Demonstrate how disparate impact statistics can fail to recover known disparities

Max Griswold

2024-01-11

Introduction

Disparate impact cases require demonstrating if a disparity in outcomes for a protected class is statistically significant and sufficiently sized. This project aims to use simulation to show a range of cases where the conventional statistical rules for determining a disparate impact -the four-fifths rule and significance rule - can lead to different conclusions due

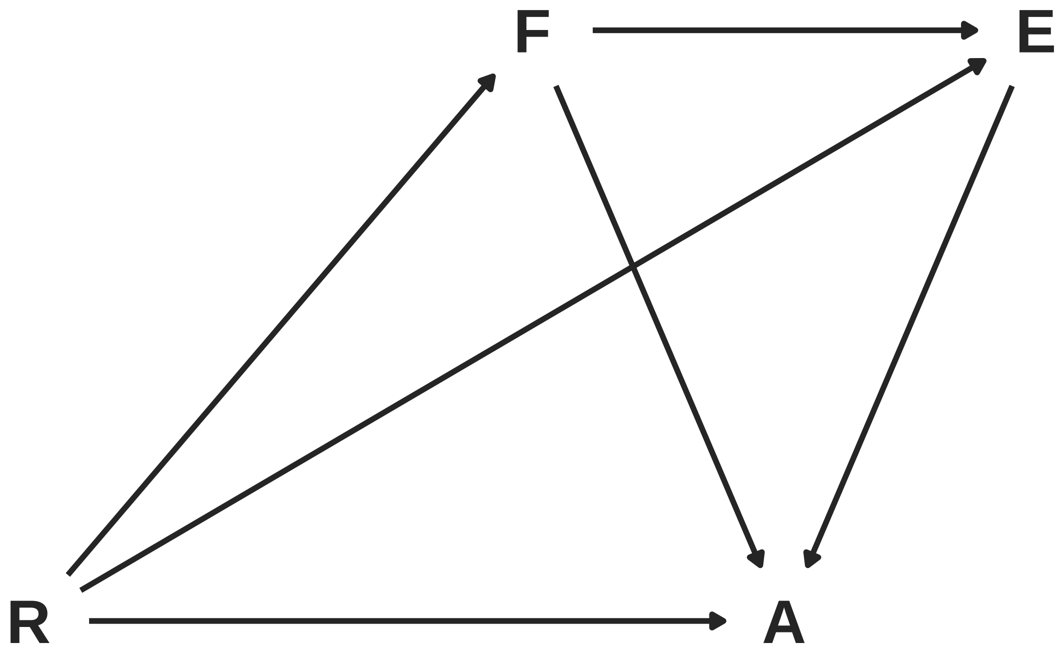
This notebook demonstrates a simplified example of how we might conduct simulations. The setup concerns a landlord determining whether to accept a tenant's rental application, based on their income and former-incarceration status. In each of the three scenarios below, the landlord aims to deny tenancy to applicants based on their race.

Problem setup

To start, I generate data according to the following data generating process (figure 1). I assume that an applicant's race is correlated both with their monthly earnings and with the probability of being formerly-incarcerated. I also assume monthly earnings also depend on former incarceration status. I assume in each scenario that these three variables uniquely determine if a landlord

I determined the probability of former incarceration status using data from DOJ-OJP, monthly earnings based off census income, and the effect of incarceration status on monthly earnings, stratified by race, from estimates in Western, 2017.

Figure 1: Synthetic Data Generating Process



R = Race, J = Former−incarceration status,
E = Monthly Earnings, A = Application status

I set up the following decision scenarios for the landlord:

1. The landlord engages in direct discrimination: If the applicant is black, they offer a unit to them 25% less often than a comparable white candidate.
2. The landlord attempts to obscure their motive by only denying tenancy to a fraction of applicants. Every third black applicant they encounter, they deny their tenancy 75% more often.
3. The landlord puts more weight on former-incarceration status to decide if they deny an applicant tenancy. They choose a probability of denying tenancy based on former-incarceration status so they deny tenancy to an equal proportion of black-to-white tenants as in the first two scenarios.

Across all scenarios, the same ratio of black-to-white applicants are denied tenancy. I used monte-carlo simulation to generate 500 synthetic datasets and varied the sample size of applicants between 200 and 4000.

For each simulation and scenario, I regressed a tenant's application status (accepted or denied) on the tenant's race, former-incarceration status, and monthly earnings using a binomial regression with a logit link. This is a routine model used to detect disparate impact (Ayres, 2010; Jung et al., 2023)

I then obtained the estimated p-value for the beta-coefficient on race and used the regression equation to calculate an estimate for the ratio of black-to-white applicants denied tenancy. Finally, I took the mean of these estimates across simulation iterations.

```r
prob_ratio <- function(p1, p2){

  p1 <- sum(p1)/length(p1)
  p2 <- sum(p2)/length(p2)

  return(p1/p2)
```

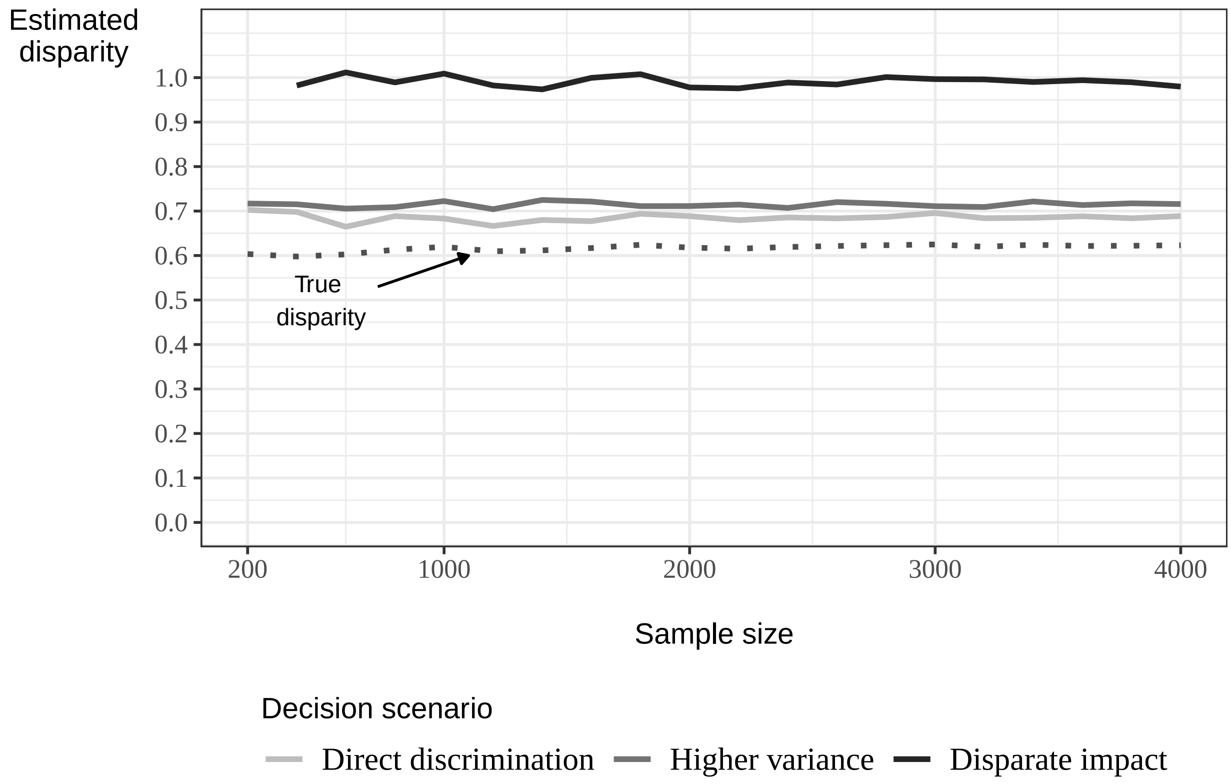
```
}

inv_logit <- function(v){
    return(exp(v)/(1 + exp(v)))
}

sim_scenarios <- function(n, iter, sum_res = T){

  # Create population by race/ethnicity
  race <- rbinom(n = n, 1, prob = 0.14)

  # Set up Jail probability. Note this is only determined by race in the DAG, with
  # some added noise to ensure these variables aren't perfectly correlated. I used
  # https://bjs.ojp.gov/content/pub/pdf/Llgsfp.pdf to set up base probabilities

  prob_jail <- 0.04 + race*0.241 + runif(n, -0.03, 0.03)
  jail <- rbinom(n = n, 1, prob = prob_jail)

  # Set used monthly earning amounts based on distribution of
  # usa income + modifiers based on race/incarceration status from:
  # https://scholar.harvard.edu/files/brucewestern/files/racial_inequality_in_employment_and_earnings_a

  earn <- rgamma(n, 2, 3)*1000
  earn <- earn*(1 - (0.2*jail + 0.35*race))

  # Set up scenarios to determine probability of offering an apartment to the
  # applicant. Start with baseline probabilities for the first two cases for
  # income and incarceration status:
  dp_jail   <- 1 - 0.5*jail

  # Rescaling income between 0 and 1 works out to be a
  # reasonable probability
  dp_income <- (log(earn) - min(log(earn)))/(max(log(earn)) - min(log(earn)))

  # Scenario 1: For black applicants, offer the apartment 25% less often
  dp1 <- (1 - 0.25*race)*dp_income*dp_jail
  d1 <- rbinom(n, 1, dp1)

  # Scenario 2: For 1/3rd of random black applicants, offer the apartment
  # 75% less often
  bad_luck <- rbinom(n, 1, prob = ifelse(race == 1, 0.3, 0))
  dp2 <- (1 - 0.75*bad_luck)*dp_income*dp_jail
  d2  <- rbinom(n, 1, dp2)

  # Scenario 3: Almost always reject formerly-incarcerated individuals.
  dp_jail_alt <- 1 - 0.99*jail
  dp3 <- dp_income*dp_jail_alt
  d3  <- rbinom(n, 1, dp3)

  df_sim <- data.table('race' = race,
                       'earn' = earn,
                       'jail' = jail,
                       'luck' = bad_luck,
                        'p1' = dp1,
                        'p2' = dp2,
                        'p3' = dp3,
                        'd1' = d1,
                        'd2' = d2,
                        'd3' = d3,
                        "iter" = iter,
                        "n" = n)

  # Hardcoding this to move quickly; I should be using a for loop for this step
  # (and in a few previous steps too):
  df_sim[, rr1 := prob_ratio(df_sim[race == 1,]$d1, df_sim[race == 0,]$d1)]
  df_sim[, rr2 := prob_ratio(df_sim[race == 1,]$d2, df_sim[race == 0,]$d2)]
  df_sim[, rr3 := prob_ratio(df_sim[race == 1,]$d3, df_sim[race == 0,]$d3)]

  # Using a poisson with a log link + robust standard errors so I can interpret the
  # coefficients as relative risks
  mod1 <- glm(d1 ~ race + earn + jail, data = df_sim, family = binomial(link = logit))
  mod2 <- glm(d2 ~ race + earn + jail, data = df_sim, family = binomial(link = logit))
  mod3 <- glm(d3 ~ race + earn + jail, data = df_sim, family = binomial(link = logit))

  res <- list(mod1, mod2, mod3)

  # Extract estimated probability of acceptance by race and associated p-value.
  for (i in 1:3){
    est_name  <- paste0("est", i)
    pval_name <- paste0("pval", i)

    p1 <- predict(res[[i]], newdata = data.frame(race = 1, jail = 0, earn = 0), type = 'response')[[1]]
    p2 <- predict(res[[i]], newdata = data.frame(race = 0, jail = 0, earn = 0), type = 'response')[[1]]
    est_prob_ratio <- p1/p2

    df_sim[, (est_name)  := est_prob_ratio]
    df_sim[, (pval_name) := summary(res[[i]])$coefficients['race','Pr(>|z|)']]

  }

  # Summarize results and return a reduced dataset
  if (sum_res == T){
    df_sim <- unique(df_sim[, .(iter, n, est1, est2, est3,
                                pval1, pval2, pval3,
                                rr1, rr2, rr3)])

    df_sim <- melt(df_sim, id.vars = c("iter", "n"), measure = patterns("^est", "^pval", "^rr"),
                value.name = c("est", "pval", "rr"), variable.name = "scenario")
  }

  return(df_sim)

}
```

Results and discussion

To meet the standards for disparate impact, the protected class needs to be selected as an applicant less than 80% of the time (the four-fifths rule) and this effect needs to be significant at the 0.05 significance level. Since we set up the scenarios, we know for a fact that landlords are discriminating against black applicants 25% more often than white applicants. However, can the models detect this effect?

Figure 2 displays the estimated rate at which black applicants are selected for a unit compared to white applicants. For scenario 1 and 2, the model is able to detect a disparity of roughly 70%. For scenario 3, the estimated disparity is close to 100%. While scenario 1 and 2 are able to meet the four-fifths rule, all scenarios underestimate the true disparity. This can be seen by examining the data generating process. Since the models include controls, the effects of race in these models will not estimate the mediating effect of race on earnings and former-incarceration status. Accordingly, all estimates are unable to detect the true disparity observed in the outcome. Increasing sample size does not improve the ability of the models to detect the true disparity.

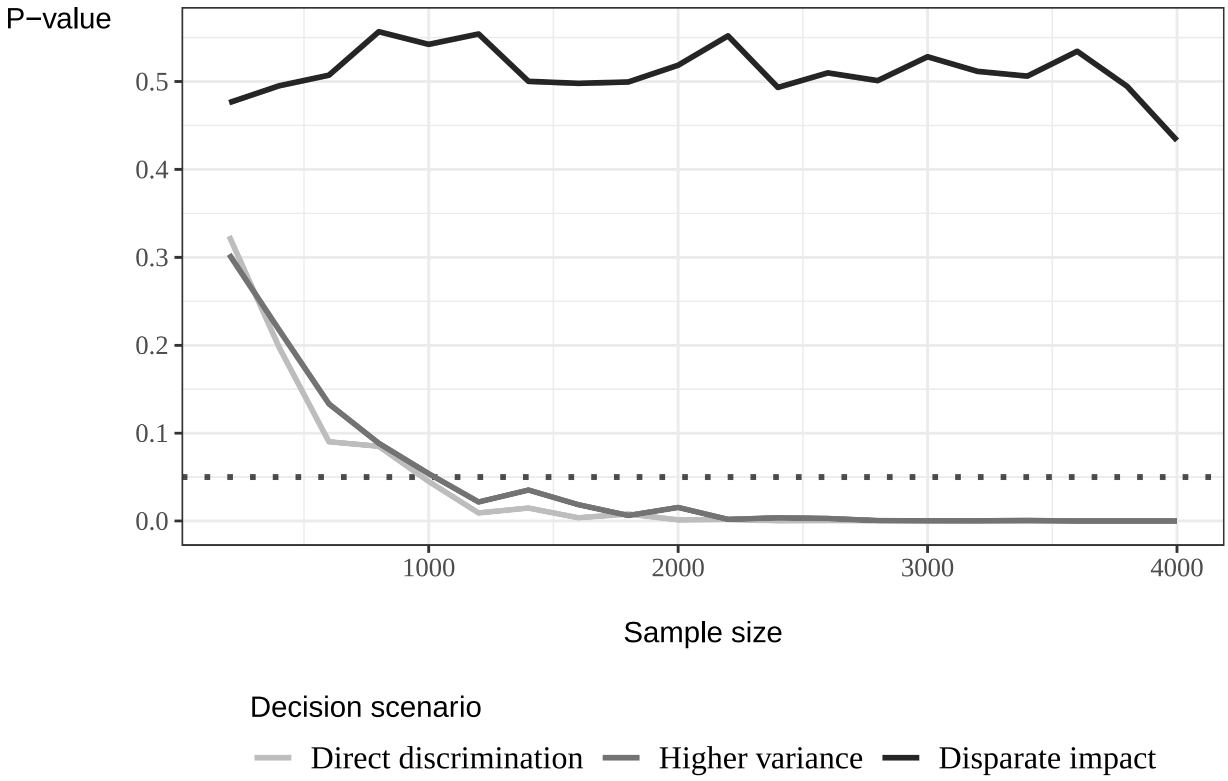
Figure 2: Estimated disparity ratio in applicant selection



Additionally, the estimating equations vary substantially in terms of detecting effects at a 0.05 significance level. Figure 3 below displays the average estimated p-value by scenario and sample sizes. Scenario 1 is identified by the model once there's a sample size of 1000 Scenario 2, which is more variable in how the the landlord makes a discriminatory decision, unsurprisingly requires a larger sample size to detect the effect, needing a minimum sample size of 1200.

However, in scenario 3, the model is unable to detect a significant effect at any sample size. This scenario is the canonical case where disparate impact tests would be most useful since the landlord's decision-making did not depend on race per se (i.e. no intent) but led to a disparate impact. This result should be unsurprising based on the data generating process above (since the model controls for both earnings and former incarceration status, there is no possibility for the race variable to generate an effect).

Figure 3: Estimated p−value for race coefficient, by scenario and sam


P−value
Sample size
Decision scenario
Direct discrimination
Higher variance
Disparate impact