years; (B-W Diff=0)
  - Judge Dredd: Bally gets 10 years; Wally gets 6 years; (B-W Diff=4)
  - Competing interpretations of variation in B-W diff (from 0 yrs to 4 yrs)
    - H_0: No discrimination: This reflects variation in cases (Bally worse than Billy; Wally better than Willy)
    - H_1: Discrimination: Reflects variation in racial bias (Dredd more biased than Judy)
- Randomization test:

	Judge Judy			Judge Dredd			Range of B-W diffs
	Assignment		B-W diff	Assignment		B-W diff	
Data	**Billy (8)**	**Willy (8)**	**0**	**Bally (10)**	**Wally (6)**	**4**	**4 yrs**
Alt. 1	Billy (8)	Wally (6)	2	Bally (10)	Willy (8)	2	0 yrs
Alt. 2	Billy (8)	Willy (8)	0	Bally (10)	Wally (6)	4	4 yrs
Alt. 3	Bally (10)	Wally (6)	4	Billy (8)	Willy (8)	0	4 yrs
Alt. 4	Bally (10)	Willy (8)	2	Billy (8)	Wally (6)	2	0 yrs
Alt. 5	*Billy (8)*	*Bally (10)*	*X*	*Wally (6)*	*Willy (8)*	*X*	*X*
Alt. 6	*Wally (6)*	*Willy (8)*	*X*	*Billy (8)*	*Bally (10)*	*X*	*X*



Distribution of Range of B-W diffs Under H_0

- Under the null of no discrimination, p=0.5 that a judge will look 4 years "more racist"

Problem with Abrams: Re-sampling with Replacement

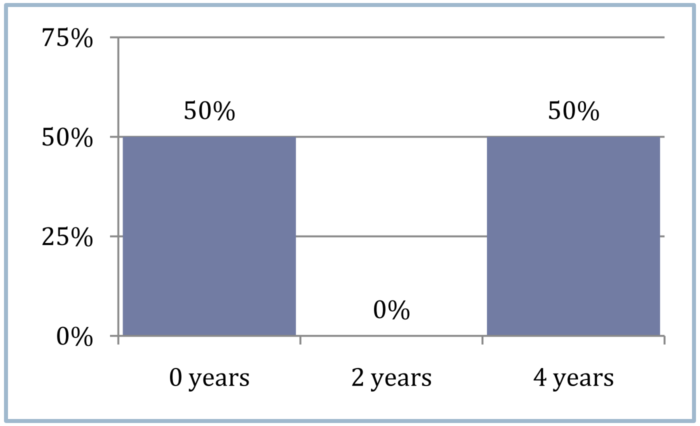
- ❑ Implement a randomization test, *with replacement*

	Judge Judy			Judge Dredd			Range of B-W diffs
	Assignment		B-W diff	Assignment		B-W diff	
Data	**Billy (8)**	**Willy (8)**	**0**	**Bally (10)**	**Wally (6)**	**4**	**4 years**
Alt. 1	Billy (8)	Wally (6)	2	Bally (10)	Willy (8)	2	0 years
Alt. 1b	*Billy (8)*	*Wally (6)*	*2*	*Bally (10)*	*Wally (6)*	*4*	*2 years*
Alt. 1c	*Billy (8)*	*Wally (6)*	*2*	*Billy (8)*	*Willy (8)*	*0*	*2 years*
Alt. 1d	*Billy (8)*	*Wally (6)*	*2*	*Billy (8)*	*Wally (6)*	*2*	*0 years*
Alt. 2	Billy (8)	Willy (8)	0	Bally (10)	Wally (6)	4	4 years
Alt. 2b	*Billy (8)*	*Willy (8)*	*0*	*Bally (10)*	*Willy (8)*	*2*	*2 years*
Alt. 2c	*Billy (8)*	*Willy (8)*	*0*	*Billy (8)*	*Willy (8)*	*0*	*0 years*
Alt. 2d	*Billy (8)*	*Willy (8)*	*0*	*Billy (8)*	*Wally (6)*	*2*	*2 years*
Alt. 3	Bally (10)	Wally (6)	4	Billy (8)	Willy (8)	0	4 years
…3b, 3c, 3d							*0, 2, 2 years*
Alt. 4	Bally (10)	Willy (8)	2	Billy (8)	Wally (6)	2	0 years
…4b, 4c, 4d							*0, 2, 2 years*

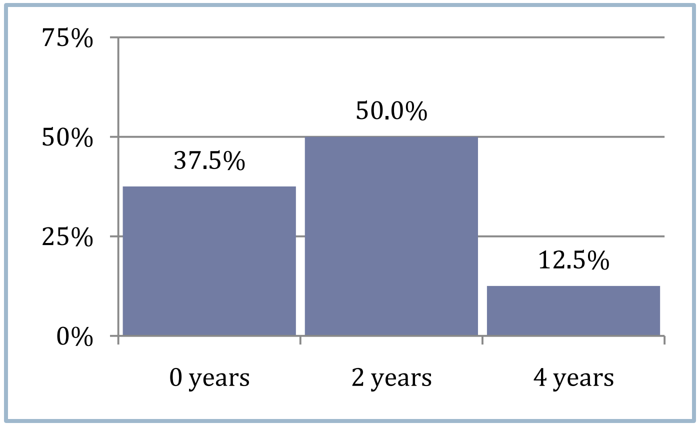
Implication of re-sampling with replacement

- Under the null of no discrimination: Estimate range of B-W diffs

True distribution

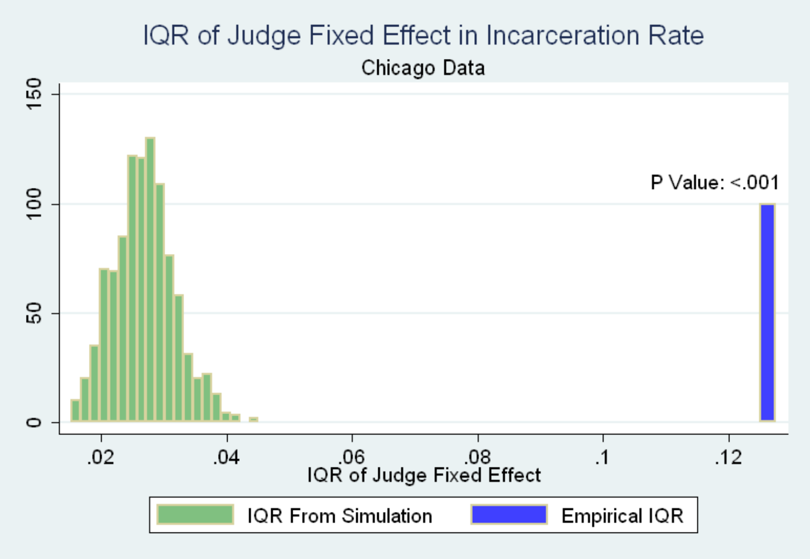


Estimated dist’n with replacement

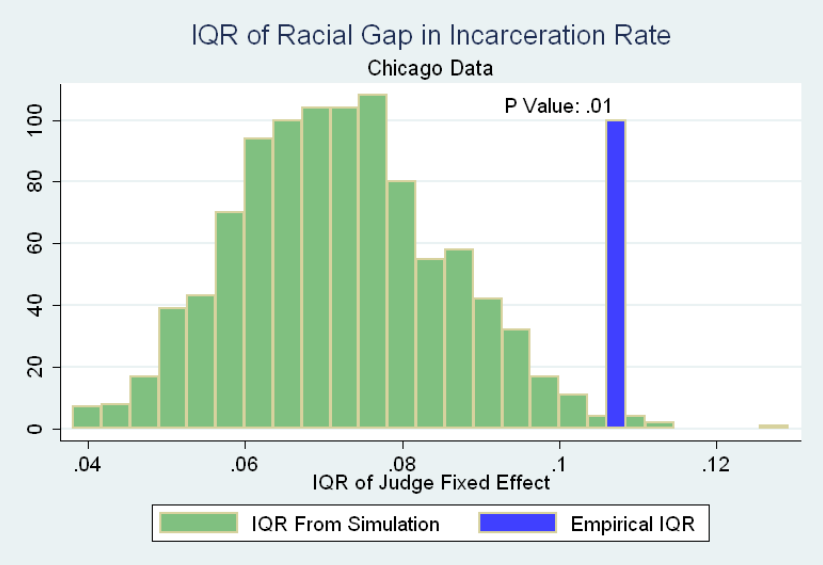


- Implies: Too easy to reject null
- Intuition: Allowing replacement gives judges the same case
  - Simulation: Judges have similar records because they get same cases
  - Reality: Judges get very different cases
    ⇒ This randomness leads even unbiased judges to have very different records

Implications for Key Findings


IQR of Judge Fixed Effect in Incarceration Rate

Chicago Data


IQR of Racial Gap in Incarceration Rate

Chicago Data

Table 7A: Dispersion of Racial Gap in Sentencing and Incarceration Rate

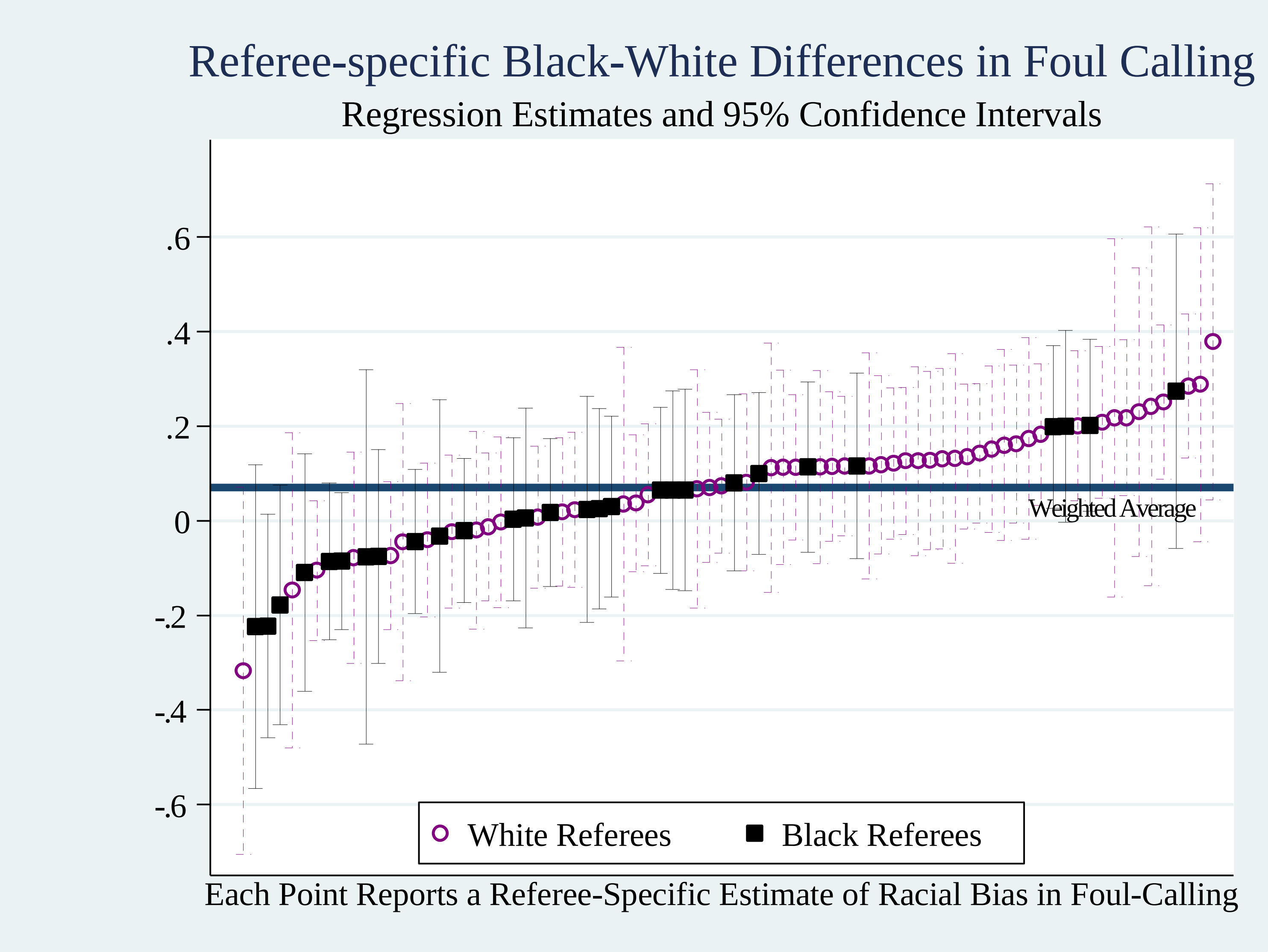
Variable Name	Empirical IQR	Simulation Mean	Simulation St Dev	P Value	Observations
jail	0.11	0.07	0.01	0.01	34298
sentence	90.50	150.35	29.17	0.98	34298
sentence2	238.36	295.21	53.51	0.85	16825

- Open question: What is the quantitative importance of these findings?
  - Waldfogel's findings:
    - Offense and offender characteristics explain 34% of variation in sentence length
    - ...interacted with judge fixed effects raises this to 43%

Substantive Interpretation

- Fact: Treatment of blacks v. whites varies across judges
- Implication: Racism exists
  - Reject null that no judges are racially biased
  - Reject null that all judges are equally racially biased
  - Question from a Bayesian: After reading this paper, is there:
    - Evidence of more discrimination than I previously thought?
    - Evidence of less discrimination than I previously thought?
- Questions about interpretation
  - Is race the mediating variable?
    - Disparate impact v. disparate treatment
    - Disparate treatment interpretation rests heavily on no omitted variables
  - Are some judges anti-black, or are some judges pro-black?
  - What is the role of judge's characteristics (own-race bias)?

Own-Race Bias: NBA “Judges”


Referee-specific Black-White Differences in Foul Calling
Regression Estimates and 95% Confidence Intervals
Weighted Average
○ White Referees
■ Black Referees
Each Point Reports a Referee-Specific Estimate of Racial Bias in Foul-Calling