# Testing for Racial Discrimination in Police Searches of Motor Vehicles*

Camelia Simoiu
Stanford University

Sam Corbett-Davies
Stanford University

Sharad Goel
Stanford University

**Abstract**

In the course of conducting traffic stops, officers have discretion to search motorists for drugs, weapons, and other contraband. There is concern that these search decisions are prone to racial bias, but it has proven difficult to rigorously assess claims of discrimination. Here we develop a new statistical method—the threshold test—to test for racial discrimination in motor vehicle searches. We use geographic variation in stop outcomes to infer the effective race-specific standards of evidence that officers apply when deciding whom to search, an approach we formalize with a hierarchical Bayesian latent variable model. This technique mitigates the problems of omitted variables and infra-marginality associated with benchmark and outcome tests for discrimination. On a dataset of 4.5 million police stops in North Carolina, we find that the standard for searching black and Hispanic drivers is considerably lower than the standard for searching white and Asian drivers, a pattern that holds consistently across the 100 largest police departments in the state.

*We thank Cheryl Phillips and Vignesh Ramachandran of the Stanford Computational Journalism Lab for compiling the North Carolina traffic stop data, and the John S. and James L. Knight Foundation for partial support of this research. We also thank Stefano Ermon, Avi Feller, Seth Flaxman, Andrew Gelman, Lester Mackey, Jan Overgoor, and Emma Pierson for helpful comments.

# 1 Introduction

The police conduct more than 20 million traffic stops across the United States each year, making it one of the most common ways in which the public interacts with law enforcement [Langton and Durose, 2013]. The vast majority of these stops stem from routine traffic violations, such as speeding, and typically result in issuance of a citation or warning, without further police action. However, when officers suspect more serious criminal activity, they have latitude to search both driver and vehicle for drugs, weapons, and other contraband. Such discretionary police searches have been criticized as being prone to implicit and explicit racial bias [Epp et al., 2014], but it has been challenging to empirically corroborate these claims.

The difficulty of rigorously assessing claims of bias in police searches is due in large part to well-known problems with the two most common statistical tests for discrimination. In the first, termed *benchmarking*, one compares the rate at which minorities are searched to that of whites, with higher search rates of minorities suggestive of discrimination. However, if minorities in reality carry contraband at higher rates than whites, then the disparities in search rates may simply result from good police work rather than from racial profiling. This limitation of benchmarking is referred to in the literature as the *qualified pool* or *denominator* problem [Ayres, 2002], and is a specific instance of omitted variable bias.[1]

Addressing this shortcoming of benchmarking, Becker [1957, 1993] proposed the *outcome test*, which is based not on the search rate, but on the *hit rate*, the proportion of searches that successfully turn up contraband. Becker argued that even if minority drivers are more likely to carry contraband, searched minorities, absent discrimination, should still be found to have contraband at the same rate as searched whites. If searches of minorities are less likely to be

[1] To give a more concrete example, suppose officers conduct searches when they observe signs of drug use (e.g., the smell of marijuana). If stopped minorities are more likely than whites to exhibit such signs, then minorities would be searched at a higher rate, even in the absence in discrimination. The benchmark test can thus lead to spurious conclusions when one fails to account for all legitimate factors that may prompt a search.

successful than searches of whites, it suggests that officers are applying a double standard, searching minorities on the basis of less evidence. Outcome tests, however, are also imperfect barometers of discrimination. To see this, suppose that there are two, easily distinguishable types of white drivers: those who have a 1% chance of carrying contraband, and those who have a 75% chance. Similarly assume that black drivers have either a 1% or 50% chance of carrying contraband. If officers, in a race-neutral manner, search individuals who are at least 10% likely to be carrying contraband, then searches of whites will be successful 75% of the time whereas searches of blacks will be successful only 50% of the time. This simple example shows that outcome tests can suggest discrimination even when there is none, a subtle failure known as the problem of *infra-marginality* [Ayres, 2002], a phenomenon we return to and discuss in detail below.

Our contribution in this paper is two-fold. First, we develop a new test for discrimination—the *threshold test*—that mitigates the limitations of both benchmark and outcome analysis. We specifically use geographic variation in search and hit rates to infer the effective race-specific standards of evidence that officers apply when deciding whom to search, an approach we formalize via a hierarchical Bayesian latent variable model [Gelman and Hill, 2006]. In the process of developing this test, we demonstrate that the issues with benchmark and outcome analysis noted above are not limited to pathological cases, but rather arise naturally in many settings. Second, we apply our technique to a dataset of 4.5 million traffic stops conducted by the 100 largest police departments in North Carolina between 2009 and 2014. We find that nearly every department applies a lower standard of evidence when searching blacks and Hispanics than when searching whites and Asians, suggestive of racial bias in search decisions. We note that it is not clear whether these disparities are due to racial animus, implicit bias, or statistical errors in officer judgement.

*Related work.* Benchmark analysis is the most common statistical method for assessing racial bias in police stops and searches. The key methodological challenge with this approach

is estimating the race distribution of the at-risk, or benchmark, population. Traditional benchmarks include the residential population, licensed drivers, arrestees, and reported crime suspects [Engel and Calnon, 2004]. Alpert et al. [2004] estimated the race distribution of drivers on the roadway by considering not-at-fault drivers involved in two-vehicle crashes. Others have looked at stops initiated by aerial patrols [McConnell and Scheidegger, 2001], and those based on radar and cameras [Lange et al., 2001], arguing that such stops are less prone to potential bias and thus more likely to reflect the true population of traffic violators. Studying police stops of pedestrians in New York City, Gelman et al. [2007] use a hierarchical Bayesian model to construct a benchmark based on neighborhood- and race-specific crime rates. Ridgeway [2006] studies post-stop police actions by creating benchmarks based on propensity scores, with minority and white drivers matched using demographics and the time, location, and purpose of the stops. Grogger and Ridgeway [2006] construct benchmarks by considering stops at night, when a "veil of darkness" masks race. Antonovics and Knight [2009] use officer-level demographics in a variation of the standard benchmark test: they argue that search rates that are higher when the officer's race differs from that of the suspect is evidence of discrimination.[2] Finally, "internal benchmarks" have been used to flag potentially biased officers by comparing each officer's stop decisions to those made by others patrolling the same area at the same time [Ridgeway and MacDonald, 2009, Walker, 2003].

Given the inherent limitations of benchmark analysis, researchers have more recently turned to outcome tests to investigate claims of police discrimination. For example, Goel et al. [2016b] use outcome analysis to test for racial bias in New York City's stop-and-frisk policy. While outcome tests mitigate the problem of omitted variables faced by benchmark analysis, they suffer from their own limitations, most notably infra-marginality. The problem of infra-marginality in outcome tests was first discussed in detail by Ayres [2002], although

[2]While intuitively appealing, this approach fails to detect discrimination when officers of all races are similarly prejudiced. Anwar and Fang [2006] further show that the test can indicate discrimination even in the absence of any.

previous studies of discrimination [Carr et al., 1993, Galster, 1993] indicate awareness of the issue. An early attempt to address the problem was presented by Knowles et al. [2001], who developed an economic model of behavior in which drivers balance their utility for carrying contraband with the risk of getting caught, while officers balance the utility of finding contraband with the cost of searching. Under equilibrium behavior, Knowles et al. argue that the hit rate is identical to the search threshold, and so one can reliably detect discrimination with the standard outcome test.[3] Though an interesting theoretical argument, Engel and Tillyer [2008] note that the model of Knowles et al. requires strong assumptions, including that drivers and officers are rational actors, and that every driver has perfect knowledge of the likelihood that he will be searched. Anwar and Fang [2006] propose a hybrid test of discrimination that is based on the rankings of race-contingent search and hit rates as a function of officer race: if officers are not prejudiced, they argue, then these rankings should be independent of officer race. This approach circumvents the problems of omitted variables and infra-marginality in certain cases, but it cannot pick up discrimination when officers of different races are similarly biased.

## 2 A New Test for Discrimination in Vehicle Searches

### 2.1 A model of officer behavior

We begin by introducing a stylized model of officer behavior that is the basis of our statistical approach, and which also illustrates the problem of infra-marginality. During each stop, officers observe a myriad of contextual factors—such as the age, gender, and race of the driver, and behavioral indicators of nervousness or evasiveness—and, based on these factors, decide whether or not to conduct a search. We imagine that officers distill all these complex signals down to a single number $p$ that represents their subjective estimate of the likelihood that the driver is carrying contraband. We further imagine that these estimates are calibrated,

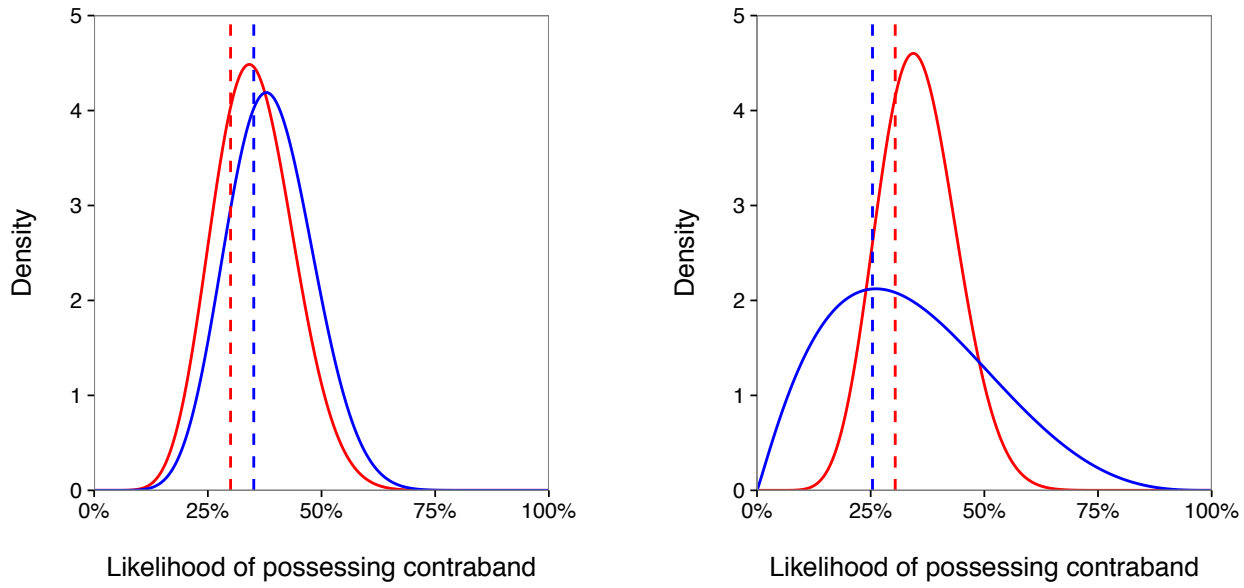
[3]This is because in equilibrium, and in the absence of discrimination, all drivers who have positive utility of carrying contraband are equally likely to do so.

meaning that the predictions are statistically consistent with the observations [Gneiting et al., 2007]: when officers think a driver has probability $p$ of having contraband, $p$-percent of such drivers do in fact have contraband. These estimates are thus analogous to a weather forecaster's prediction that there is a 30% chance of rain tomorrow: they are based on the available evidence, but there is no purely objective or "true" probability of the event occurring. Finally, we assume that officers conduct a search if and only if their subjective estimate of finding contraband exceeds a fixed, race-specific search threshold $t_r$. Under this model, if officers have a lower threshold for searching minorities than whites—for example, $t_{\text{black}} < t_{\text{white}}$, indicating that they are willing to search blacks on the basis of less evidence than whites—then we would say that minorities are being discriminated against. In the economics literature, this is often referred to as *taste-based discrimination* [Becker, 1957].[4] We treat both the subjective probabilities and the search thresholds as latent, unobserved quantities, and our goal is to infer them from data.

### 2.2 The problem of infra-marginality

Fig. 1a illustrates the setup described above for two hypothetical race groups, where the curves show race-specific *signal distributions* (i.e., the distribution of officers' subjective estimates of guilt across all stopped motorists of that race), and the vertical lines indicate race-specific search thresholds. In this example, the red vertical line (at 30%) is to the left of the blue vertical line (at 35%), and so the red group, by definition, is being discriminated against. Under our model, the search rate for each race equals the area under the group's signal distribution to the right of the corresponding race-specific threshold, which in this case is 71% for the red group and 64% for the blue group. The hit rate (i.e., the search success rate) for each race equals the mean of the group's signal distribution conditional on

[4]Taste-based discrimination stands in contrast to *statistical discrimination* [Arrow, 1973, Phelps, 1972], in which officers might use a driver's race to improve their estimate that he is carrying contraband. Regardless of whether such information increases the efficiency of searches, officers are legally barred from using race to inform search decisions outside of circumscribed situations (e.g., when acting on specific and reliable suspect descriptions that include race among other factors). As is standard in the empirical literature on racial bias, we test only for taste-based discrimination.




(a) (b)

Figure 1: *Hypothetical signal distributions (solid curves) and search thresholds (dashed vertical lines) that illustrate how the benchmark and outcome tests can give misleading results.*[5] *Under the model of Section 2.1, the search rate for a given race is equal to the area under the signal distribution above the threshold, and the hit rate is the mean of the distribution conditional on being above the threshold. Situations (a) and (b) are observationally equivalent: in both cases, red drivers are searched more often than blue drivers (71% vs. 64%), while searches of red drivers recover contraband less often than searches of blue drivers (39% vs. 44%). Thus, the outcome and benchmark tests suggest that red drivers are being discriminated against in both (a) and (b). This is true in (a), because red drivers face a lower search threshold than blue drivers. However, blue drivers are subject to the lower threshold in (b), contradicting the results of the benchmark and outcome tests.*

being above the group's search threshold, 39% for the red group and 44% for the blue group. The red group is thus searched at a higher rate (71% vs. 64%), and when searched, found to have contraband at a lower rate (39% vs. 44%) than the blue group. Both the benchmark and outcome tests correctly indicate that the red group is being discriminated against.

Fig. 1b shows an alternative, hypothetical situation that is observationally equivalent to the one depicted in Fig. 1a, meaning that the search and hit rates of the red and blue groups are exactly the same in both settings. Accordingly, both the benchmark and outcome tests

[5]The depicted signal curves are beta distributions. The parameters for the red curves are: (a) $\alpha = 10.2$, $\beta = 18.8$; and (b) $\alpha = 10.8$, $\beta = 19.8$. The parameters for the blue curves are: (a) $\alpha = 10.3$, $\beta = 16.2$; and (b) $\alpha = 2.1$, $\beta = 4.1$.

again suggest that the red group is being discriminated against. In this case, however, blue drivers face a lower search threshold (25%) than red drivers (30%), and therefore the true discrimination present is exactly the opposite of the discrimination suggested by the outcome and benchmark tests.

What went wrong in this latter example? It is easier in the blue group to distinguish between innocent and guilty individuals, as indicated by the signal distribution of the blue group having higher variance. Consequently, those who are searched in the blue group are more likely to be guilty than those who are searched in the red group, resulting in a higher hit rate for the blue group, throwing off the outcome test. Similarly, it is easier in the blue group to identify low-risk individuals, who need not be searched, in turn lowering the overall search rate of the group and leading to spurious results from the benchmark test. In this example, the search and hit rates are poor proxies for the search thresholds.

The key point about Fig. 1b is that it is not a pathological case; to the contrary, it seems quite ordinary, and a variety of mechanisms could lead to this situation. If innocent minorities anticipate being discriminated against, they might display the same behavior—nervousness and evasiveness—as guilty individuals, making it harder to distinguish those who are innocent from those who are guilty. Alternatively, one group may simply be more experienced at concealing criminal activity, again making it harder to distinguish guilty from innocent. Given that one cannot rule out the possibility of such signal distributions arising in real-world examples (and indeed we later show that such cases do occur in practice), the benchmark and outcome tests are at best partial indicators of discrimination. We overcome this so-called problem of infra-marginality by directly estimating the search thresholds themselves, instead of simply considering the search and hit rates.

### 2.3 Inferring search thresholds

We now describe our *threshold test* for discrimination, which mitigates the most serious shortcomings of benchmark and outcome tests. For each stop $i$, we assume that we observe:

(1) the race of the driver, $r_i$; (2) the department of the officer, $d_i$; (3) whether the stop resulted in a search, indicated by $S_i \in \{0, 1\}$; and (4) whether the stop resulted in a "hit" (i.e., a successful search), indicated by $H_i \in \{0, 1\}$. Since a hit, by definition, can only occur if there was a search, $H_i \leq S_i$. Given a fixed set of stops annotated with the driver's race and the officer's department, we assume $S_i$ and $H_i$ are random outcomes resulting from a parametric process of search and discovery that formalizes the model of Section 2.1, described in detail below. We take a Bayesian approach to estimating the parameters of this process, and our primary goal is to infer race-specific search thresholds for each department. We interpret lower search thresholds for one group relative to another as evidence of discrimination. For example, if we were to find black drivers face a lower search threshold than white drivers, we would say blacks are being discriminated against.

Consider a single stop of a motorist of race $r$ conducted by an officer in department $d$. Upon stopping the driver, the officer assesses all the available evidence and concludes the driver has probability $p$ of possessing contraband. Even though officers may make these judgements deterministically, there is uncertainty in who is pulled over in any given stop. We thus model $p$ as a random draw from a race- and department-specific signal distribution, which captures heterogeneity across stopped drivers. This formulation side-steps the omitted variables problem of benchmark tests by allowing us to express information from all unobserved covariates as variation in the signal distribution. We can, in other words, think of the signal distribution as the marginal distribution over all unobserved variables.

We assume the signal $p$ is drawn from a beta distribution parameterized by its mean $\phi_{rd}$ (where $0 < \phi_{rd} < 1$) and total count parameter $\lambda_{rd}$ (where $\lambda_{rd} > 0$).[6] Thus, $\phi_{rd}$ is the overall probability that a stopped driver of race $r$ in department $d$ has contraband, and $\lambda_{rd}$ characterizes the heterogeneity across stopped drivers of that race in that department. We further assume $\phi_{rd}$ and $\lambda_{rd}$ are functions of parameters that depend only on a motorist's race ($\phi_r$ and $\lambda_r$), and those that depend only on an officer's department ($\phi_d$ and $\lambda_d$):

[6] In terms of the standard count parameters $\alpha$ and $\beta$ of the beta distribution, $\phi = \alpha/(\alpha+\beta)$ and $\lambda = \alpha+\beta$.

$$\phi_{rd} = \text{logit}^{-1}(\phi_r + \phi_d) \tag{1}$$

$$\lambda_{rd} = \exp(\lambda_r + \lambda_d) \tag{2}$$

where for identifiability we set $\phi_d$ and $\lambda_d$ equal to zero for the largest department.[7] Thus, if there are $D$ departments and $R$ races, the collection of $D \times R$ signal distributions is parameterized by $2(D+R-1)$ latent variables. Turning to the search thresholds, we assume that officers in a department apply the same threshold $t_{rd}$ to all drivers of a given race, but we allow these thresholds to vary by driver race and by department. Given the randomly drawn signal $p$, we assume officers deterministically decide to search a motorist if and only if $p$ exceeds $t_{rd}$; and if a search is conducted, we assume that contraband is found with probability $p$.

In summary, for each stop $i$, the data-generating process for $(S_i, H_i)$ proceeds in three steps, as follows:

1. Given the race $r_i$ of the driver and the department $d_i$ of the officer, the officer observes a signal $p_i \sim \text{beta}(\phi_{r_i d_i}, \lambda_{r_i d_i})$, where $\phi_{r_i d_i}$ and $\lambda_{r_i d_i}$ are defined according to Eqs. (1) and (2).

2. $S_i = 1$ (i.e., a search is conducted) if and only if $p_i \geq t_{r_i d_i}$.

3. If $S_i = 1$, then $H_i \sim \text{Bernoulli}(p_i)$; otherwise $H_i = 0$.

This generative process is parameterized by $\{\phi_r\}$, $\{\lambda_r\}$, $\{\phi_d\}$, $\{\lambda_d\}$ and $\{t_{rd}\}$. To complete the Bayesian model specification, we put weakly informative $\text{N}(0, 2)$ priors on $\phi_r$ and

[7] Without this constraint, the posterior distributions of the parameters would still be well-defined, but in that case the model would be identified by the priors rather than by the data. Moreover, the posterior distribution of $\phi_r$ would be highly correlated with that of $\phi_d$ (and likewise for $\lambda_r$ and $\lambda_d$), which makes inference computationally difficult.

$\lambda_r$, and hierarchical priors on $\phi_d$, $\lambda_d$, and $t_{rd}$. Specifically, we set

$$\phi_d \sim \mathrm{N}(\mu_\phi, \sigma_\phi)$$

where $\mu_\phi \sim \mathrm{N}(0, 2)$ and $\sigma_\phi \sim \mathrm{N}_+(0, 2)$ (i.e., $\sigma_\phi$ has a half-normal distribution). We similarly set

$$\lambda_d \sim \mathrm{N}(\mu_\lambda, \sigma_\lambda)$$

where $\mu_\lambda \sim \mathrm{N}(0, 2)$ and $\sigma_\lambda \sim \mathrm{N}_+(0, 2)$. Finally, for each race $r$, we put a logit-normal prior on every department's search threshold:

$$t_{rd} \sim \mathrm{logit}^{-1}\left(\mathrm{N}(\mu_{t_r}, \sigma_{t_r})\right)$$

where the race-specific hyperparameters $\mu_{t_r}$ and $\sigma_{t_r}$ have hyperpriors $\mu_{t_r} \sim \mathrm{N}(0, 2)$ and $\sigma_{t_r} \sim \mathrm{N}_+(0, 2)$. This hierarchical structure allows us to make reasonable inferences even for departments with a relatively small number of stops. We note that our results are robust to the exact specification of priors. Figure 2 shows this process represented as a graphical model [Jordan, 2004].

The number of observations $\mathcal{O} = \{(S_i, H_i)\}_i$ equals the number of stops—which could be in the millions—and so it can be computationally difficult to naively estimate the posterior distribution of the parameters. We can, however, dramatically improve the speed of inference by re-expressing the model in terms of the total number of searches ($S_{rd}$) and hits ($H_{rd}$) for drivers of each race in each department:

$$S_{rd} = \sum_{T_{rd}} S_i$$
$$H_{rd} = \sum_{T_{rd}} H_i$$

where $T_{rd} = \{i \,|\, r_i = r \text{ and } d_i = d\}$. Given that we fix the number of stops $n_{rd}$ of drivers

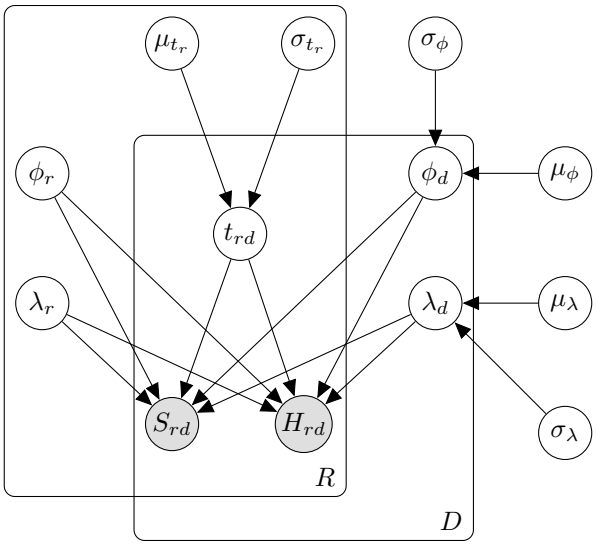



Figure 2: *Graphical representation of our generative model of traffic stops and searches. Observed search and hit rates are shaded, and unshaded nodes are latent variables that we infer from data.*

of race $r$ in department $d$, the quantities $\{S_{rd}\}$ and $\{H_{rd}\}$ are sufficient statistics for the process, and there are now only $2DR$ quantities to consider, regardless of the number of stops. This aggregation is akin to switching from Bernoulli to binomial response variables in a logistic regression model.

The distributions of $S_{rd}$ and $H_{rd}$ are readily computed for any parameter setting as follows. Let $I_x(\phi, \lambda)$ be the cumulative distribution function for the beta distribution. Then,

$$S_{rd} \sim \text{binomial}(p_{rd}, n_{rd})$$

where $p_{rd} = 1 - I_{t_{rd}}(\phi_{rd}, \lambda_{rd})$ is the probability that the signal is above the threshold. Similarly,

$$H_{rd} \sim \text{binomial}(q_{rd}, S_{rd})$$

where for $p \sim \text{beta}(\phi_{rd}, \lambda_{rd})$, $q_{rd} = \mathbb{E}[p \,|\, p \geq t_{rd}]$ is the likelihood of finding contraband when

a search is conducted. A straightforward calculation shows that

$$q_{rd} = \phi_{rd} \cdot \frac{1 - I_{t_{rd}}(\mu_{rd},\, \lambda_{rd} + 1)}{1 - I_{t_{rd}}(\phi_{rd},\, \lambda_{rd})} \tag{3}$$

where $\mu_{rd} = (\phi_{rd}\lambda_{rd} + 1)/(\lambda_{rd} + 1)$. With this reformulation, it is computationally tractable to run the threshold test on large datasets.

Having formally described our estimation strategy, we conclude by offering some additional intuition for our approach. Each race-department pair has three key parameters: the threshold $t_{rd}$ and two parameters ($\phi_{rd}$ and $\lambda_{rd}$) that define the beta signal distribution. Our model is thus in total governed by $3DR$ terms. However, we only effectively observe $2DR$ outcomes, the search and hit rates for each race-department pair. We overcome this information deficit in two ways. First, we restrict the form of the signal distributions according to Eqs. (1) and (2), representing the collection of $DR$ signal distributions with $2(D + R - 1)$ parameters. With this restriction, the process is now fully specified by $2(D + R - 1) + DR$ total terms, which is fewer than the $2DR$ observations when $R \geq 3$ and $D \geq 5$. Second, we regularize the parameters via hierarchical priors, which lets us efficiently pool information across races and departments. In this way, we leverage heterogeneity across jurisdictions to simultaneously infer signal distributions and thresholds for all race-department pairs.

## 3 An Empirical Analysis of North Carolina Traffic Stops

Using the approach described above, we now test for discrimination in vehicle searches by North Carolina police officers.

### 3.1 The data

We consider a comprehensive dataset of 9.5 million traffic stops conducted in North Carolina between January 2009 and December 2014 that was obtained via a public records request filed with the state. Several variables are recorded for each stop, including the race of the

driver (white, black, Hispanic, Asian, Native American, or "other"), the officer's department, the reason for the stop, whether a search was conducted, the type of search, the legal basis for that search, and whether contraband (e.g., drugs, alcohol, or weapons) was discovered during the search.[8] Due to lack of data, we exclude Native Americans from our analysis, who comprise 0.8% of all stops; we also exclude the 1.2% of stops where the driver's race was not recorded or was listed as "other".

We say that a stop resulted in a search if any of four listed types of searches (driver, passenger, vehicle, or property) were conducted. There are five legal justifications for searches recorded in our dataset: (1) the officer had probable cause that the driver possessed contraband; (2) the officer had reasonable suspicion—a weaker standard than probable cause—that the driver presented a danger, and searched the passenger compartment of the vehicle to secure any weapons that may be present (a "protective frisk"); (3) the driver voluntarily consented to the officer's request to search the vehicle; (4) the search was conducted after an arrest was made to look for evidence related to the alleged crime (a search "incident to arrest"); and (5) the officer was executing a search warrant. To gauge discrimination in search decisions, it is customary to only consider those situations in which searches are conducted at the discretion of officers. Probable cause and protective frisk searches are unambiguously discretionary. There is debate over whether consent searches should be considered, with Engel and Tillyer [2008] arguing that they are not fully discretionary since drivers can in principle decline, and Maclin [2008] claiming that consent searches give officers "discretion to conduct an open-ended search with virtually no limits." In our dataset, 22% of consent searches lead to the discovery of contraband, which suggests drivers do not regularly decline to be searched. We consequently consider consent searches to be discretionary and include them in our analysis. The data also suggest that searches incident to arrest are discretionary—as opposed to being conducted after every arrest—as only 42% of drivers that are arrested

[8]In our analysis, "Hispanic" includes anyone whose ethnicity was recorded as Hispanic, irrespective of their recorded race (e.g., it includes both white and black Hispanics). For consistency with the North Carolina data, we throughout use the term "Hispanic" when referring to this group, but we note that many people of Latin American descent prefer the term "Latino," and some object to either term.

are subsequently searched. It is impossible from our data to tell how often search warrants lead to actual searches, and hence whether such searches are discretionary, but since these cases comprise just 0.1% of all searches, we include them in our analysis to avoid unnecessary filtering. We thus ultimately do not exclude any searches based on the recorded legal justification.

There are 287 police departments in our dataset, including city departments, departments on college campuses, sheriffs' offices, and the North Carolina State Patrol. We find that state patrol officers conduct 47% of stops but carry out only 12% of all searches, and recover only 6% of all contraband found. State patrol officers search vastly less often than other officers, and the relatively few searches they do carry out are less successful. Given these qualitative differences, we exclude state patrol searches from our primary analysis. We further restrict to the 100 largest local police departments (by number of recorded stops), which in aggregate comprise 91% of all non-state-patrol stops. We are left with 4.5 million stops that we use for our primary analysis. Among this set of stops, 50% of drivers are white, 40% are black, 8.5% are Hispanic, and 1.5% are Asian. The overall search rate is 4.1%, and 29% of searches turn up contraband.

We note that our main results are robust to the specific set of stops we consider. For example, we find qualitatively similar patterns if we include only the clearly discretionary categories of probable cause and reasonable suspicion searches, or if we include state patrol stops.

### 3.2 Results from benchmark and outcome tests

We start with standard benchmark and outcome analyses of North Carolina traffic stops. Table 1 shows that the search rate for black drivers (5.4%) and Hispanic drivers (4.1%) is higher than for whites drivers (3.1%). Moreover, when searched, the rate of recovering contraband on blacks (29%) and Hispanics (19%) is lower than when searching whites (32%). Thus both the benchmark and outcome tests point to discrimination in search decisions

| Driver race | Stop count | Search rate | Hit rate |
|---|---|---|---|
| White | 2,227,214 | 3.1% | 32% |
| Black | 1,810,608 | 5.4% | 29% |
| Hispanic | 384,186 | 4.1% | 19% |
| Asian | 67,508 | 1.7% | 26% |

Table 1: *Summary of the traffic stops conducted by the 100 largest police departments North Carolina. Relative to white drivers, the benchmark test (comparing search rates) finds discrimination against blacks and Hispanics, while the outcome test (comparing hit rates) finds discrimination against blacks, Hispanics, and Asians.*

against blacks and Hispanics. The evidence for discrimination against Asians is mixed. Asian drivers are searched less often than whites (1.7% vs. 3.1%), but these searches also recover contraband at a lower rate (26% vs. 32%). Therefore, relative to whites, the outcome test finds discrimination against Asians but the benchmark test does not.

Adding resolution to these aggregate results, Fig. 3 compares search and hit rates for minorities and whites in each department. In the vast majority of cases, the top panel shows that blacks and Hispanics are searched at higher rates than whites. Benchmark analysis thus suggests a widespread pattern of discrimination against these groups. Asians, however, are consistently searched at lower rates than whites—indicating an absence of discrimination against Asians—in line with the aggregate results discussed above. The department-level outcome analysis is shown in the bottom panel of Fig. 3. In most departments, when Hispanics are searched, they are found to have contraband less often than searched whites, indicative of discrimination. However, hit rates for blacks and Asians are comparable to, or even higher than, hit rates for whites in a substantial fraction of cases, suggesting a lack of discrimination against these groups in many departments.

Table 2 summarizes the department-level evidence from the benchmark and outcome tests, where we count the number of departments in which the search rates for minorities are higher than for whites, and similarly, the number of times hit rates are lower.[9] Both

[9]The table indicates only the ordering of search and hit rates between minorities and whites, not whether the observed differences are statistically significant.

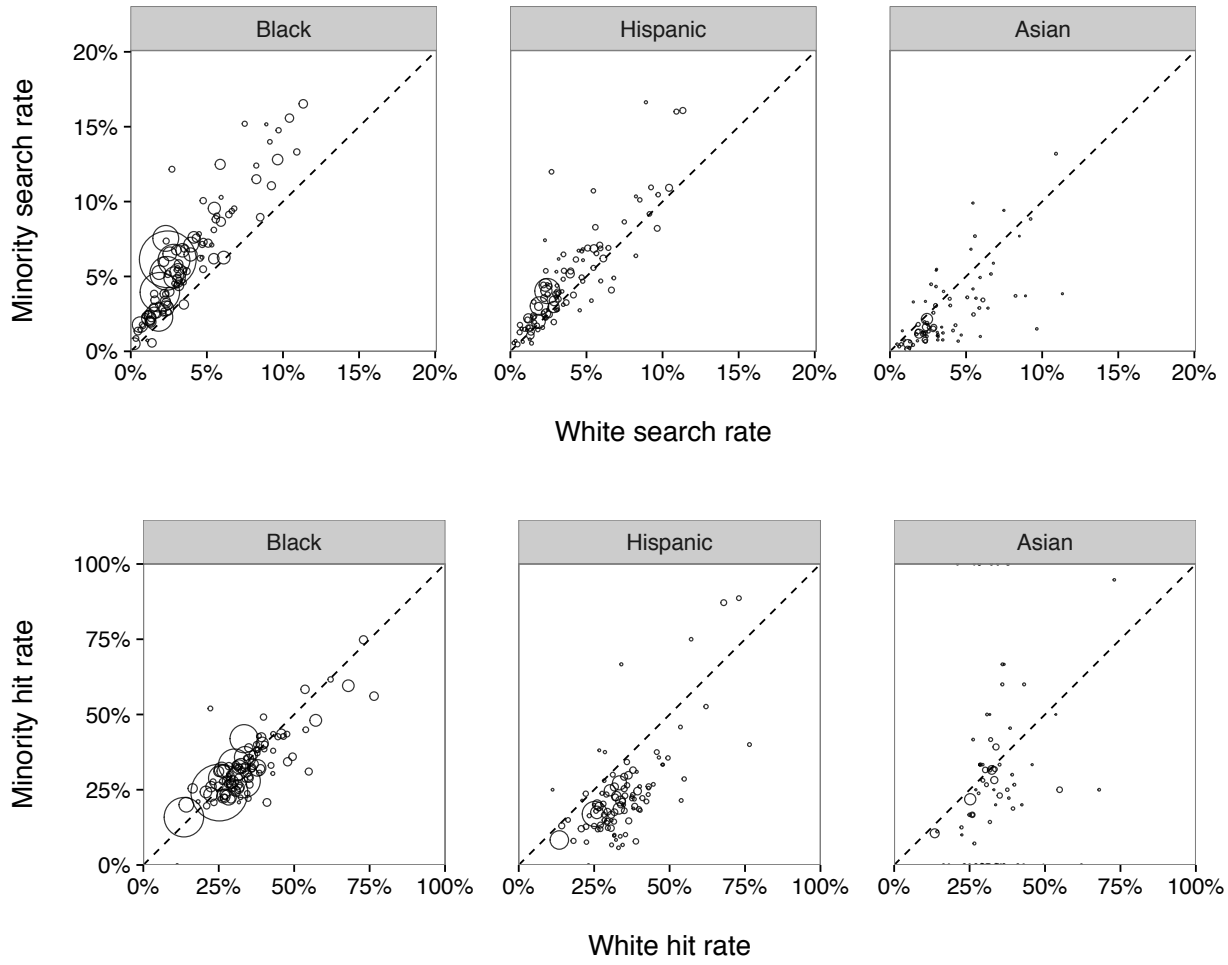



Figure 3: *Results of benchmark and outcome tests on a department-by-department basis. Each point in the top panel compares search rates of minority and white drivers for a single department. In the vast majority of departments, blacks and Hispanics are searched at higher rates than whites. In the bottom panel, each point compares the corresponding department-level hit rates. While Hispanics have consistently lower hit rates than whites, black and white hit rates are comparable in many departments; in particular, the outcome test thus suggests an absence of discrimination against blacks in many departments. Points in all the plots are scaled to the number of times the minority race was stopped by the department.*

tests suggest that discrimination exists against blacks and Hispanics in the majority of police departments, but they yield conflicting results in a significant number of cases. For example, both tests are indicative of discrimination against blacks in 57 of the top 100 departments, but offer ambiguous evidence in 42 departments; in only one department do both the outcome and benchmark tests point to an absence of discrimination against black drivers.

| | Benchmark test | Outcome test | |
|---|---|---|---|
| | | No discrimination | Discrimination |
| Black drivers | | | |
| | No discrimination | 1 | 2 |
| | Discrimination | 40 | 57 |
| Hispanic drivers | | | |
| | No discrimination | 2 | 24 |
| | Discrimination | 7 | 67 |
| Asian drivers | | | |
| | No discrimination | 20 | 41 |
| | Discrimination | 3 | 12 |

Table 2: *Outcome and benchmark analysis for the 100 largest police departments in North Carolina. Rows compare search rates between white and minority drivers (benchmark test), and columns compare hit rates (outcome test). The bottom right quadrant for each race group thus counts the number of departments in which both tests suggest discrimination against minorities.*

### 3.3 Results from the threshold test

We next use our threshold test to infer race- and department-specific standards of evidence for searching stopped drivers. Given the observed data, we estimate the posterior distribution of the search thresholds via Hamiltonian Monte Carlo (HMC) sampling [Duane et al., 1987, Neal, 1994], a form of Markov chain Monte Carlo sampling [Metropolis et al., 1953]. We specifically use the No-U-Turn sampler (NUTS) [Hoffman and Gelman, 2014] as implemented in Stan [Carpenter et al., 2016], an open-source modeling language for full Bayesian statistical inference. To assess convergence of the algorithm, we sampled five Markov chains in parallel and computed the potential scale reduction factor $\hat{R}$ [Gelman and Rubin, 1992]. We found that 2,500 warmup iterations and 2,500 sampling iterations per chain were sufficient for convergence, as indicated by $\hat{R}$ values less than 1.05 for all parameters, as well as by visual inspection of the trace plots.

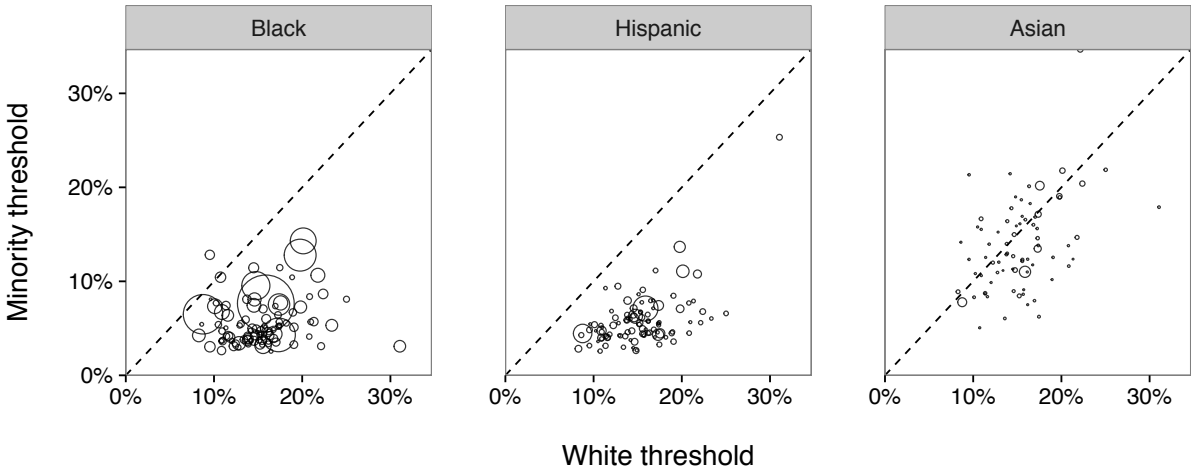



Figure 4: *Inferred search thresholds in the 100 largest North Carolina police departments. Each point compares the search thresholds applied to minority and white drivers in a department, where points are scaled to the number of times the minority race was stopped by the department. In nearly every department, black and Hispanic drivers are subject to lower search thresholds than whites, suggestive of discrimination.*

Figure 4 shows the posterior mean search thresholds for each race and department. Each point in the plot corresponds to a department, and compares the search threshold for whites (on the $x$-axis) to that for minorities (on the $y$-axis). In nearly every one of the 100 departments we consider, we find that black and Hispanic drivers are subject to a lower search threshold than whites, suggestive of discrimination against these groups. In many departments, we find the disparity is quite large, with the threshold for searching minorities 10 or even 20 percentage points lower than for searching whites. For Asians, in contrast, the inferred search thresholds are generally in line with those of whites, indicating an absence of discrimination against Asians in search decisions.

Figure 5 displays the average, state-wide inferred signal distributions and thresholds for whites, blacks, Hispanics, and Asians. These averages are computed by weighting the department-level results by the number of stops in the department. Specifically, the overall race-specific threshold $t_r$ is given by $(\sum_d t_{rd} \cdot n_d) \,/\, \sum_d n_d$, where $n_d$ is the number of stops in department $d$. Similarly, the aggregate signal distributions show the department-weighted distribution of subjective probabilities of possessing contraband. As is visually apparent, and

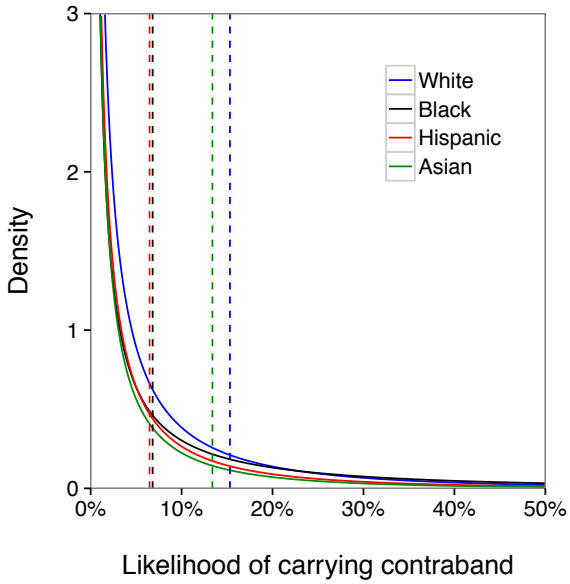



Figure 5: *Race-specific search thresholds and signal distributions, averaged over all departments and where we weight by the total number of stops conducted by the department. We find that black and Hispanic drivers face substantially lower search thresholds than white and Asian drivers.*

also summarized in Table 3, the thresholds for searching whites (15%) and Asians (13%) are considerably higher than the thresholds for searching blacks (7%) and Hispanics (6%). These thresholds are estimated to about ±2%, as indicated by the 95% credible intervals listed in Table 3.

To put these results in more concrete terms, we follow Antonovics and Knight [2009] and estimate how many fewer black and Hispanic drivers might have been searched had they been held to the same standard as whites. To do this, in each department we first use the estimated signal distributions and thresholds to compute the number of stopped black and Hispanic drivers with signals above the department's threshold for searching whites; we then compare this estimate to the actual number of black and Hispanic drivers searched. Across the six years in our dataset, this analysis suggests that over 30,000 searches of black drivers would not have been conducted had officers uniformly applied the white threshold, making up one-third of all searches of black drivers. We similarly find that more than half of the searches of Hispanic drivers would not have been conducted, totaling about 8,000 searches

| Driver race | Search threshold | 95% credible interval |
|---|---|---|
| White | 15% | (14%, 16%) |
| Black | 7% | (3%, 10%) |
| Hispanic | 6% | (5%, 8%) |
| Asian | 13% | (11%, 16%) |

Table 3: *Inferred search thresholds for stops conducted by the 100 largest police departments in North Carolina. For each race group, we report the average threshold across departments, weighting by the number of stops conducted by the department. We find black and Hispanic drivers face significantly lower search thresholds than white and Asian drivers.*

over the six years we consider. In this hypothetical world, the search rate of black drivers would drop to 3.5% (compared to the existing search rate of 5.4%), in line with the 3.1% search rate of whites. We estimate that the search rate of Hispanic drivers would drop even further, to 1.9%, below that of white drivers. It bears emphasis that this exercise is inherently speculative, and while useful for understanding the magnitude of the threshold differences we find, it is difficult to accurately predict what would have happened under such a counterfactual scenario.

Why is it that our threshold test shows substantial and consistent discrimination against blacks and Hispanics when benchmark and outcome analysis suggests a more ambiguous story? To understand this dissonance, we examine the specific case of the Raleigh Police Department, the second largest department in North Carolina by number of stops recorded in our dataset. Black drivers in Raleigh are searched at a much higher rate that whites (3.9% vs. 1.9%), but when searched, blacks are also found to have contraband at a higher rate (16% vs. 13%). The benchmark and outcome tests thus yield conflicting assessments of whether black drivers face discrimination. Figure 6 shows the inferred signal distributions and thresholds for white and black drivers in Raleigh, and sheds light on these seemingly contradictory results. Crucially, the signal distribution for black drivers has a much heavier right tail—for example, there is four times more mass above 20% than in the white distribution. This suggests that officers can more easily determine which black drivers are carrying contraband, which causes their searches of blacks to be more successful than their searches of whites. In spite of

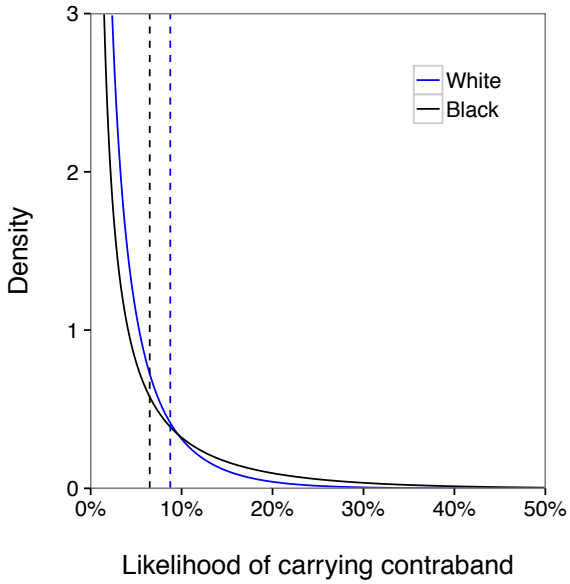



Figure 6: *The inferred search thresholds and signal distributions for black and white drivers stopped by the Raleigh Police Department. The problem of infra-marginality is in effect: the heavier tail of the black signal distribution means that searches of blacks have a higher hit rate despite black drivers being subject to a lower search threshold than whites. Hence, the outcome test concludes white drivers are being discriminated against, whereas the threshold test finds discrimination against black drivers.*

the higher hit rate for black drivers, we find that blacks still face a lower search threshold (6.4%) than whites (8.8%), suggesting discrimination against blacks. The Raleigh Police Department thus provides a real-world example of how the problem of infra-marginality can lead outcome tests to produce spurious results.

Despite the theoretical advantages of the threshold test, it is hard to know for sure whether we have accurately identified racial bias in Raleigh. We note, though, two reasons to believe the threshold test is capturing the underlying patterns of behavior. First, looking at Hispanic drivers in Raleigh, both the benchmark and outcome tests suggest they face discrimination. Hispanic drivers are searched more often than whites (3.0% vs. 1.9%), and are found to have contraband less often (11% vs. 13%). The threshold test likewise finds evidence of discrimination against Hispanics. The outcome test applied to black drivers is thus the odd one out: the benchmark, outcome, and threshold tests all point to discrimination against Hispanic drivers, and the benchmark and threshold tests suggest discrimination

against black drivers. Second, the outcome test indicates not only an absence of discrimination, but that white drivers face substantial bias; while not impossible, that conclusion is at odds with past empirical research on traffic stops [Epp et al., 2014].

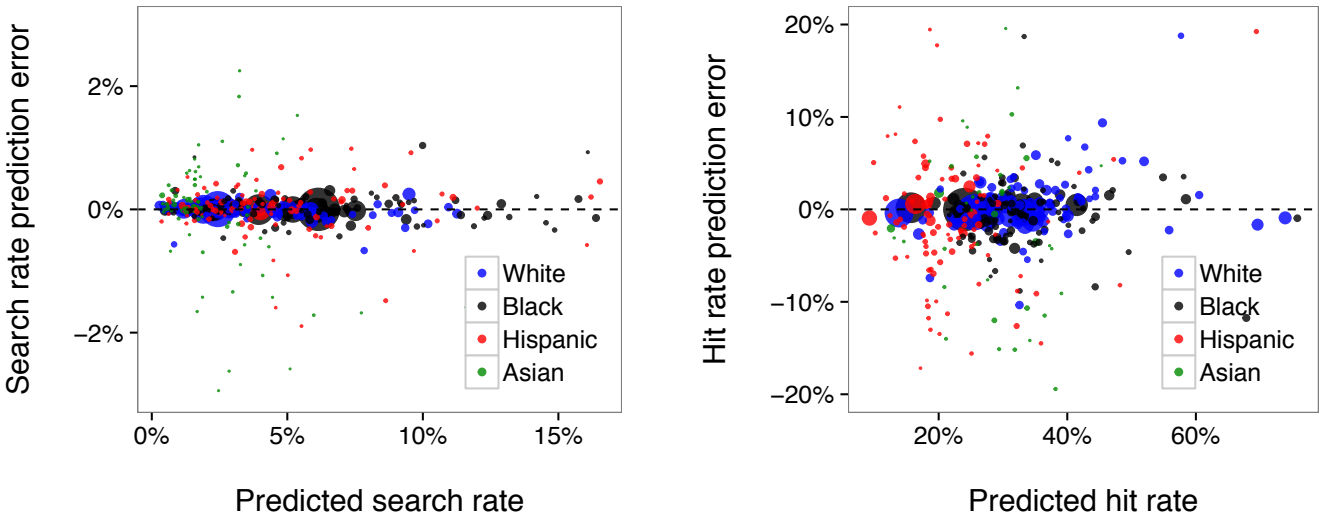
### 3.4 Model checks

We now evaluate in more detail how well our analytic approach explains the observed patterns in the North Carolina traffic stop data, and examine the robustness of our conclusions to violations of the model assumptions.

*Posterior predictive checks.* We begin by investigating the extent to which the fitted model yields race- and department-specific search and hit rates that are in line with the observed data. Specifically, for each department and race group, we compare the observed search and hit rates to their expected values under the assumed data-generating process with parameters drawn from the inferred posterior distribution. Such *posterior predictive checks* [Gelman et al., 1996, 2014] are a common approach for identifying and measuring systematic differences between a fitted Bayesian model and the data.

We compute the posterior predictive search and hit rates as follows. During model inference, our Markov chain Monte Carlo sampling procedure yields 2,500 draws from the joint posterior distribution of the parameters. For each parameter draw—consisting of $\{\phi_r^*\}$, $\{\lambda_r^*\}$, $\{\phi_d^*\}$, $\{\lambda_d^*\}$ and $\{t_{rd}^*\}$—we analytically compute the search and hit rates $s_{rd}^*$ and $h_{rd}^*$ for each race-department pair implied by the data-generating process with those parameters. Finally, we average these search and hit rates over all 2,500 posterior draws.

Figure 7 compares the model-predicted search and hit rates to the actual, observed values. Each point in the plot corresponds to a single race-department group, where groups are sized by number of stops. The fitted model recovers the observed search rates almost perfectly across races and departments. The fitted hit rates also agree with the data quite well, with the largest groups exhibiting almost no error. These posterior predictive checks thus indicate




(a) (b)

Figure 7: *Comparison of model-implied search and hit rates to the actual, observed values. Each point is a race-department pair, with points sized by number of stops. The plots show that the fitted model captures key features of the observed data. The root mean squared prediction error (weighted by stop count) is 0.1% for search rate and is 2.9% for hit rate.*

that the fitted model captures key features of the observed data.

*Miscalibration and heterogeneous search thresholds.* In describing our stylized model of officer behavior, we assumed that officers' subjective probability estimates are calibrated: when officers draw a signal $p$, the driver has contraband with probability $p$ (since the Bernoulli random variable $H_i$ has success probability $p$). But even the most knowledgeable experts are known to show patterns of miscalibration [Koehler et al., 2002]. For example, officers might exhibit *long-shot bias* [Snowberg and Wolfers, 2010], overestimating the likelihood of rare events. Officers might think they are searching those with at least a 20% chance of carrying contraband when in fact they are objectively searching those above a 10% threshold. Similarly, officers may consistently overestimate the probability that black and Hispanic drivers have contraband, perhaps because they overestimate the base rate of illegal activity in these groups. However, regardless of what officers may think they are doing, the thresholds we infer are based on the observed empirical frequencies of discovering contraband, and thus

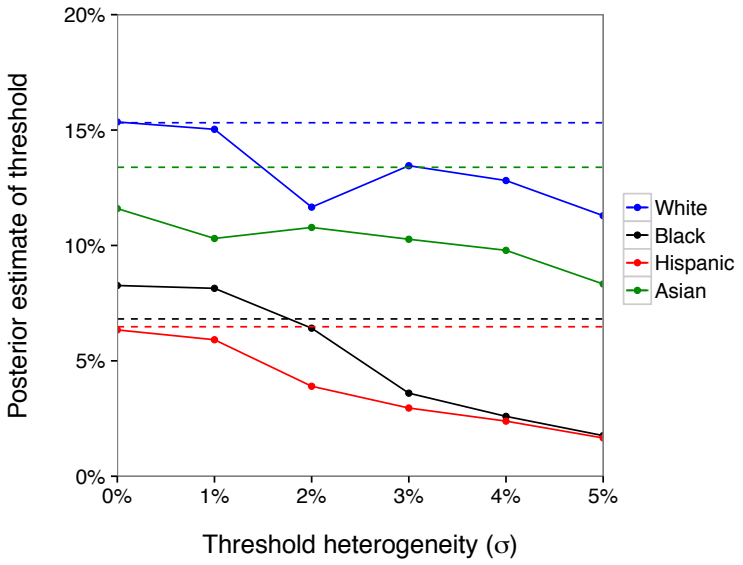



Figure 8: *Inferred race-specific search thresholds for synthetic data generated under a model in which thresholds randomly vary from one stop to the next. The dashed horizontal lines show the average of the thresholds used to generate the data. Model inferences are largely robust to stop-level heterogeneity in search thresholds.*

reflect the standards of evidence that they are actually applying.

A related point is that our behavioral model assumes there is a single search threshold for each race-department pair. In reality, officers within a department might apply different thresholds, and even the same officer might vary the threshold he or she applies from one day to the next. To investigate the robustness of our approach and results to such heterogeneity, we examine the stability of our inferences on synthetic datasets derived from a generative process with varying thresholds. Specifically, we start with the fitted model and then proceed in four steps. First, for each observed stop, we draw a signal $p$ from the inferred signal distribution for the department $d$ in which the stop occurred and the race $r$ of the motorist. Second, we set the stop-specific threshold to $T \sim \mathrm{N}(t_{rd}, \sigma)$, where $t_{rd}$ is the inferred threshold, and $\sigma$ is a parameter we set to control the degree of heterogeneity in the thresholds. Third, we assume a search occurs if and only if $p \geq T$, and if a search is conducted, we assume contraband is found with probability $p$. Finally, we use our modeling framework to infer new search thresholds $t'_{rd}$ for the synthetic dataset. Figure 8 plots the result of this exercise for $\sigma$

varying between 0 and 0.05. It shows that the inferences are relatively stable throughout this range, and in particular, that there is a persistent gap between whites and Asians compared to blacks and Hispanics. We note that a five percentage point change in the thresholds is quite large. For example, decreasing the search threshold of blacks by five points in each department would more than triple the overall state-wide search rate of blacks.

*Omitted variable bias.* As we discussed in Section 2, our approach is robust to unobserved heterogeneity that affects the signal, since we effectively marginalize over any omitted variables when estimating the signal distribution. However, we must still worry about systematic variation in the thresholds that is correlated with race. For example, if officers apply a lower search threshold at night, and black drivers are disproportionately likely to be stopped at night, then blacks would, on average, experience a lower search threshold than whites even in the absence of discrimination. Fortunately, as a matter of policy, only a limited number of factors may legitimately affect the search threshold, all of which are typically observable. As a point of comparison, there are a multitude of hard-to-quantify factors (such as a driver behaving nervously) that may, and likely do, affect the signal, but these should not affect the threshold.

Our model already explicitly accounts for search thresholds that vary by department. We now examine the robustness of our results when adjusting for possible variation across year, time-of-day, age, and gender of the driver.[10] Specifically, we disaggregate our primary dataset by year (and, separately, by time-of-day, by age, and by gender), and then independently run the threshold test on each component.[11] Figure 9 shows the results of this analysis, and illustrates two points. First, we find that the inferred thresholds do indeed vary across the different subsets of the data. Second, in every case, the thresholds for searching blacks and

[10]Gender, like race, is generally not considered a valid criterion for altering the search threshold, though for completeness we still examine its effects on our conclusions.

[11]4.7% of stops are recorded as occurring exactly at midnight, whereas 0.2% are listed at 11pm and 0.1% at 1am. It is likely that nearly all of the midnight stops are recorded incorrectly, and so we exclude these from our time-of-day analysis. Similarly, in a small fraction of stops (0.1%) the driver's age is recorded as either less than 16 or over 105 years old; we exclude these from our age analysis.

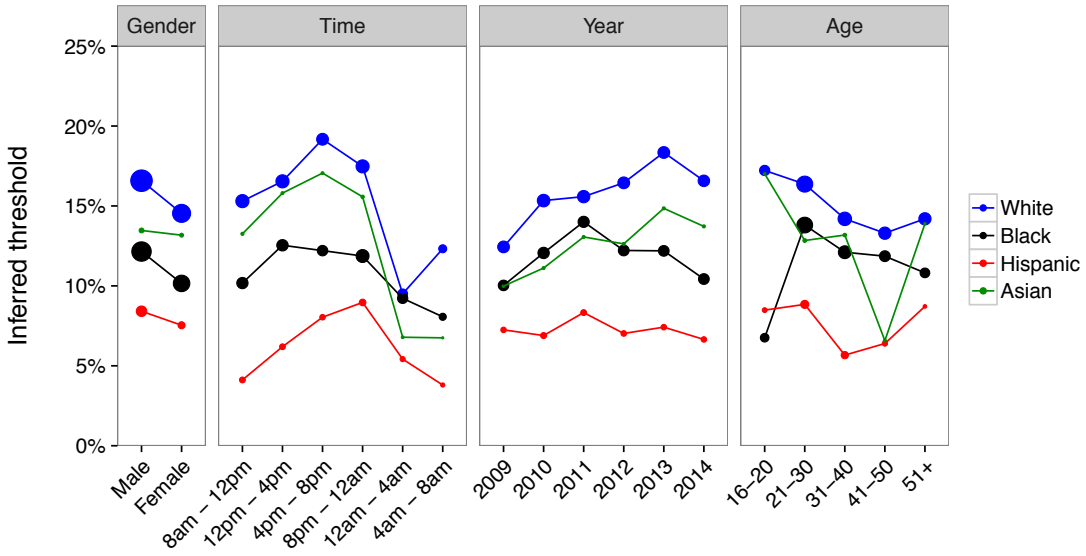



Figure 9: *Inferred search thresholds by race when the model is fit separately on various subsets of the data. Points indicate posterior means and are sized according to the number of stops in the subset. We consistently observe that blacks and Hispanics are subject to lower search thresholds than whites.*

Hispanics are lower than the threshold for searching whites, corroborating our main results.

Another potential confound is that searches of blacks and Hispanics may be prompted by concern for different types of illegal activity than searches of whites and Asians. For example, if officers have a lower threshold for searching drivers when they suspect possession of weapons rather than drugs, and black and Hispanic drivers are disproportionately likely to be suspected of carrying weapons, then we could again mistakenly infer discrimination where there is none. The suspected offense motivating a search is not recorded in our data, and so we cannot directly test for such an effect. However, the type of contraband (e.g., drugs, alcohol, weapons, or money) discovered in a search is recorded, which we treat as a proxy for the suspected crime. We find that the distribution of recovered contraband indeed differs across race groups, but the observed differences do not seem sufficiently large, absent discrimination, to account for the overall gap in search thresholds. In particular, drugs are discovered in the plurality of cases for all groups: 46% of searches of whites turn up drugs, compared to 49% for blacks, 34% for Hispanics, and 42% for Asians. And weapons are only

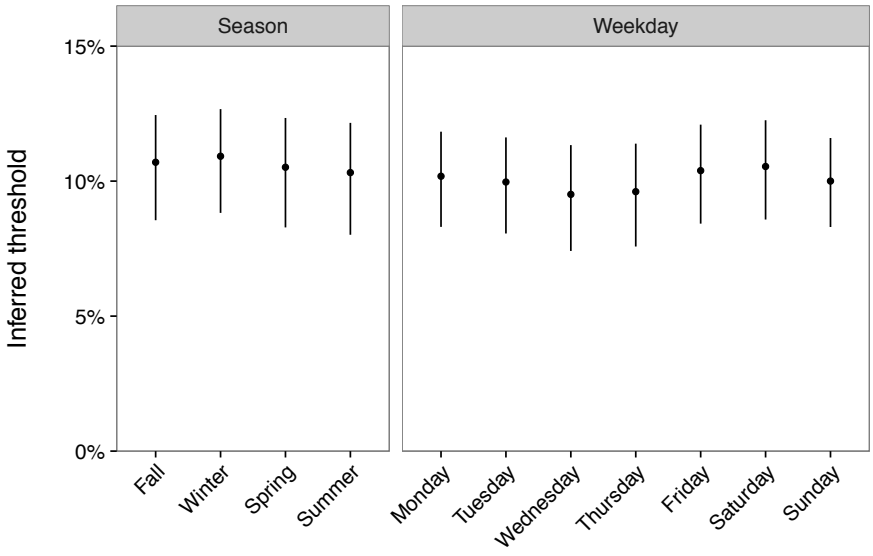



Figure 10: *Results of placebo tests, in which we examine how search thresholds vary by season and day-of-week. Points show the posterior means, and the bars indicate 95% credible intervals. The threshold test accurately suggests a lack of "discrimination" in these cases.*

infrequently recovered: they are found 6% of the time for whites, 9% for blacks, 7% for Hispanics, and 11% for Asians.

*Placebo tests.* Finally, we conduct two *placebo tests*, where we rerun our threshold test with race replaced by day-of-week, and separately, with race replaced by season. The hope is that the threshold test accurately captures a lack of "discrimination" based on these factors. Figure 10 shows that the model indeed finds that the threshold for searching individuals is relatively stable by day-of-week, with largely overlapping credible intervals. We similarly find only small differences in the inferred seasonal thresholds. We note that some variation is expected, as officers might legitimately apply slightly different search standards throughout the week or year.

## 4 Discussion and Conclusion

There is increasing public concern that police actions are racially biased, yet theoretical limitations with the two most widely used tests for discrimination—the benchmark and

outcome tests—have stymied rigorous investigation of the issue. For example, in assessing claims of bias in airport searches, Judge Easterbrook, of the U.S. Court of Appeals for the Seventh Circuit, noted that the problem of infra-marginality raised doubts that higher search rates and lower hit rates for black women were indicative of discrimination [Anderson v. Cornejo, 2004].

Addressing these challenges, we have developed an alternative statistical approach to detecting and quantifying bias that builds on the strengths of the benchmark and outcome tests and mitigates the shortcomings of both. On a dataset of 4.5 million motor vehicle stops in North Carolina, our threshold test suggests that officers apply a double standard when deciding whom to search, with black and Hispanic drivers searched on the basis of less evidence than whites and Asians. We consistently observe this pattern of behavior across the 100 largest police departments in the state. By specifically examining the Raleigh Police Department, we further find that the problem of infra-marginality is more than a theoretical possibility, and likely caused the outcome test to mistakenly conclude that Raleigh officers discriminated against white drivers (our analysis suggests it is more likely that black motorists were discriminated against).

Our empirical results appear robust to reasonable violations of the model assumptions, and we have attempted to rule out the most obvious legitimate reasons for which thresholds might vary, including search policies that differ across department, year, or time of day. We cannot, however, definitively conclude that the disparities we see stem from racial bias. For example, officers might instead be applying lower search thresholds to those from lower socio-economic backgrounds, a demographic that is disproportionately black and Hispanic. At the very least, though, our results indicate that there are concerning irregularities in the search decisions we study.

It bears emphasis that lower search thresholds do not imply racial animus, and it is likely that the effects we observe are at least partially attributable to implicit bias [Eberhardt et al., 2004]. For example, officers might mistakenly believe that blacks who behave

nervously are more likely to have contraband than similarly situated whites. This is a particularly important legal distinction, as claims of discrimination brought under the Equal Protection Clause of the Fourteenth Amendment typically require evidence of discriminatory intent [Goel et al., 2016a]. Moreover, depending on the underlying mechanism, attempts to remedy such disparities would likely take different forms.

Aside from police searches, our threshold test could be applied to detect discrimination in a variety of settings where benchmark and outcome analysis is the status quo, including mortgage lending, hiring, and publication decisions. Looking forward, we hope our methodological approach and substantive results spur further investigation into the theoretical properties of statistical tests of discrimination, as well as their practical application.

## References


Geoffrey P Alpert, Michael R Smith, and Roger G Dunham. Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research. *Justice Research and Policy*, 6(1):43–69, 2004.

Anderson v. Cornejo. 355 F.3d 1021 (7th Cir.), 2004.

Kate Antonovics and Brian G Knight. A new look at racial profiling: Evidence from the Boston police department. *The Review of Economics and Statistics*, 91(1):163–177, 2009.

Shamena Anwar and Hanming Fang. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *The American Economic Review*, 2006.

Kenneth Arrow. The theory of discrimination. In *Discrimination in Labor Markets*. Princeton University Press, 1973.

Ian Ayres. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.

Gary S Becker. *The economics of discrimination*. University of Chicago Press, 1957.

Gary S Becker. Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, pages 385–409, 1993.

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 2016.

James H Carr, Isaac F Megbolugbe, et al. *The Federal Reserve Bank of Boston study on mortgage lending revisited*. Fannie Mae Office of Housing Policy Research, 1993.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

Jennifer L Eberhardt, Phillip Atiba Goff, Valerie J Purdie, and Paul G Davies. Seeing black: race, crime, and visual processing. *Journal of Personality and Social Psychology*, 87(6): 876, 2004.

Robin S Engel and Jennifer M Calnon. Comparing benchmark methodologies for police-citizen contacts: Traffic stop data collection for the Pennsylvania State Police. *Police Quarterly*, 7(1):97–125, 2004.

Robin S Engel and Rob Tillyer. Searching for equilibrium: The tenuous nature of the outcome test. *Justice Quarterly*, 25(1):54–71, 2008.

Charles R Epp, Steven Maynard-Moody, and Donald P Haider-Markel. *Pulled over: How police stops define race and citizenship*. University of Chicago Press, 2014.

George C Galster. The facts of lending discrimination cannot be argued away by examining default rates. *Housing Policy Debate*, 4(1):141–146, 1993.

Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472, 1992.

Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760, 1996.

Andrew Gelman, Jeffrey Fagan, and Alex Kiss. An analysis of the New York City Police Department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479), 2007.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Taylor & Francis, 2nd edition, 2014.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Sharad Goel, Maya Perelman, Ravi Shroff, and David Sklansky. Combatting police discrimination in the age of big data. *New Criminal Law Review*, 2016a. Forthcoming.

Sharad Goel, Justin M Rao, and Ravi Shroff. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*, 2016b.

Jeffrey Grogger and Greg Ridgeway. Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, 101(475), September 2006.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15 (Apr):1593–1623, 2014.

Michael I Jordan. Graphical models. *Statistical Science*, pages 140–155, 2004.

John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1), 2001.

Derek J Koehler, Lyle Brenner, and Dale Griffin. The calibration of expert judgment: Heuristics and biases beyond the laboratory. In *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press, 2002.

James E Lange, Kenneth O Blackman, and Mark B Johnson. *Speed violation survey of the New Jersey Turnpike: Final report*. Public Services Research Institute, 2001.

Lynn Langton and Matthew Durose. Police behavior during traffic and street stops, 2011. Technical report, U.S. Department of Justice, 2013.

Tracey Maclin. Good and bad news about consent searches in the Supreme Court. *McGeorge L. Rev.*, 39:27, 2008.

Elizabeth H McConnell and Amie R Scheidegger. Race and speeding citations: Comparing speeding citations issued by air traffic officers with those issued by ground traffic officers. In *Annual meeting of the Academy of Criminal Justice Sciences, Washington, DC*, 2001.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Radford M Neal. An improved acceptance procedure for the hybrid Monte Carlo algorithm. *Journal of Computational Physics*, 111(1):194–203, 1994.

Edmund S Phelps. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661, 1972.

Greg Ridgeway. Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of Quantitative Criminology*, 22(1):1–29, 2006.

Greg Ridgeway and John M MacDonald. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104(486):661–668, 2009.

Erik Snowberg and Justin Wolfers. Explaining the favorite-longshot bias: Is it risk-love or misperceptions? *Journal of Political Economy*, 118(4):723–746, 2010.

Samuel Walker. Internal benchmarking for traffic stop data: An early intervention system approach. Technical report, Police Executive Research Forum, 2003.