The threshold test: Testing for racial bias in vehicle searches by police

Claims of biased decision making occur in many disciplines, including lending, hiring, and policing. However, it is typically difficult to rigorously assess these claims. This is in large part because of well-known problems with the two most common statistical tests for discrimination: benchmarking and outcome tests. We develop a new statistical test of discrimination — the "threshold test", that mitigates some of the problems affecting existing tests. Our test uses a hierarchical latent Bayesian model to jointly estimate the decision thresholds applied by decision makers, and the risk distributions for different populations. We apply our test to a dataset of 4.5 million police stops in North Carolina, where we find that officers apply lower standards of evidence when searching minorities than when searching whites.

Introduction

More than 20 million Americans are stopped each year for traffic violations, making this one of the most common ways in which the public interacts with the police (Langton and Durose, 2013). During routine traffic stops, officers have latitude to search both the driver and vehicle for drugs, weapons, and other contraband when they suspect more serious criminal activity. As a result, there have been concerns that racial bias plays a role in officers' decisions. Attempts to measure any such bias have so far been hindered by well-known problems with the two most common statistical tests for discrimination.

In the first test, termed benchmarking, one compares the rate at which whites and minorities are searched. If minority drivers are searched more often than whites, that may be the result of bias against minorities. However, if minorities in reality are more likely to carry contraband or engage in illegal activities, then such disparities in search rates may simply reflect routine police work. 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.

Addressing this shortcoming of benchmarking, Becker (1993, 1957) proposed the outcome test, which is based not on the rate at which search decisions are made, but on the success rate of those searches. Becker argued that even if different groups vary in their propensity to carry contraband, discrimination could be detected if searches of minorities yield contraband less often than searches of whites. Such a finding suggests that officers apply a lower standard of evidence when searching minorities, discriminating against them.

Outcome tests, however, are imperfect barometers of bias. 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 illustrates a subtle failure of outcome tests known as the problem of infra-marginality (Ayres, 2002), a phenomenon we discuss in detail below.

To address the shortcomings of the outcome test, we built on Becker's ideas to develop a more robust statistical measure of discrimination: the threshold test. The threshold test combines information on both search rates and hit rates, and lets us directly infer the standard of evidence officers require before carrying out a search. If this standard differs for drivers of different races, police officers are discriminating in their search decisions.

The Data

We consider a dataset of 4.5 million stops conducted by the 100 largest local police departments (by number of recorded stops) in North Carolina. The data was obtained via a public records request filed with the state and includes the complete set of stops were conducted between January 2009 and December 2014 . Several variables are recorded for each stop, including the race of the driver (white, black, Hispanic), the officer's department, whether a search was conducted, and whether contraband (e.g., drugs, alcohol, or weapons) was discovered during the search. Among this set of stops, 50% of drivers are white, 40% are black, 8.5% are Hispanic. The overall search rate is 4.1%, and 29% of searches turn up contraband.

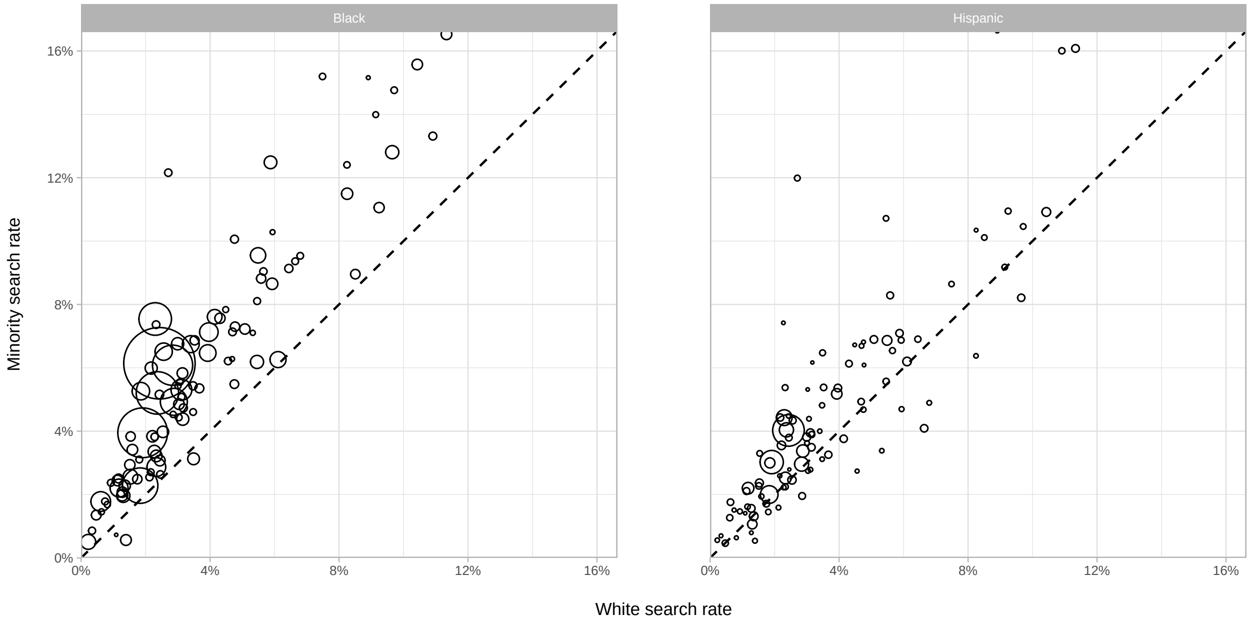
Existing tests of discrimination

The two most common statistical tests for discrimination, termed "benchmarking" and "outcome" suffer from well-known limitations. In this section we apply the tests and discuss their deficiencies.

Benchmarking test

A large body of work makes use of benchmark tests to assess racial bias in police stops and searches. The benchmark test compares, for each group of interest, the rate at which a decision is made relative to some benchmark. For example, bias in search decisions could be investigated by determining whether, conditional on being pulled over, minority drivers are searched more often than white drivers. Higher search rates for minorities would suggest discrimination. In the figure below, we see that this is the case for the largest 100 departments in North Carolina.

```r
plot_benchmark_test = function(obs) {
  races = as.character(levels(obs$driver_race))
  mx = max(obs$search_rate)
  df = obs %>%
    filter(driver_race == 'White') %>%
    right_join(obs %>% filter(driver_race != 'White'), by = 'police_department')

  ggplot(df) + geom_point(aes(x=search_rate.x, y=search_rate.y, size=num_stops.y), shape =1 ) +
    geom_abline(slope=1, intercept=0, linetype='dashed') +
    scale_y_continuous('Minority search rate\n', limits=c(0, mx), labels=percent, expand = c(0, 0)) +
    scale_x_continuous('\nWhite search rate', limits=c(0, mx), labels=percent, expand = c(0, 0)) +
    scale_size_area(max_size=15) +
    theme(legend.title = element_blank(),
          legend.background = element_rect(fill = 'transparent'),
          panel.margin.x=unit(1.5, "cm")) +
    guides(size=FALSE) + facet_grid( .~driver_race.y)
}

plot_benchmark_test(stops)
```



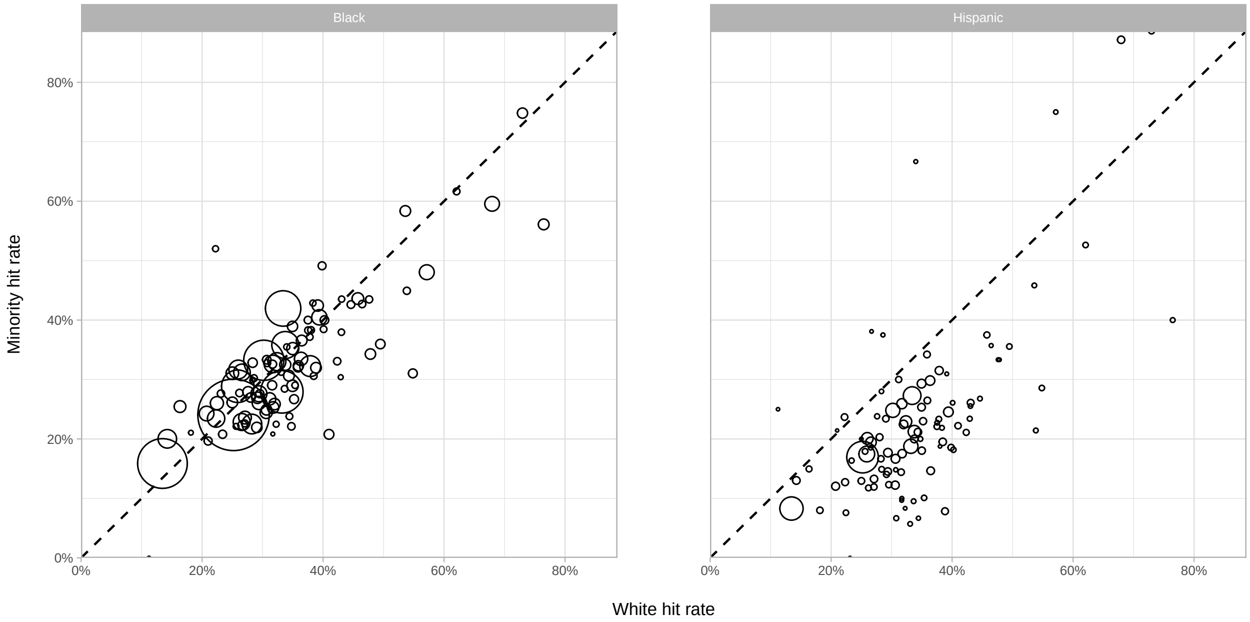
These disparities, however, are not necessarily the product of discrimination. Minority drivers might carry contraband at higher rates than whites, resulting in a commensurate increase in searches of minorities. More generally, we do not observe the factors that caused an officer to conduct a search. There may be factors indicative of guilt that are correlated with race, which officers use to determine who to search. If the benchmark does not include these variables (e.g. because they are unavailable to researchers), benchmarking tests will be unable to determine whether it was the driver's race that affected the search decision, or these omitted variables that are correlated with race.

Outcome test

Addressing this shortcoming of benchmarking tests, Becker (1993, 1957) proposed the outcome test, which is based not on the rate at which decisions are made, but on the success rate of those decisions. In other words, the outcome test 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, absent discrimination searched minorities should still be found to have contraband at the same rate as searched whites. If searches of minorities are less often successful than searches of whites, it suggests that officers may be applying a double standard, searching minorities on the basis of less evidence.

```
plot_outcome_test = function(obs) {
  races = as.character(levels(obs$driver_race))
  mx = max(obs$hit_rate)
  df = obs %>% filter(driver_race == 'White') %>%
    right_join(obs %>% filter(driver_race != 'White'), by='police_department')

  ggplot(df) + geom_point(aes(x=hit_rate.x, y=hit_rate.y, size=num_stops.y), shape = 1) +
    geom_abline(slope=1, intercept=0, linetype='dashed') +
    scale_y_continuous('Minority hit rate\n', limits=c(0, mx), labels=percent, expand = c(0, 0)) +
    scale_x_continuous('\nWhite hit rate', limits=c(0, mx), labels=percent, expand = c(0, 0)) +
    scale_size_area(max_size=15) +
    theme(legend.title = element_blank(),
          legend.background = element_rect(fill = 'transparent'),
          panel.margin.x=unit(1.5, "cm")) +
    guides(size=FALSE, color = FALSE) + facet_grid( .~driver_race.y)
}
```

```
plot_outcome_test(stops)
```



The aggregate results of the benchmark and outcome tests across all departments are given below:

```
stops %>%
  group_by(driver_race) %>%
  summarise(total_stops = sum(num_stops),
            total_searches = sum(num_searches),
            total_hits = sum(num_hits),
            search_rate = round(total_searches / total_stops * 100, 1),
            hit_rate = round(total_hits/ total_searches * 100, 1)) %>%
  setNames(c('Driver Race', 'Total Stops', 'Total Searches',
             'Total Hits', 'Search Rate', 'Hit Rate'))
```

```
## # A tibble: 3 x 6
##   `Driver Race` `Total Stops` `Total Searches` `Total Hits` `Search Rate`
##          <fctr>         <int>            <int>        <int>         <dbl>
## 1         White       2227214            68199        22114           3.1
## 2         Black       1810608            98538        28137           5.4
## 3      Hispanic        384186            15727         3056           4.1
## # ... with 1 more variables: `Hit Rate` <dbl>
```

The problem of infra-marginality

The outcome test is intuitively appealing, however, it is also an imperfect barometer of bias. In particular, it is known to suffer from the problem of infra-marginality (Anwar and Fang, 2006; Ayres, 2002): even absent discrimination, the search success rates for minority and white drivers might differ if the two groups have different risk distributions. Thus, at least in theory, outcome tests can fail to accurately measure discrimination.

Stylized model of officer behavior

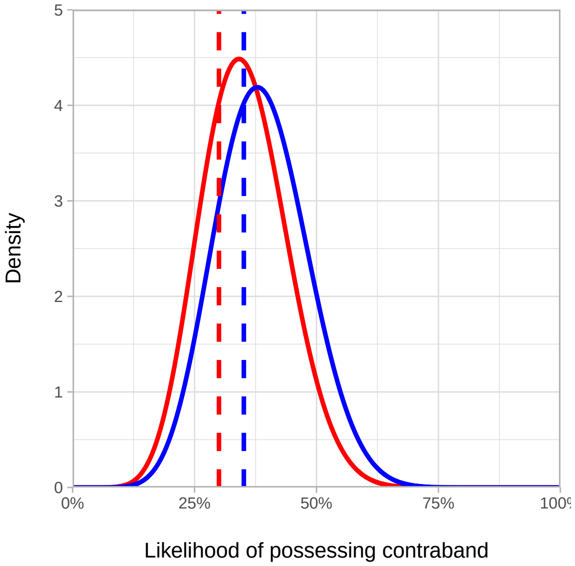
To introduce our new test we develop the following stylized model of officer behaviour during traffic stops. We assume that the officer observes noisy but informative signals about whether or not a driver carries contraband. These might include age, gender, race, contextual information such as location, time of day, any odours, and behavioural cues such as signs of evasiveness or nervousness, etc. The officer then uses this information to estimate the probability that the driver is carrying contraband. We can think of this estimate as a random draw from a *risk distribution*, that summarizes the risks of all drivers in a population. If the estimated probability exceeds some threshold, the officer conducts a search, and the chance of finding contraband is equal to the probability (i.e. the officer's estimates are calibrated).

Example illustrating problem of infra-marginality

We can use our model of police behavior to better illustrate the problem of infra-marginality. In the two figures below, consider two hypothetical risk distributions (solid curves) and search thresholds (dashed vertical lines) that illustrate how the benchmark and outcome tests can give misleading results. Under the model described above, the search rate for a given group is equal to the area under the risk 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 cases. This is true in the top plot, because red drivers face a lower search threshold than blue drivers. However, blue drivers are subject to the lower threshold (bottom plot), contradicting the results of the benchmark and outcome tests.

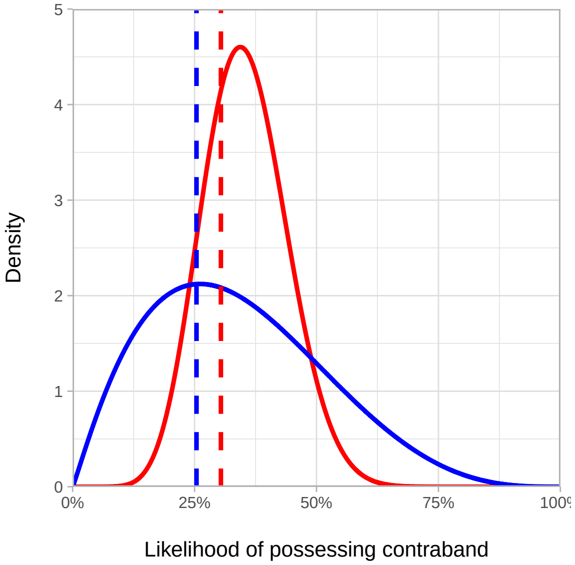
```r
beta_conditional_mean <- function(x, a, b) {
  return ((1-pbeta(x, a+1, b)) / (1-pbeta(x, a, b)) * a / (a+b))
}

plot_example = function(t, p, l1, l2, t2 = t, p2=p, ylim = 3) {
  x = seq(0.0001, 0.9999, 0.0001)
  print(round(c(
    1-pbeta(t, l1*p, l1*(1-p)),
    beta_conditional_mean(t, l1*p, l1*(1-p))), digits = 2))
  print(round(c(
    1-pbeta(t2, l2*p2, l2*(1-p2)),
    beta_conditional_mean(t2, l2*p2, l2*(1-p2))), digits = 2))

  y1 = dbeta(x, l1*p, l1*(1-p))
  y2 = dbeta(x, l2*p2, l2*(1-p2))
  x = c(0,x,1)
  y1 = c(0,y1,0)
  y2 = c(0,y2,0)

  plt = ggplot() + geom_line(aes(x=x, y=y1, color = '1'), size = 0.8) +
    geom_line(aes(x=x, y=y2, color = '2'), size = 0.8) +
    scale_y_continuous('Density\n', expand = c(0,0), limits = c(0, ylim)) +
    scale_x_continuous('\nLikelihood of possessing contraband', labels=percent, expand = c(0, 0)) +
    guides(color=FALSE)
  if (t == t2) {
    plt = plt + geom_vline(xintercept = t, linetype = 'dashed', size = 0.8)
  } else {
    plt = plt + geom_vline(aes(color = '1', xintercept = t), linetype = 'dashed', size = 0.8) +
      geom_vline(aes(color = '2', xintercept = t2), linetype = 'dashed', size = 0.8)
  }
  plt + scale_color_manual(values = c('red', 'blue'))
}

# discrimination against red
plot_example(0.3, 0.352, 28.98, t2=0.351, p2=0.3878, l2=26.48, ylim = 5)
```

```
## [1] 0.71 0.39
## [1] 0.64 0.44
```



```
# discrimination against blue
plot_example(0.304, 0.354, 30.59, t2=0.254, p2=0.3385, l2=6.19, ylim = 5)
```

```
## [1] 0.71 0.39
## [1] 0.64 0.44
```



The threshold test model

We now describe the full Bayesian model for a single stop i. The probability of carrying contraband x_i is drawn from a risk distribution which depends on the race of the driver and the department of the stop. While any distribution over $(0, 1)$ would be reasonable, for computational reasons we use the discriminant distribution (Pierson *et al.*, 2017):

$x_i \sim \mathrm{disc}(\phi_{r_i d_i}, \delta_{r_i d_i})$.

The parameters of the risk distribution are allowed to vary with race and department as follows. ϕ_{rd} is the mean of the risk distribution (the overall propensity to carry contraband), and is defined as:

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

δ_{rd} measures how easy it is to distinguish between innocent and guilty drivers, and is defined as:

$\delta_{rd} = \exp(\delta_r + \delta_d)$.

Given their estimate of probability of guilt x_i, the officer deterministically conducts search if it exceeds the race- and department-specific search threshold $t_{r_i d_i}$:

$S_i = 1$ iff $x_i > t_{r_i d_i}$.

If a search is conducted (i.e. if $S_i = 1$), contraband is found with probability x_i:

$H_i \sim Bernoulli(x_i)$

Under this model, if officers have a lower threshold for searching blacks than whites in some department d (i.e., $t_{black,d} < t_{white,d}$), we would say that black drivers are being discriminated against in that department.

In the model above the number of observations equals the number of stops. Since we have 4.5 million stops, it is intractable to naively estimate the posterior distribution of the parameters. So, instead of modeling the stop-level outcomes as Bernoulli varibables, we switch to an equivalent binomial parametrization, re-expressing the model in terms of the total number of stops (N_{rd}), searches (S_{rd}), and hits (H_{rd}) for drivers of each race in each department:

$S_{rd} = \sum_{i=1}^{N} S_i$

$H_{rd} = \sum_{i=1}^{N} H_i$

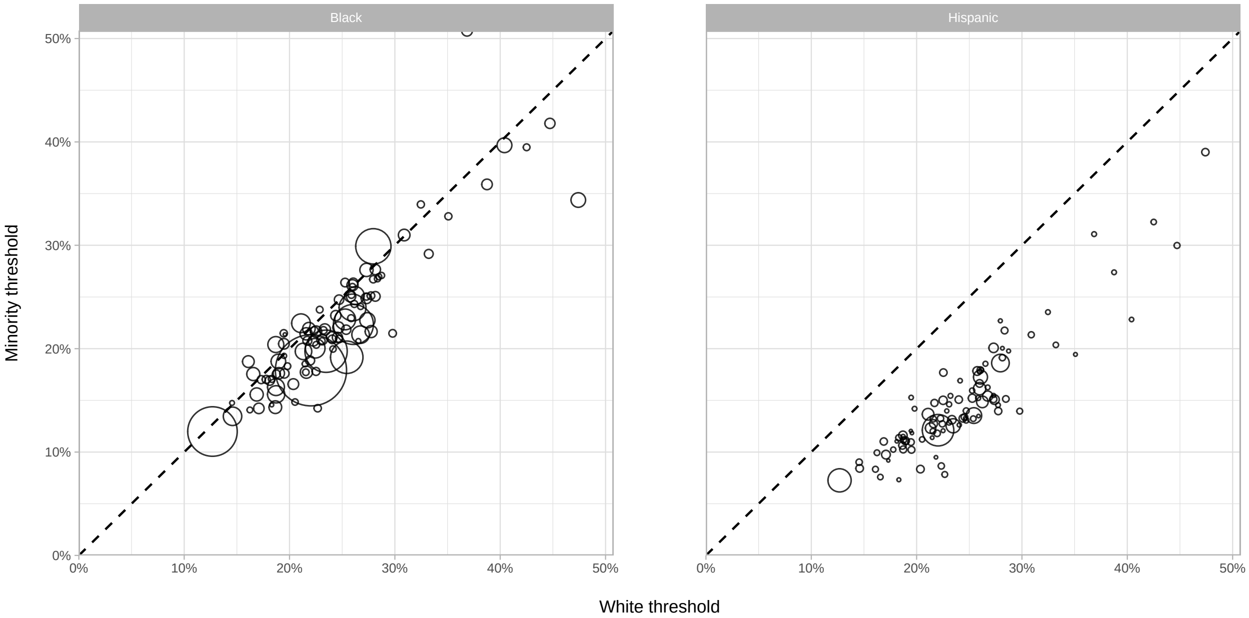
The search thresholds t_{rd} and risk distribution parameters (ϕ_r, ϕ_d, δ_r, δ_d) are latent variables that are inferred using Stan. We sample five Markov chains of 5,000 iterations (with the first half for warmup) in parallel, and found this was sufficient for convergence, as indicated by $\hat{R}$ values less than 1.05 for all parameters.

```
data {
  int<lower=1> N; // number of observations
  int<lower=1> R; // number of suspect races
  int<lower=1> D; // number of counties

  int<lower=1,upper=R> r[N]; // race of suspect
  int<lower=1,upper=D> d[N]; // county where stop occurred

  int<lower=1> n[N]; // # of stops
    int<lower=0> s[N]; // # of searches
    int<lower=0> h[N]; // # of successful searches (hits)
}

parameters {
  // hyperparameters
  real<lower=0> sigma_t; #standard deviation for the normal the thresholds are drawn from.
  // search thresholds
  vector[R] t_r;
  vector[N] t_i_raw;

  // parameters for risk distribution
  vector[R] phi_r;
  vector[D-1] phi_d_raw;
  real mu_phi;

  vector[R] delta_r;
  vector[D-1] delta_d_raw;
  real mu_delta;
}

transformed parameters {
  vector[D] phi_d;
  vector[D] delta_d;
  vector[N] phi;
  vector[N] delta;
  vector[N] t_i;
  vector<lower=0, upper=1>[N] search_rate;
  vector<lower=0, upper=1>[N] hit_rate;
  real successful_search_rate;
  real unsuccessful_search_rate;

  phi_d[1]      = 0;
  phi_d[2:D]    = phi_d_raw;
  delta_d[1]   = 0;
  delta_d[2:D] = delta_d_raw;

  t_i = t_r[r] + t_i_raw * sigma_t;

  for (i in 1:N) {

    // phi is the fraction of people of race r, d who are guilty (ie, carrying contraband)
    phi[i]    = inv_logit(phi_r[r[i]] + phi_d[d[i]]);

    // mu is the center of the guilty distribution.
    delta[i] = exp(delta_r[r[i]] + delta_d[d[i]]);

    successful_search_rate = phi[i] * (1 - normal_cdf(t_i[i], delta[i], 1));
    unsuccessful_search_rate = (1 - phi[i]) * (1 - normal_cdf(t_i[i], 0, 1));
    search_rate[i] = (successful_search_rate + unsuccessful_search_rate);
    hit_rate[i] = successful_search_rate / search_rate[i];
   }
}

model {
  // Draw threshold hyperparameters
  sigma_t ~ normal(0, 1);

  // Draw race parameters. Each is centered at a mu, and we allow for inter-race heterogeneity.
  mu_phi ~ normal(0, 1);
  mu_delta ~ normal(0, 1);
  phi_r    ~ normal(mu_phi, .1);
  delta_r ~ normal(mu_delta, .1);
  t_r ~ normal(0, 1);

  // Draw department parameters (for un-pinned departments)
  phi_d_raw    ~ normal(0, .1);
  delta_d_raw ~ normal(0, .1);

  // Thresholds
  t_i_raw ~ normal(0, 1);

  s ~ binomial(n, search_rate);
  h ~ binomial(s, hit_rate);
}
```

```r
stan_data = with(stops, list(
   N = nrow(stops),
   D = length(unique(police_department)),
   R = length(unique(driver_race)),
   d = as.integer(police_department),
   r = as.integer(driver_race),
   n = num_stops,
   s = num_searches,
   h = num_hits))

model <- stan_model(file = 'threshold_test.stan')

fit <- sampling(
  model, data = stan_data, iter=5000,
  init = 'random', chains=5,
  cores=5, refresh=50, warmup = 2500,
  control = list(adapt_delta = 0.95,
                 max_treedepth = 12,
                 adapt_engaged = TRUE))

post = rstan::extract(fit)
```

The figure below shows the differences in the thresholds faced by white and minority drivers in each department. We see that Hispanic drivers face substantially lower thresholds than white drivers in all departments, while black drivers face slightly lower thresholds in a majority of departments.

```r
signal_to_p = function(x, phi, delta){
  #Checked. Converts x -> p.
  p = phi * dnorm(x, delta, 1) / (phi * dnorm(x, delta, 1) + (1 - phi) * dnorm(x, 0, 1));
  return(p)
}

plot_department_thresholds = function(obs, post) {
  colors = c('blue', 'black', 'red')
  races = as.character(levels(obs$driver_race))
  obs$thresholds = colMeans(signal_to_p(post$t_i, post$phi, post$delta))
  mx = max(obs$thresholds)
  df = obs %>% filter(driver_race == 'White') %>%
    right_join(obs %>% filter(driver_race != 'White'), by = 'police_department')
  ggplot(df) +
    geom_point(aes(x=thresholds.x, y=thresholds.y, size = num_stops.y), alpha=0.8, shape = 1) +
    geom_abline(slope=1, intercept=0, linetype='dashed') +
    scale_y_continuous('Minority threshold\n', limits=c(0,mx), labels=percent, expand=c(0, 0)) +
    scale_x_continuous('\nWhite threshold', limits=c(0,mx), labels=percent, expand=c(0, 0)) +
    scale_size_area(max_size=15) +
    theme(legend.position=c(0.0,1.0),
          legend.justification=c(0,1),
          legend.title = element_blank(),
          legend.background = element_rect(fill = 'transparent'),
          panel.margin.x=unit(1.5, "cm")) +
    scale_color_manual(values = colors[-1], labels=races[-1]) +
    guides(size=FALSE) + facet_grid(.~driver_race.y)
}



plot_department_thresholds(stops, post)
```



To compute the average threshold faced by North Carolina drivers of each race, we weigh the department thresholds by the number of stops conducted by that department (of drivers of all races). Black and Hispanic drivers face lower thresholds than whites.

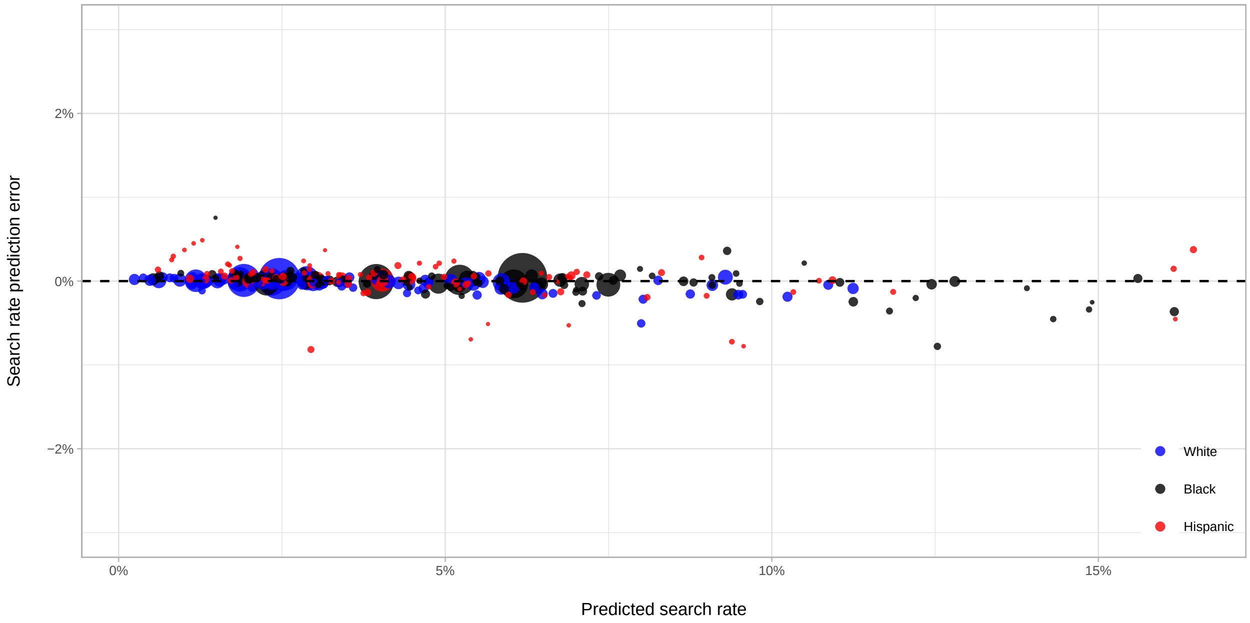
```
  stops = stops %>%
    mutate(thresholds = colMeans(signal_to_p(post$t_i, post$phi, post$delta))) %>%
    group_by(police_department) %>%
    mutate(total_stops = sum(num_stops)) %>%
    ungroup()

  na_replace = function(x, r) ifelse(is.finite(x), x, r)

  accumrowMeans = function(M, i, w = rep(1, nrow(M)), imax = max(i)) {
    t(sapply(1:imax, function(j) (i == j)*na_replace(w/sum(w[i == j]),0))) %*% M
  }
  avg_thresh = accumrowMeans(t(signal_to_p(post$t_i, post$phi, post$delta)),
                             as.integer(stops$driver_race), stops$total_stops)

  data.frame(levels(stops$driver_race),
             sprintf('%.3f', rowMeans(avg_thresh)),
             apply(rowQuantiles(avg_thresh, probs = c(0.025, 0.975)), 1,
                   function(x) paste0('(', paste0(sprintf('%.3f',x), collapse = ', '), ')'))
             ) %>%
    setNames(c('Driver Race', 'Average Threshold', '95% Credible Interval'))
```

```
##   Driver Race Average Threshold 95% Credible Interval
## 1       White             0.230        (0.222, 0.238)
## 2       Black             0.206        (0.197, 0.215)
## 3    Hispanic             0.137        (0.130, 0.144)
```

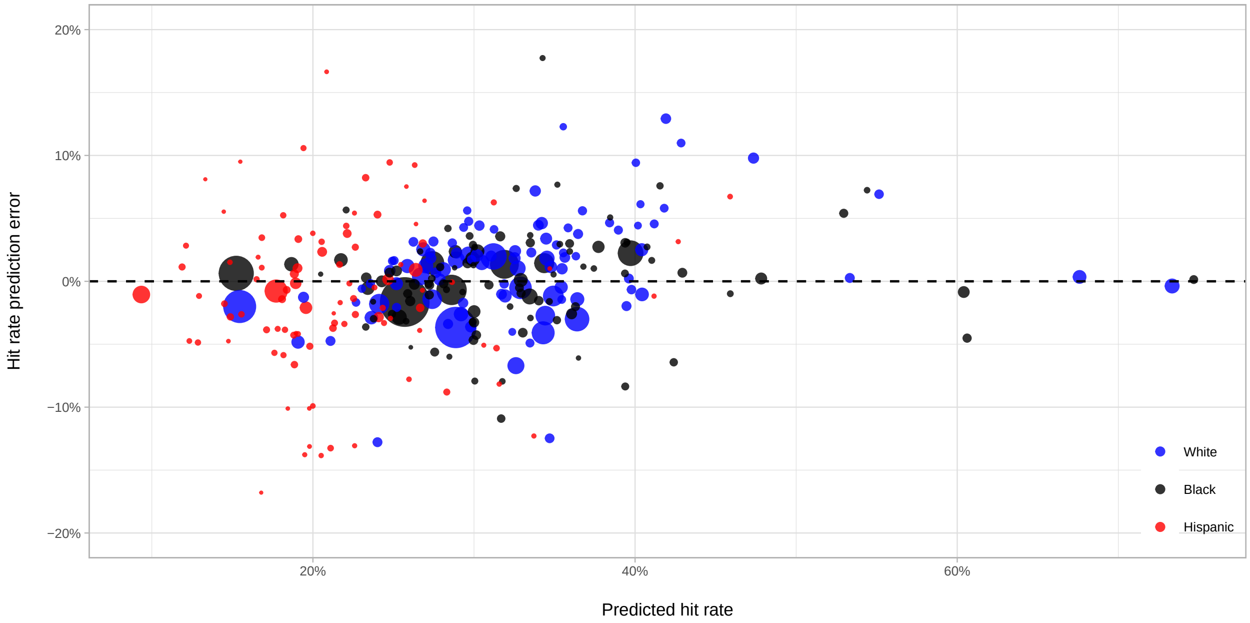
Posterior Predictive Checks

We investigate 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. 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, we analytically compute the resulting search and hit rates and average these over the 2,500 posterior draws. The figures below compare the model-predicted search and hit rates to the actual, observed values and indicates that the fitted model recovers the observed search rates almost perfectly across races and departments.

```
search_rate_ppc <- function(obs, post, ylim = 0.03) {
  obs$pred_search_rate = colMeans(post$search_rate)
  ggplot(data=obs, aes(x=pred_search_rate, y=pred_search_rate-search_rate)) +
    geom_point(aes(size=num_stops, color=driver_race), alpha = 0.8) +
    scale_size_area(max_size=10) +
    scale_x_continuous('\nPredicted search rate', labels=percent)+
    scale_y_continuous('Search rate prediction error\n', labels=percent, limits=c(-ylim, ylim)) +
    geom_abline(slope=0, intercept=0, linetype='dashed') +
    theme(legend.position=c(1.0,0),
          legend.justification=c(1,0),
          legend.title = element_blank(),
          legend.background = element_rect(fill = 'transparent')) +
    scale_color_manual(values=c('blue','black','red', 'green4')) +
    guides(size=FALSE)
}

hit_rate_ppc <- function(obs, post, ylim = 0.2) {
  obs$pred_hit_rate = colMeans(post$hit_rate)
  ggplot(data=obs, aes(x=pred_hit_rate, y=hit_rate-pred_hit_rate)) +
    geom_point(aes(size=num_stops, color=driver_race), alpha=0.8) +
    scale_size_area(max_size=10) +
    scale_x_continuous('\nPredicted hit rate', labels=percent) +
    scale_y_continuous('Hit rate prediction error\n', labels=percent, limits = c(-ylim, ylim)) +
    geom_abline(slope=0, intercept=0, linetype='dashed') +
    theme(legend.position=c(1.0,0),
          legend.justification=c(1,0),
          legend.title = element_blank(),
          legend.background = element_rect(fill = 'transparent'))+
    scale_color_manual(values=c('blue','black','red', 'green4')) +
    guides(size=FALSE)
}

search_rate_ppc(stops, post)
```



```
hit_rate_ppc(stops, post)
```



Finally, we check convergence of the model fit by inspecting Rhat and the number of effective samples.

```
s <- summary(fit)
Rhat <- s$summary[,'Rhat']
```

```
n_eff <- s$summary[,'n_eff']
summary(Rhat, na.rm=T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.9996  0.9998  1.0000  1.0014  1.0003  1.0344       2
```

```
summary(n_eff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   244.2  7789.8 12500.0  9930.2 12500.0 12500.0
```

Conclusions

Theoretical limitations with the two most widely used tests for discrimination—the benchmark and outcome tests—have hindered investigations of bias. Addressing this challenge, we have developed a new statistical approach to detecting discrimination that builds on the strengths of the benchmark and outcome tests and that mitigates the shortcomings of both. On a dataset of 4.5 million motor vehicle stops in North Carolina, our threshold test suggests that black and Hispanic motorists face discrimination in search decisions.

In our published paper on this method (Simoiu *et al.*, 2017) we check the robustness to reasonable violations of the model assumptions, including noise in estimates of the likelihood a driver is carrying contraband. We also attempt to rule out some of the more obvious legitimate reasons for which thresholds might vary, including search policies that differ across department, year, or time of day. However, as with all tests of discrimination, there is a limit to what one can conclude from such statistical analysis alone. For example, if search policies differ not only across but also within department, then the threshold test could mistakenly indicate discrimination where there is none. Such within-department variation might result from explicit policy choices, or as a by-product of deployment patterns; in particular, the marginal cost of conducting a search may be lower in heavily policed neighborhoods, potentially justifying a lower search threshold in those areas. To a large extent, such limitations apply equally to past tests of discrimination, and as with those tests, caution is warranted when interpreting the results.

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

Bibliography

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.” Princeton University Press, 1973.

Ian Ayres. “Outcome tests of racial disparities in police practices.” Justice Research and Policy, 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, 1993

Andrew Gelman, Xiao-Li Meng, and Hal Stern. “Posterior predictive assessment of model fitness via realized discrepancies.” Statistica Sinica, 1996

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. “Fast threshold tests for detecting discrimination.” Forthcoming, 2018.

Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. “The problem of infra-marginality in outcome tests for discrimination.” Annals of Applied Statistics, 2017.