



# Internal benchmarking using propensity scores for detecting racial bias in police traffic stops

Greg Ridgeway  
RAND

John MacDonald  
RAND & U. Penn

September 7, 2006

# ***Racial profiling is a growing concern***

## Introduction

❖ Racial profiling is a growing concern

❖ Analytic quality is weak

## Internal benchmarking

## Flagging officers

## Conclusions

- I-95 “turnpike” studies in the mid-1990s raised public concern about racial profiling
  - ❖ Concrete evidence of racial profiling policies
- Public concern has led to state and local-level action
  - ❖ At least 25 states have passed legislation
  - ❖ Many localities collect data voluntarily; some are compelled by the Justice Department
  - ❖ Hundreds of police agencies now compile race data on all stopped motorists
- Congress considering the End of Racial Profiling Act
  - ❖ Mandates data collection to receive Federal funds

# *Analytic quality is weak*

## Introduction

❖ Racial profiling is a growing concern

❖ Analytic quality is weak

## Internal benchmarking

## Flagging officers

## Conclusions

- A growing number of studies claim racial profiling based on analysis of data collected
  - ❖ **Texas:** Concluded that “75% of agencies stop more black and Latino drivers than white drivers”
- And some studies hastily conclude no profiling occurs based on analyzed data
  - ❖ **Sacramento:** Found that the percentage of black drivers stopped matched the percentage of blacks among crime suspect descriptions
- Even those that are carefully executed provide little guidance to departments about how they should adapt

# Internal benchmark

Introduction

Internal  
benchmarking

❖ Internal  
benchmark

❖ Internal  
benchmark

❖ Propensity score  
weighting

Flagging officers

Conclusions

- Consider a particular officer #534
- 71% of this officer's stops involve a black driver

|       |            | Percentage |
|-------|------------|------------|
| Time  | (12-4pm]   | 9          |
|       | (4-8pm]    | 57         |
|       | (8pm-12am] | 34         |
|       |            |            |
| Day   | Mon        | 20         |
|       | Tue        | 12         |
|       | Wed        | 12         |
|       | ⋮          | ⋮          |
|       |            |            |
| Month | Jan        | 12         |
|       | Feb        | 14         |
|       | Mar        | 7          |
|       | Apr        | 6          |
|       | May        | 8          |
|       | ⋮          | ⋮          |
|       |            |            |
| Area  | J          | 49         |
|       | K          | 33         |
|       | L          | 5          |
|       | M          | 11         |
|       |            |            |

# Internal benchmark

Introduction

Internal  
benchmarking

❖ Internal  
benchmark

❖ Internal  
benchmark

❖ Propensity score  
weighting

Flagging officers

Conclusions

- 46% of similarly situated stops made by other officers involved black drivers

|       |            | Percentage | Comparison |
|-------|------------|------------|------------|
| Time  | (12-4pm]   | 9          | 9          |
|       | (4-8pm]    | 57         | 56         |
|       | (8pm-12am] | 34         | 35         |
| Day   | Mon        | 20         | 20         |
|       | Tue        | 12         | 11         |
|       | Wed        | 12         | 12         |
|       | ⋮          | ⋮          | ⋮          |
| Month | Jan        | 12         | 12         |
|       | Feb        | 14         | 15         |
|       | Mar        | 7          | 7          |
|       | Apr        | 6          | 6          |
|       | May        | 8          | 7          |
|       | ⋮          | ⋮          | ⋮          |
| Area  | J          | 49         | 48         |
|       | K          | 33         | 34         |
|       | L          | 5          | 5          |
|       | M          | 11         | 11         |

# Propensity score weighting

Introduction

Internal  
benchmarking

❖ Internal  
benchmark

❖ Internal  
benchmark

❖ Propensity score  
weighting

Flagging officers

Conclusions

- Reweight stops that other officers made so that they have the same distribution of features

$$f(\mathbf{x}|t = 1) = w(\mathbf{x})f(\mathbf{x}|t = 0)$$

- Solving for  $w(\mathbf{x})$  yields the propensity score weight

$$w(\mathbf{x}) = \frac{f(t = 1|\mathbf{x})}{f(t = 0|\mathbf{x})} K = \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} K$$

where  $p(\mathbf{x})$  is the probability that a stop with features  $\mathbf{x}$  involves the officer in question

- Estimate  $p(\mathbf{x})$  using a flexible, non-parametric version of logistic regression
- Compare the percentage of black drivers among the officer's stops with the weighted percentage of black drivers among other stops using weights

$$w_i = p(\mathbf{x}_i)/(1 - p(\mathbf{x}_i))$$

# Results

Introduction

Internal  
benchmarking

Flagging officers

❖ Results

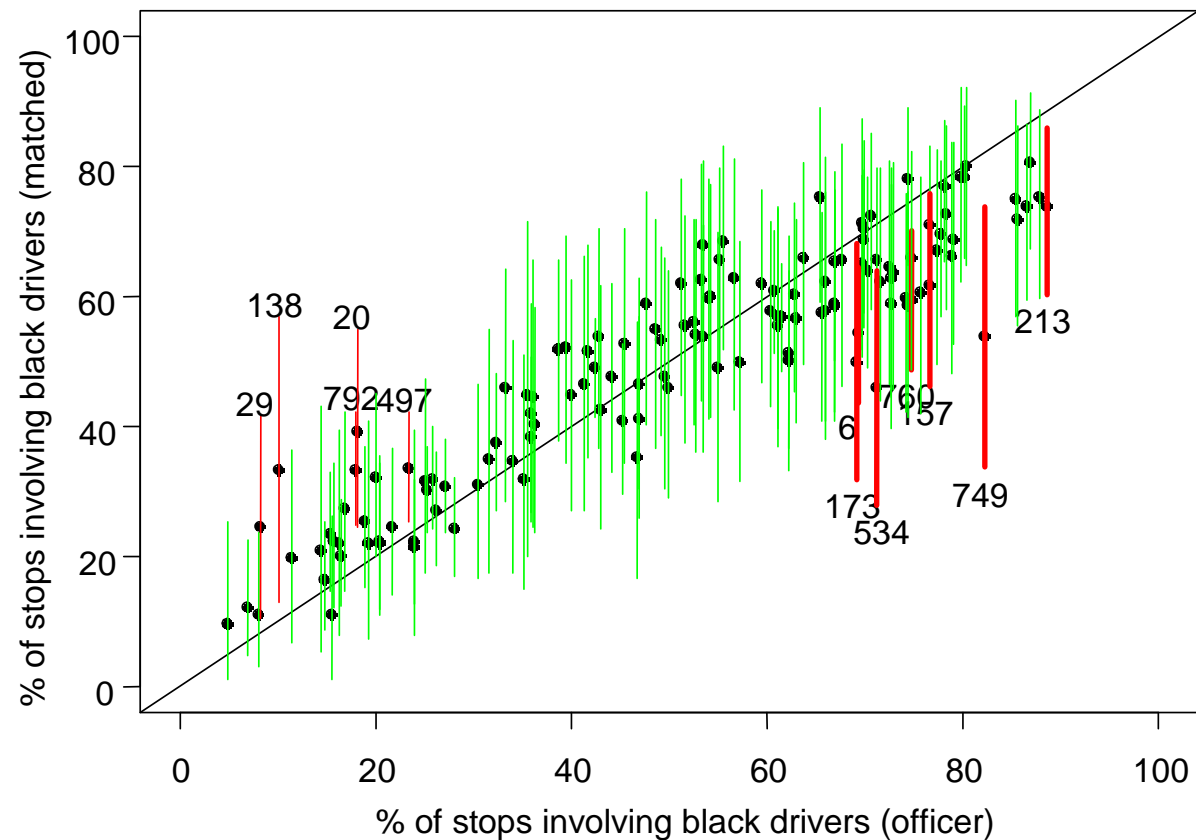
❖ Common  
approach

❖ False discovery  
rate

❖ Estimating fdr

Conclusions

- Seven officers have a substantially greater fraction of stopped black drivers than their internal benchmark





# Common approach

Introduction

Internal  
benchmarking

Flagging officers

❖ Results

❖ Common  
approach

❖ False discovery  
rate

❖ Estimating fdr

Conclusions

- A common approach is to compute z-statistics for each officer

$$z = \frac{p_t - p_c}{\sqrt{\frac{p_t(1-p_t)}{n_t} + \frac{p_c(1-p_c)}{ESS}}}$$

- In the absence of racial bias this would be distributed  $N(0,1)$  and a cutoff of 2.0 would be reasonable
- With 133 officers and 133 correlated  $z$ s an appropriate reference distribution can be much wider (Efron 2006).

# False discovery rate

Introduction

Internal  
benchmarking

Flagging officers

❖ Results

❖ Common  
approach

❖ False discovery  
rate

❖ Estimating fdr

Conclusions

- Benjamini and Hochberg (1995) pioneered the use of the false discovery rate (fdr)

$$\begin{aligned} P(\text{problem}|z) &= 1 - P(\text{no problem}|z) \\ &= 1 - \frac{f(z|\text{no problem})f(\text{no problem})}{f(z)} \\ &\geq 1 - \frac{f_0(z)}{f(z)} \end{aligned}$$

- If the fraction of problem officers is small then the last inequality is a tight bound

# Estimating fdr

Introduction

Internal  
benchmarking

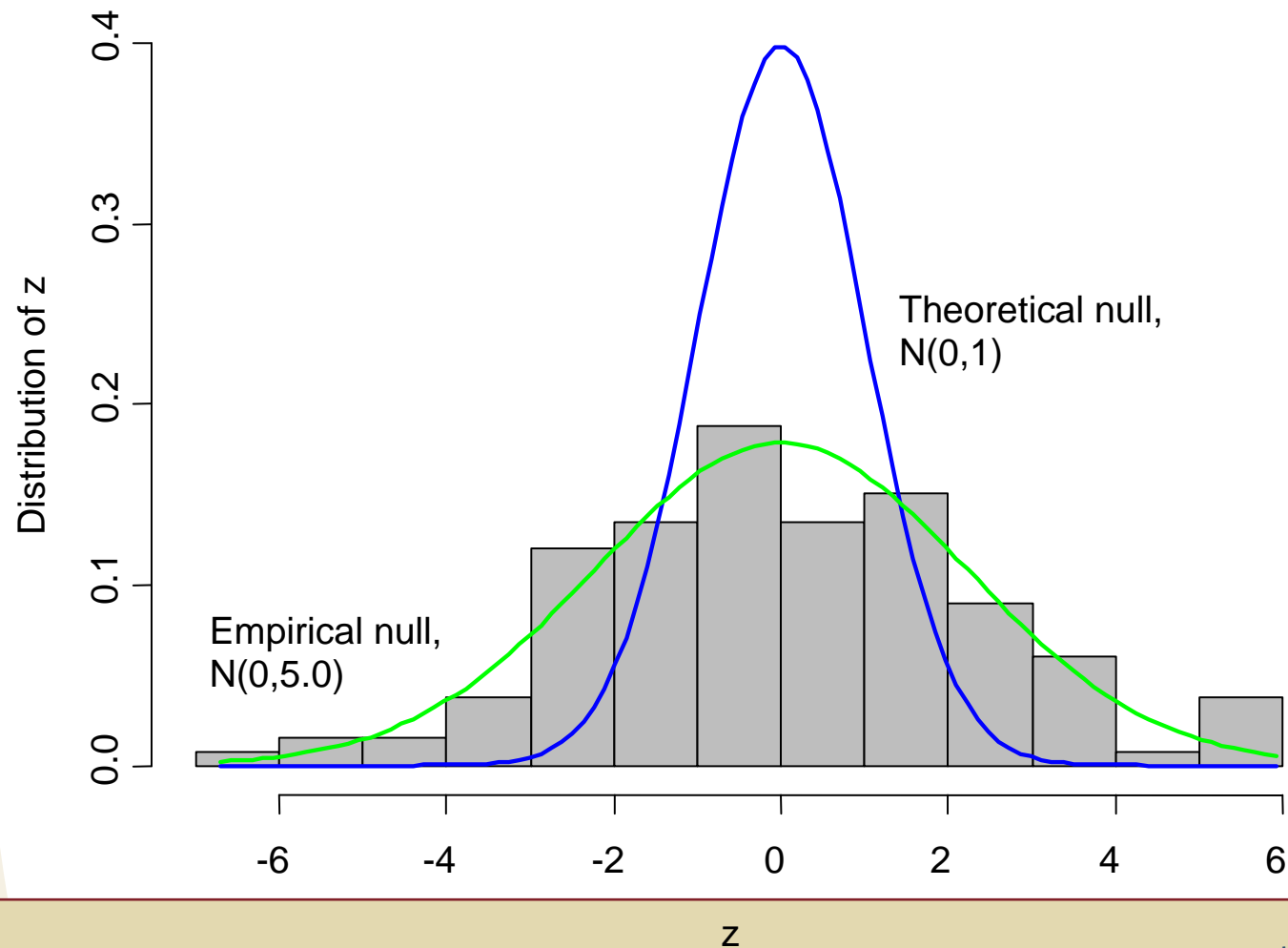
Flagging officers

❖ Results  
❖ Common  
approach  
❖ False discovery  
rate

❖ Estimating fdr

Conclusions

- Estimate  $f_0(z)$  and  $f(z)$  from the observed  $z$ s
- Right tail consists of 5 officers with “problem officer” probabilities ranging from 70% to 86%



# Conclusions

Introduction

Internal  
benchmarking

Flagging officers

Conclusions

❖ Conclusions

- Internal benchmarking can help identify problem officers
- Propensity score weighting offers a sound process for constructing the internal benchmark
- Flagging particular officers requires dealing with the issues of massive multiple comparisons
- False discovery rate offer a promising direction