The threshold test

A new test for biased decision-making

Camelia Simoiu, Sam Corbett-Davies
Emma Pierson, Sharad Goel

Stanford University

Traffic stops

- Traffic stops are the primary way in which the public interacts with law enforcement
- Concern of racial bias in police actions
- Seemingly reasonable tests of discrimination can give misleading results

Our contribution

- Novel test for discrimination, “threshold test” to measure racial bias in officers' **decision to search**

- Are minorities subjected to a search on the basis of less evidence than whites ?

- Bayesian hierarchical latent variable model

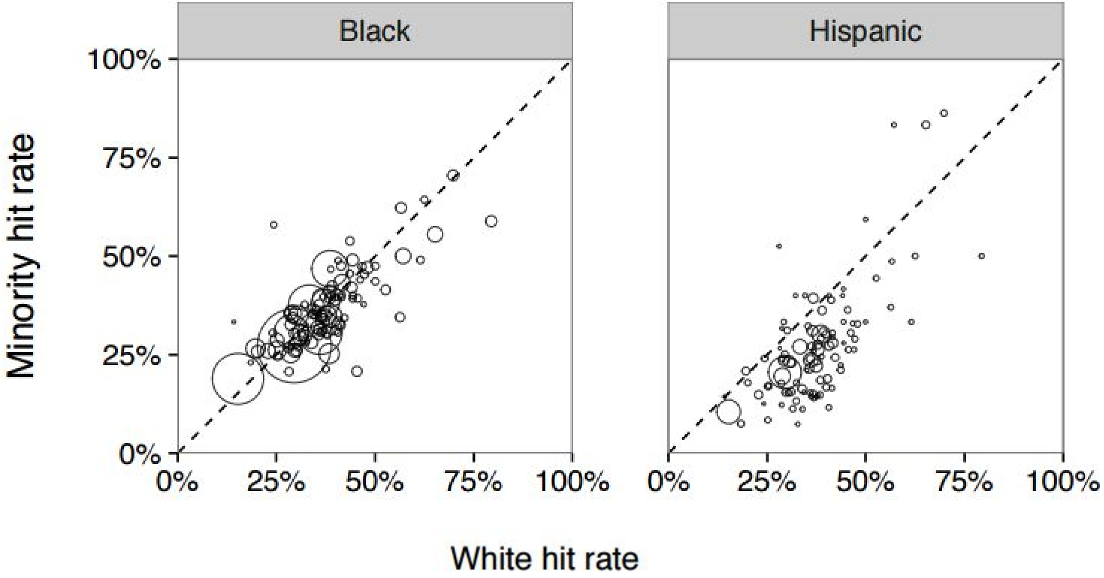
North Carolina Data Set

- 4.5 million stops
- 6 year observation period: 2009-2014
- Largest 100 local police departments
  - account for 90% of local stops
- 3 race groups (White, Black, Hispanic)
- Search rates for each race group typically range between 4% - 8%

Standard Tests of Discrimination

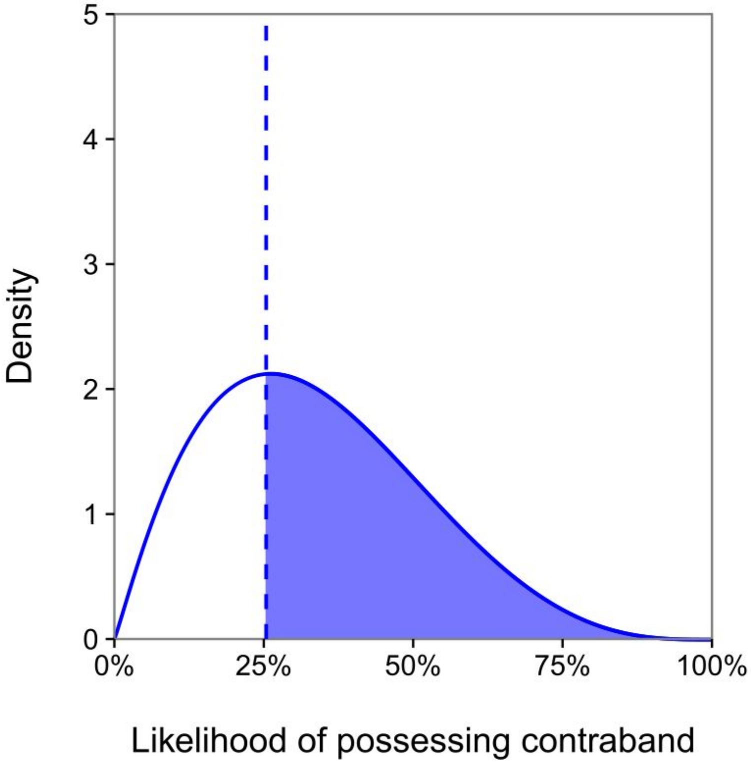
Outcome Test [Becker 1957, 1992]

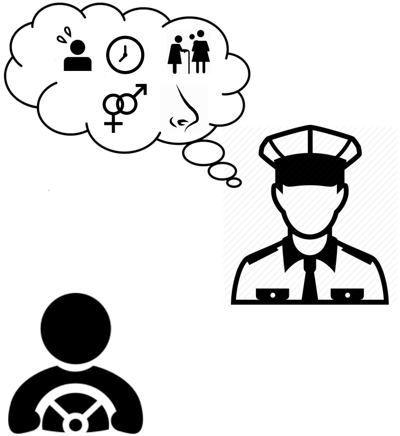
Compare the search success (hit) rate across race groups



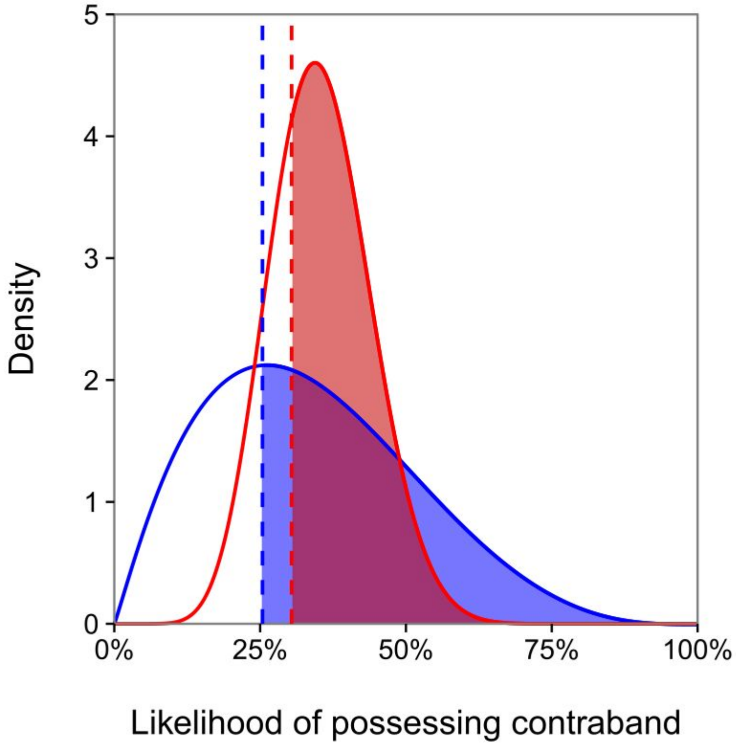
Race	Hit Rate
White	36%
Black	32%
Hispanic	23%

Problem of infra-marginality [Ayers, 2002]





Problem of infra-marginality [Ayers, 2002]



Discrimination against Blue by construction.

Outcome test fails to identify discrimination against Blue and suggests discrimination against red.

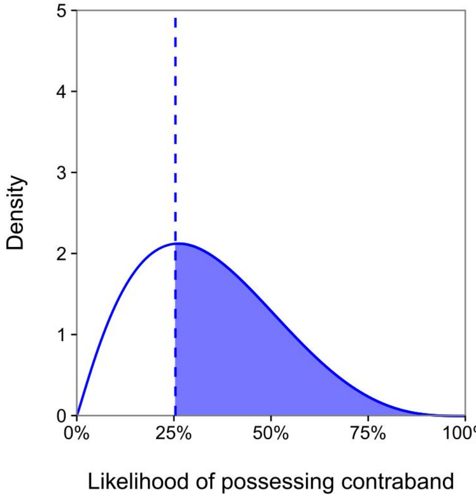
	Red	**Blue**
Search rate	71%	64%
Hit rate	39%	44%

Threshold Model

Modeling a Traffic Stop

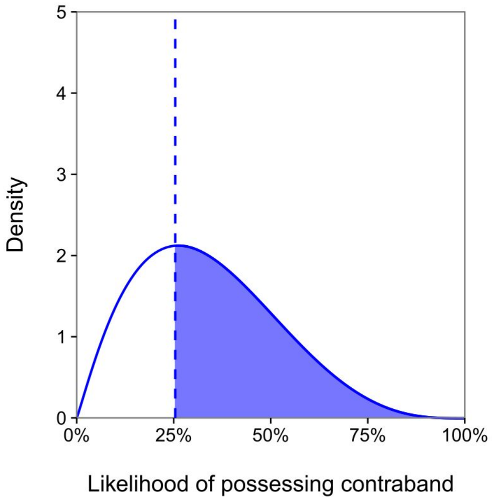
- Officer in department *d* stops a driver of race *r*
- Each driver has a risk of possessing contraband:

$$x_i \sim \mathrm{Beta}(\Phi_{rd}, \lambda_{rd})$$



Modeling a Traffic Stop

- Officer in department *d* stops a driver of race *r*
- Each driver has a risk of possessing contraband:
  $x_i \sim \mathrm{Beta}(\Phi_{rd}, \lambda_{rd})$
- Deterministically conduct search $S_i = 1$ iff $x_i > t_{rd}$
- If $S_i = 1$: $H_i \sim \mathrm{Bernoulli}(x_i)$
- Lower t_{rd} indicate discrimination



Parameterizing the Risk Distribution

$x_i \sim \text{Beta}(\Phi_{rd}, \lambda_{rd})$

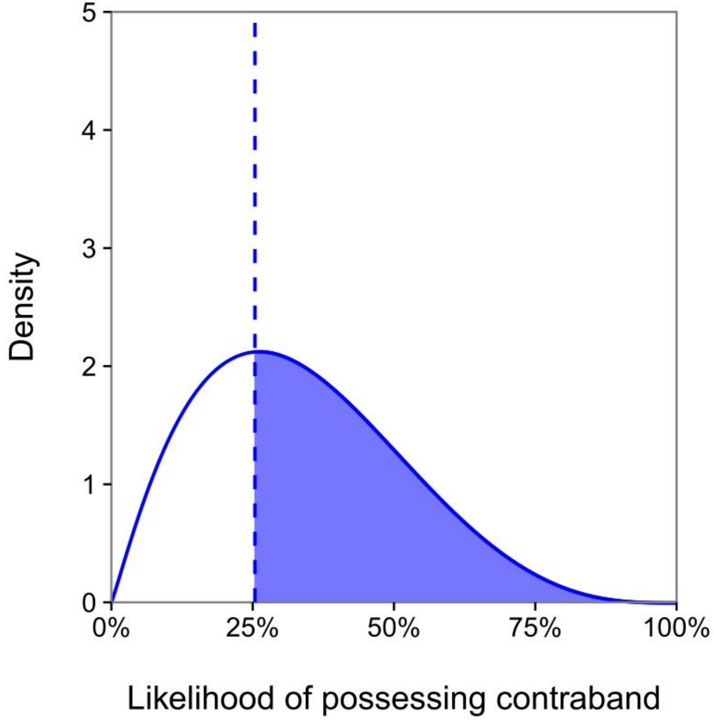
$\Phi_{rd} \sim \text{logit}^{-1}(\Phi_r + \Phi_d)$

Probability that a randomly stopped driver is carrying contraband

$\lambda_{rd} \sim \exp(\lambda_r + \lambda_d)$

Difficulty in distinguishing between guilty and innocent drivers

Simplifying inference



For a given department *d*, race *r*

Observe N_{rd} stops

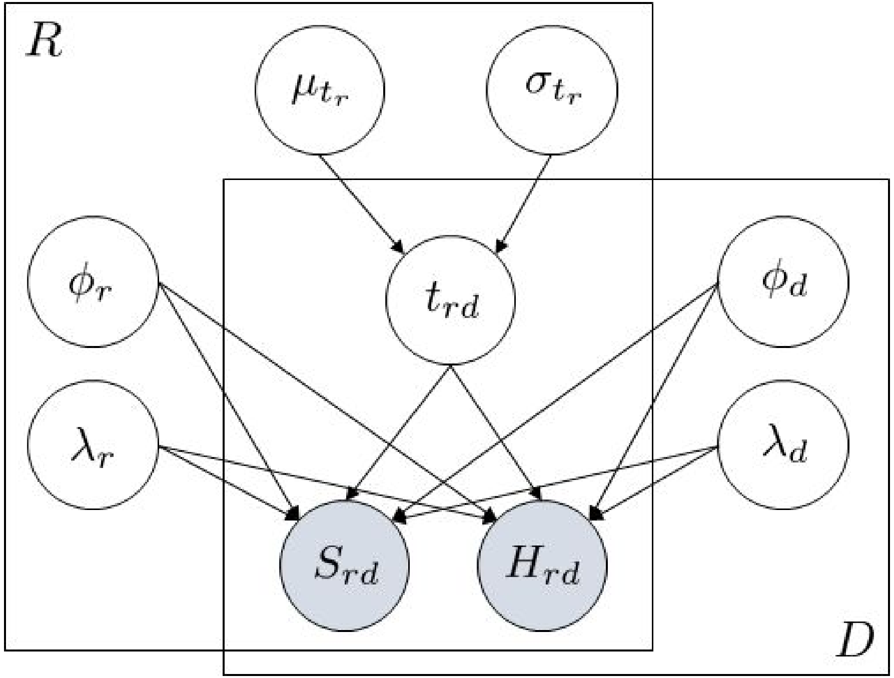
$$x_{rd} \sim \text{Beta}(\Phi_{rd}, \lambda_{rd})$$

$$\delta_{rd} = P(x_{rd} > t_{rd} \; ; \; \Phi_{rd}, \lambda_{rd})$$

$$\gamma_{rd} = E(x_{rd} \mid x_{rd} > t_{rd} \; ; \; \Phi_{rd}, \lambda_{rd})$$

$$\boxed{\begin{aligned} S_{rd} &= \text{Binomial}(\delta_{rd}, N_{rd}) \\ H_{rd} &= \text{Binomial}(\gamma_{rd}, S_{rd}) \end{aligned}}$$

Graphical Model Representation



Speeding up inference

Speeding up inference

- The threshold test requires repeated computation of complicated derivatives of beta CDFs
- MCMC can take hours (~2 hours in North Carolina)
- Can we choose an alternative to the beta distribution that has better computational properties?

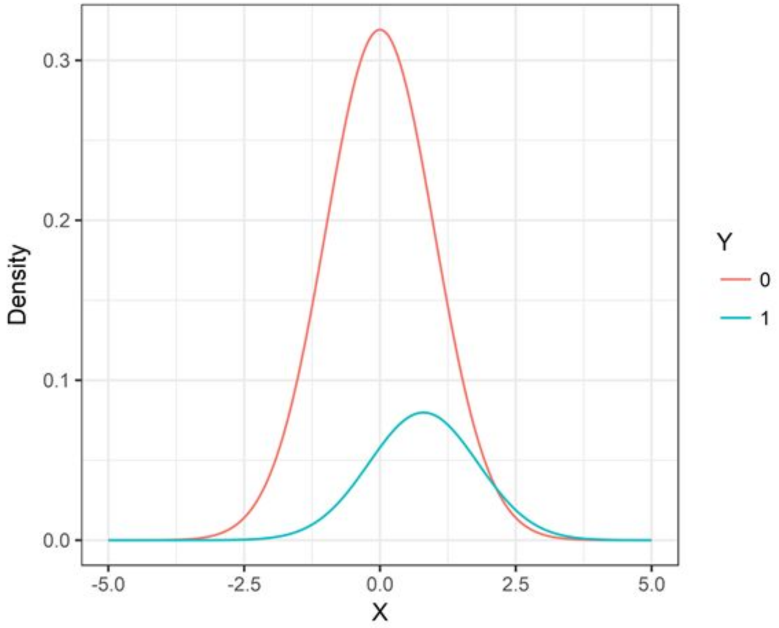
Mixture of Gaussians

$$Y \sim \text{Bernoulli}(p)$$

$$X \mid Y = 0 \sim N(\mu_0, \sigma_0)$$
$$X \mid Y = 1 \sim N(\mu_1, \sigma_1)$$

$$g(x; p, \mu_0, \mu_1, \sigma_0, \sigma_1) = P(Y = 1 | X = x)$$

$$Z = g(X)$$



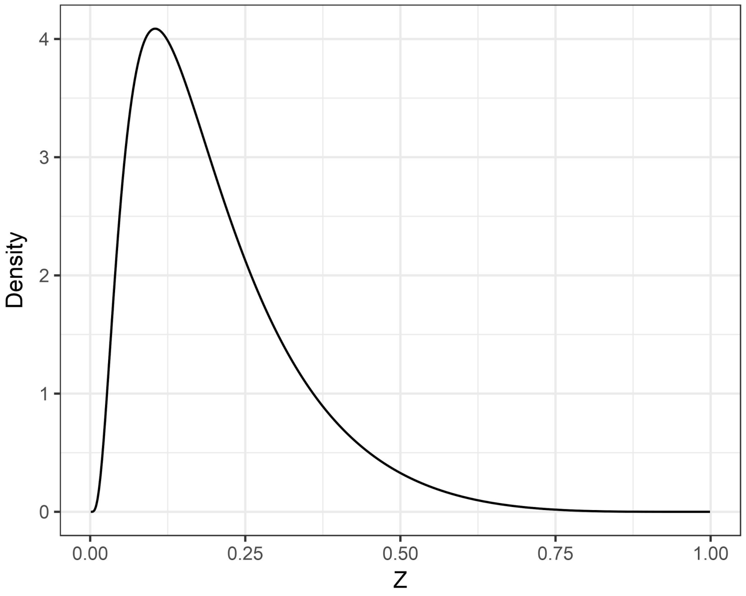
Mixture of Gaussians

$$Y \sim \text{Bernoulli}(p)$$

$$X \mid Y = 0 \sim N(\mu_0, \sigma_0)$$
$$X \mid Y = 1 \sim N(\mu_1, \sigma_1)$$

$$g(x; p, \mu_0, \mu_1, \sigma_0, \sigma_1) = P(Y = 1 | X = x)$$

$$Z = g(X)$$



- Z is distributed on (0,1)
- The distribution is defined by 5 parameters

Discriminant distributions

$$Z = g(X; p, \mu_0, \mu_1, \sigma_0, \sigma_1) = P(Y = 1|X)$$

- For the mapping $g(X)$ to be monotonic we require the mixture to be homoscedastic (ie $\boldsymbol{\sigma_0 = \sigma_1 = \sigma}$)
- Then the density of Z only depends on $\boldsymbol{p}$ and $\boldsymbol{\delta = \frac{\mu_1 - \mu_0}{\sigma}}$
- So, w.l.o.g, we can define $\boldsymbol{\mu_0 = 0}$, $\boldsymbol{\sigma = 1}$, and $\boldsymbol{\mu_1 = \delta}$, resulting in a 2-parameter family of distributions on (0,1)

Key calculations in the threshold test

- By doing all computations in signal space the search and hit rates are simple functions of the Gaussian CDF:

- Search rate $P(Z > t)$:

$$P(Z > t) = P(X > g^{-1}(t)|Y = 1)P(Y = 1) + P(X > g^{-1}(t)|Y = 0)P(Y = 0)$$

- Hit rate $E[Z|Z > t]$:

$$E[Z|Z > t] = \frac{P(X > g^{-1}(t)|Y = 1)P(Y = 1)}{P(Z > t)}$$

A threshold test practical for broad use

- 53x speed up in gradient computation
- 3x fewer gradient computations needed per sample
- **150x speed up in inference**
- 5x more “effective samples” per sample

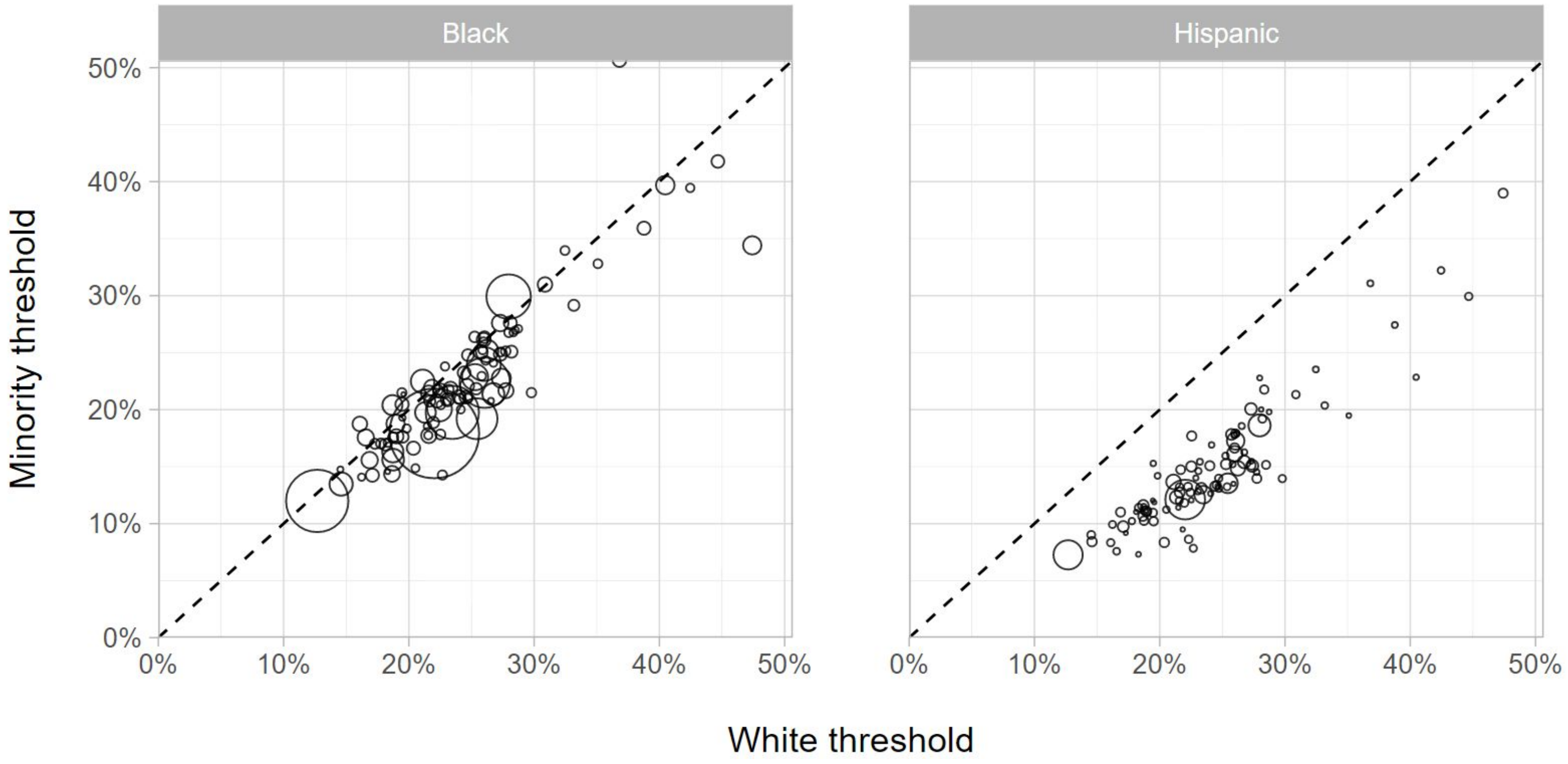
Results

Results

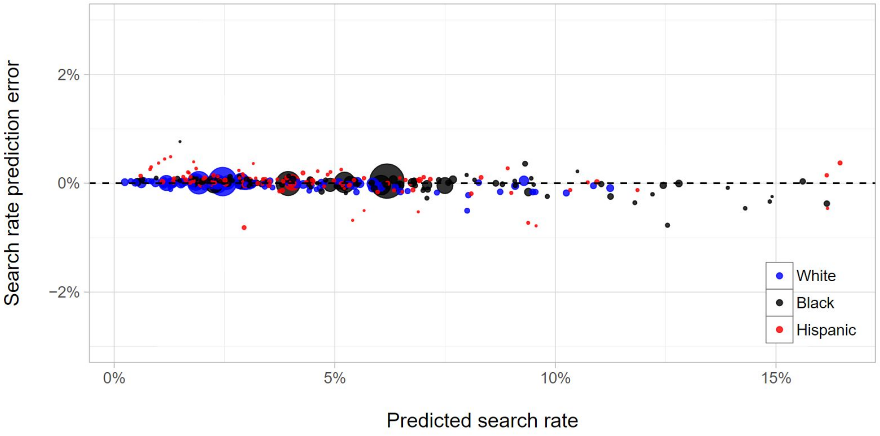
- Black and Hispanic drivers face a lower threshold than whites in North Carolina

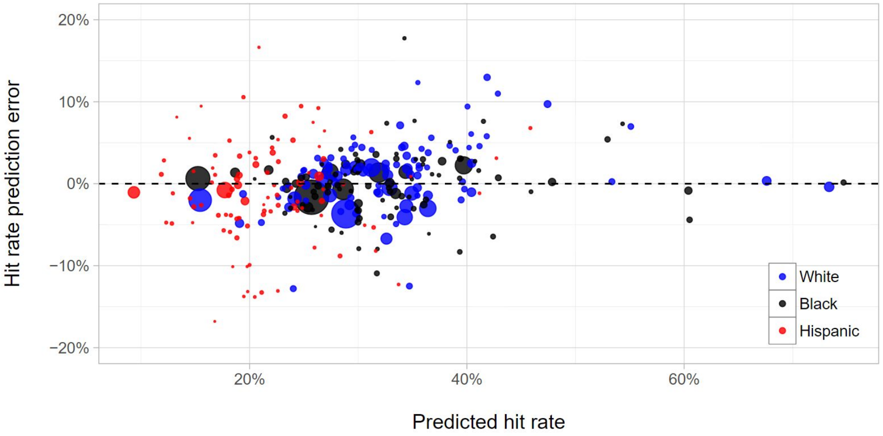
Race	Search Threshold	95% CI
White	23.0%	(22.2%, 23.8%)
Black	20.6%	(19.7%, 21.5%)
Hispanic	13.7%	(13.0%, 14.4%)

Results


Black
Hispanic
Minority threshold
White threshold
0%
10%
20%
30%
40%
50%

Posterior Predictive Checks





RMS prediction error 0.09pp

RMS prediction error 2.8pp

Conclusions

- Bayesian latent variable model allows for direct estimation of thresholds – providing a solution to the problem of infra-marginality

- We find *unjustified disparate impact* against African American and Hispanic drivers in North Carolina.

- Very few legitimate reasons to apply different search thresholds (different types of contraband)

- Could be implicit bias – overestimation of risk for minority drivers

openpolicing.stanford.edu

Thank you

"The Problem of Infra-marginality in Outcome Tests for Discrimination"
Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel

"Fast Threshold Tests for Detecting Discrimination"
Emma Pierson, Sam Corbett-Davies, and Sharad Goel.