The Insufficiency of Statistics for Detecting Racial Discrimination by Police

Naftali Weinberger

October 17, 2022

Abstract

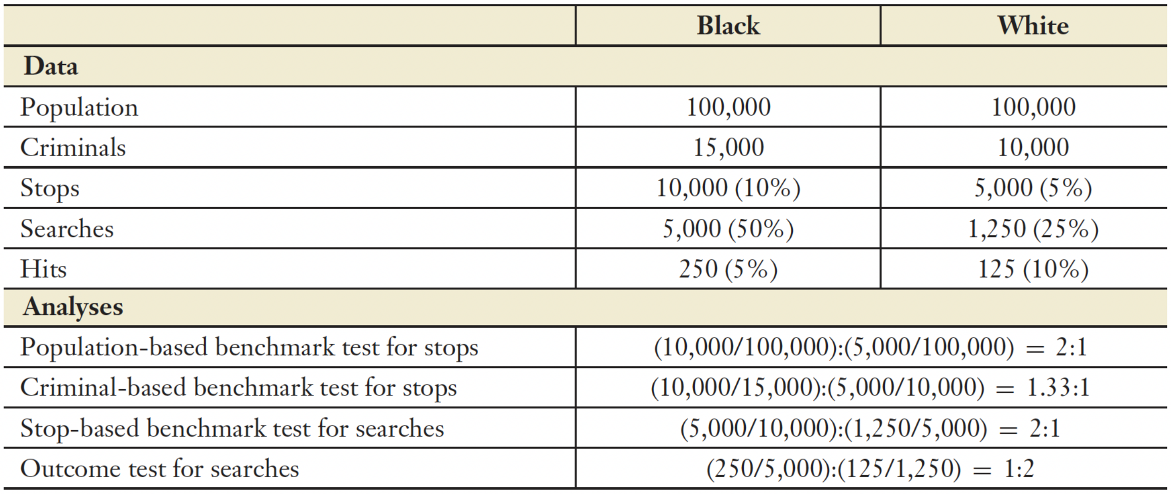
Benchmark tests are widely employed in testing for racial discrimination by police. Neil and Winship (2019) correctly point out that the use of such tests is threatened by the phenomenon of Simpson's paradox. Nevertheless, they avoid giving a causal analysis of the paradox, and as a result cannot adequately account for the relationship between statistical quantities and discrimination hypotheses. Simpson's paradox reveals that the statistics employed in benchmark tests will not, in general, be invariant to updating on new information. I argue that as a result of this, benchmark statistics should not by themselves be taken to provide *any* evidence for or against discrimination, absent additional modeling assumptions. Neil and Winship highlight ways in which benchmark statistics that appear to provide evidence for discrimination no longer do so given additional assumptions, but they lack an account of which sets of assumptions would ensure invariance. Causal models provide such an account. This motivates the use of causal models when using statistical methods as evidence for discrimination.

1 Introduction

Consider a study indicating that police performing traffic stops in Pittsburgh search minority drivers at a higher rate than non-minority drivers. This would seem to provide at least some evidence that the police are discriminating. But this is not the only possible explanation. For example, it could be that the police make stops based on observing suspicious activities, and that minority drivers disproportionately engage in such activities. Given the assumptions that A) it is legitimate to use suspicious activity as a basis for stops and B) whether a driver gets deemed "suspicious" is not just a covert re-description of their race, the police actions might not count as discriminatory. Yet the raw statistics concerning the racial disparity in stops do nothing to differentiate the discriminatory and non-discriminatory explanations. For this reason, legal and empirical studies of discrimination often employ *benchmark tests*. Such tests involve statistically conditioning on covariates that differentiate the relevant groups in order to determine what the disparity between the group stop-rates would be in the absence of discrimination. The idea, roughly, is that once one takes into account the factors that legitimately could explain the disparity, any remaining disparity is evidence of discrimination.

As Neil and Winship (2019) note, benchmark tests are fatally undermined by Simpson's Paradox (Simpson, 1951; Sprenger and Weinberger, 2021). An example of the paradox would be a case in which police stopped minorities and non-minorities at the same rate in Pittsburgh as a whole, but stopped minorities at a higher rate within every single district. Accordingly, statistical claims involving comparisons of relative rates across populations – including the rates invoked in benchmark tests – will not be robust to conditioning on additional covariates. While their discussion compellingly illustrates the limitations of benchmark tests, they say little about how to address the paradox systematically. Doing so is important not only because the paradox is widely discussed in the empirical discrimination literature (see e.g. Bickel et al., 1975; Ross et al., 2018), but also because it illuminates the role of causal assumptions in interpreting statistics relevant to discrimination.

The first general lesson I will draw from my discussion of the paradox concerns the sense in which discrimination statistics provide evidence for claims about police discrimination. One might be tempted by the position that discovering that police stop non-minorities and minorities at the same rates would count as evidence against discrimination, and that subsequently learning that minorities are stopped at a higher rate within every district would count as countervailing evidence. In contrast, I argue that the



	Black	White
Data		
Population	100,000	100,000
Criminals	15,000	10,000
Stops	10,000 (10%)	5,000 (5%)
Searches	5,000 (50%)	1,250 (25%)
Hits	250 (5%)	125 (10%)
Analyses		
Population-based benchmark test for stops	(10,000/100,000):(5,000/100,000) = 2:1	
Criminal-based benchmark test for stops	(10,000/15,000):(5,000/10,000) = 1.33:1	
Stop-based benchmark test for searches	(5,000/10,000):(1,250/5,000) = 2:1	
Outcome test for searches	(250/5,000):(125/1,250) = 1:2	

Figure 1: Hypothetical population with benchmark statistics (from Neil and Winship (2019)

statistics being cited provide no evidence for or against discrimination, absent additional substantive assumptions about the variables being modeled. Since Simpson's paradox reveals that comparisons of relative rates across populations are not robust to conditioning on additional variables, non-statistical assumptions are required to draw any conclusions about discrimination, even tentative ones.

The second lesson I draw concerns an underappreciated role of causal assumptions in empirical modeling. Causal models are often advertised as licensing inferences concerning experimental interventions. Additionally, such models can provide a framework for differentiating meaningful from non-meaningful statistical relationships. Given that statistics alone cannot provide evidence for discrimination absent additional substantive assumptions, a further framework is required for representing such assumptions in a general way. I will argue that causal models provide precisely such a framework.

2 Neil and Winship on Benchmark Tests

Neil and Winship's diagnoses of the problems with benchmark tests all appeal to the table reproduced in figure 1. The table presents a hypothetical population with some corresponding benchmark statistics. The rate at which black individuals are stopped is twice as high as that for white individuals. While one might suggest that this simply reflects the higher rate of criminality in the hypothetical black population, that rate is only 50% higher than in the white population, and thus cannot account for the difference. These considerations suggest that in a population where criminality is unevenly distributed across races, one would not expect the proportion of stops to be equal, even assuming that police are not racially discriminating. The *criminal-based benchmark test* statistically adjusts for criminality by comparing stop rates as a proportion of criminality rates in each population (see table). Since this benchmark adjusts for the different rates of criminality – and reveals that the rate of black stops will be higher even post-adjustment – it plausibly provides evidence of police discrimination. Yet Neil and Winship show that there are numerous ways of filling in further details about this population such that there either is no discrimination, or a level of discrimination that differs from that suggested by this benchmark statistic.

For now it will suffice to provide a single illustration of why benchmark statistics may mislead. Neil and Winship (p. 79) imagine a scenario in which police only stop people spending time in public spaces and that the racial distribution of public-space users consists of 40,000 blacks to 20,000 whites. Assuming that police stop individuals as the same rate of .25, regardless of race, the numbers and proportions of stops will match those given in figure 1. But, by stipulation, police are not taking race into account in choosing who to stop. Neil and Winship analyze discrimination using the "similarly situated" criterion, which entails that since here police are not differentiating among individuals who are similar in all respects other than race, they are not discriminating.[1] Accordingly, a benchmark statis-

[1] See Kohler-Hausmann (2018) for a criticism of the similarly situated criterion.

tic that may seem to provide evidence of discrimination no longer does so after specifying additional information.

The scenario in which racial disparities in stops reflect disparities in public space occupancy is just one illustration of how benchmark statistics may fail to reflect the presence, absence or degree of discrimination. I will consider some of Neil and Winship's other scenarios below. Here I want to highlight a feature of their general strategy. After pointing out that the criminal-based benchmark will fail to correspond to discrimination if black and white individuals are observed by police at different rates, they claim that those employing the benchmark are implicitly making the false assumption that differently-raced individuals are observed by the police at the same rates. More generally, in all their examples they present a scenario in which the actual degree of discrimination diverges from that of the benchmark, and then fault those employing the benchmark for making the false assumption that the relevant scenario does not obtain. While they are correct that benchmarks will only be justified given substantive assumptions, their treatment of specific benchmarks as corresponding to particular assumptions suggests that the assumptions in question are statistical. In what follows, I will use Simpson's paradox to illustrate why statistical assumptions do not suffice.

3 The Causal Analysis of Simpson's Paradox

Simpson's paradox refers to cases in which the probabilistic relationship between two variables in a population differs from that found in every one of its subpopulations, where subpopulations are derived by conditioning on different values of a variable (or set of variables). For example, in the early days if COVID-19, the fatality rate among those infected was higher in Italy than in China, but within every age group, the fatality rate was higher in China than in Italy (von Kügelgen et al., 2021). Put differently, although learning that someone was infected in Italy as opposed to China provides evidence that they are more likely to die, once one learns the person's age, this relationship reverses (no matter what age they are). Simpson's paradox is not paradoxical in the sense of being impossible. It is consistent with the axioms of probability theory and arises in actual cases. It is paradoxical only in the sense that reasoners find instances of it to be perplexing, and there is an ongoing discussion regarding which features of human reasoning lead to this reaction (Sprenger and Weinberger, 2021).

Simpson's paradox is only one of the challenges that Neil and Winship raise for benchmark tests, but a proper understanding of the paradox is relevant to all of them. Since benchmark tests compare proportions within populations partitioned using a particular set of variables, a clear understanding of how proportions change relative to one another as one conditions on additional variables is essential to their interpretation. In cases of Simpson's paradox, the relationships in the population fail to be a guide to those in the subpopulations. In this section, I will present a standard causal analysis for when the paradox arises, highlight some common misunderstandings of the paradox, and then show how these misunderstandings are present in Neil and Winship's analysis.

	Full Population, N=52		Success	Men (M), N=20		Success	Women (¬M), N=32		Success
	Success (Y)	Failure (¬Y)	Rate	Success	Failure	Rate	Success	Failure	Rate
Treatment (X)	20	20	50%	8	5	≈ 61%	12	15	≈44%
Control (¬X)	6	6	50%	4	3	≈ 57%	2	3	≈40%

Table 1: Simpson's Paradox: the type of association at the population level (positive, negative, independent) changes at the level of subpopulations. Numbers taken from Simpson's original example (1951).

I'll illustrate the paradox using numbers from the paper by the eponymous Simpson (1951). Table 1 compares the success rates of a medical treatment in a whole population as well as in the male and female subpopulations (assumed to be exhaustive). In the population, the success rate is the same in both the treatment and control groups, and thus treatment (X) is uncorrelated with success (Y). However, the treatment raises the probability of success in both the male (M) and female ($\neg M$) subpopulations. Two features of this table are crucial for seeing what is going on. First, note that there are more females than males in the population. Second, even though the treatment *raises* the probability of success in both subpopulations, men are more likely to recover than females even if they do not take the treatment.

The graph in figure 2 presents a plausible causal representation of the variables in the example. The arrows from *Gender* and *Treatment* to *Success* indicate that they are causes of success (though

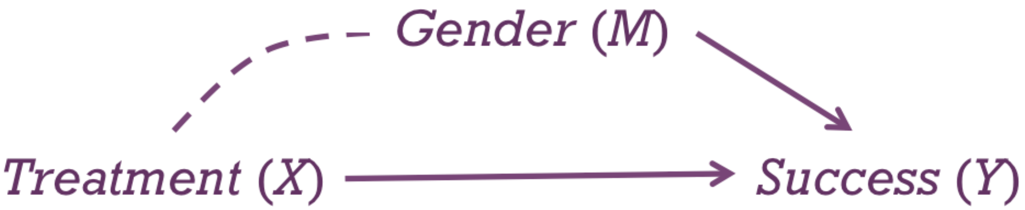


Figure 2: Caption

the precise quantitative relationship is unspecified). The dashed arrow between *Gender* and *Treatment* indicates that they are correlated. The causal explanation for this correlation is not further specified, though it is important for the analysis that it is *not* due to the treatment being a *cause* of gender.

Broadly speaking, determining whether the treatment is a cause of success involves disentangling two ways that the former may be evidentially relevant to the latter. On the one hand, experimental participants who receive the treatment may have a higher success rate because the treatment in fact causally promotes success. On the other hand, it may be merely that *learning* that someone received the treatment provides evidence of their gender (due to the stipulated correlation) and that one's gender provides information about their probability of recovering whether or not one receives the treatment. Experiments in which one *intervenes* so that whether someone receives the treatment is uncorrelated with any feature (e.g. gender) that may influence success enable one to differentiate between causal and merely evidential relevance. Formally, this distinction is marked by the introduction of the operator $do()$, where $P(Y|do(X))$ indicates the probability distribution of Y that would result from intervening on X (this may differ from the "observational" distribution $P(Y|X)$).

The key concept for understanding the relationship between causal and statistical assumptions is that of *identifiability*. Identifiability is a relationship between (i) a probability distribution, (ii) a causal graph, and (iii) a causal quantity (such as the magnitude of the effect of one variable on another). A causal quantity is identifiable iff, given the causal assumptions represented in a graph, it is possible to uniquely determine its value from the probability distribution. To illustrate, if there was an unmeasured common cause of the treatment and success, then the effect of the former on the latter would *not* be identifiable, since no matter how much information one had regarding the probability distribution, one could not determine the extent to which any correlation between treatment and success is due to the causal relationship as opposed to the common cause. In contrast, if the causal model is that in figure 1 - indicating that the only evidential relationship between the treatment and success is that via *gender* – then the effect of the treatment on success would be identified conditional on gender.

More specifically, given the assumptions in the causal graph – most crucially that *treatment* does not cause *gender* – the causal effect of the treatment X on success Y can be identified using the following formula:

$$P(Y|\mathrm{do}(X)) = \sum_{M} P(Y|X, M)\, P(M) \tag{1}$$

This indicates that the effect of X on Y in the population can be derived as a weighted average of the conditional probabilities in the subpopulations partitioned based on gender. The significance of this expression is that although the non-causal conditional probability $P(Y|X)$ may differ arbitrarily from $P(Y|X, M)$ and $P(Y|X, \neg M)$, the causal conditional probability $P(Y|do(X))$ must average over conditional probabilities in the subpopulations (which here also correspond to the subpopulation-specific effects). Given this averaging, it is impossible for X to causally raise the probability of Y in the population, but not in *any* of the subpopulations. The crux of the causal explanation for why reasoners find cases of Simpson's paradox paradoxical is that they conflate causal and probabilistic relevance. While it is entirely possible for X to be evidentially relevant to Y in a population, but not in any of its subpopulations, it is not possible for X to causally raise Y's probability in the population, but not it any subpopulations.

As emphasized, the adjustment formula in equation (1) only applies assuming that M is not an effect of X. Suppose instead that we were considering a different example in which X does cause M – say, because M is a certain blood chemical that mediates the way that the treatment is effective. In such a case, one should *not* condition on M when identifying the effect of X on Y. So whether one should condition on M – and, correspondingly, whether one should consult the population or the M-partitioned subpopulations in identifying the relevant effect – depends on the relationships in the causal model. Note when X and M are direct causes of Y, the model in which M is a common cause of X and Y is statistically indistinguishable from that in which it is a mediator, in the sense that any data generated from one model could have been generated from the other. This highlights the importance of causal information in determining whether to consider populations or subpopulations.

While one should not condition on mediators in evaluating the *total* effect of X on Y along all paths, one does condition on mediators (in particular ways) when evaluating the influence of X on Y along particular paths. We will return to this point below.

This brief discussion of Simpson's paradox will suffice to rebut several common misconceptions about it. The first misconception is the paradox teaches us to avoid mixing heterogeneous populations. This interpretation seems plausible in the first example above, where the statistical relationships in the male and female populations – but not in the whole population – reflected the causal ones. But partitioning a population into subpopulations is not a general solution to the paradox, since where M is a mediator, then the effect of X on Y is not identified in M-specific subpopulations. Whether one should partition on a variable depends on its causal relationships to the others. Moreover, there is nothing problematic about average effects. Assuming one can identify an effect in a population – e.g. by using the formula in equation (1) – the population effect will be an average of those in the causally dissimilar male and female subpopulation. But, being an *average* of the subpopulation effects, there is no possibility of (e.g.) its being positive while the effects in all subpopulations are negative.

A second misconception is that Simpson's paradox is a reference-class problem (Cartwright, 1979, p. 426). Reference class problems are those in which the value of a particular quantity (such as a probability) depends on how that quantity is described. This description sensitivity is taken to threaten the objectivity of what is being quantified. But Simpson's paradox does not threaten the objectivity of causation. Although, as in the case just described, the magnitude of an effect can differ across populations, this reflects that the effect is *in fact* different across the populations, as a result of variation in background factors. As long as one is dealing with genuine causal effects, as opposed to non-causal conditional probabilities, the quantities measured will reflect the effect in the relevant population.

Both misconceptions are alive and well in Neil and Winship's discussion. After presenting some examples in which benchmark tests are undermined due to Simpson's paradox, they present the following explanation:

> When the police behave differently across the strata of some variable, but a researcher's analysis uses data that ignores and aggregates across this distribution, Simpson's paradox threatens to give outcome statistics that are inconsistent with reality. As we have shown, this problem can bias benchmark tests, as well as the outcome test, for stops and searches. (2019, p. 85)

In talking about Simpson's paradox as one of aggregation, they are assuming that the basic problem is one of mixing heterogeneous populations. Their comment regarding statistics that are "inconsistent with reality" suggests a concern about objectivity. Finally, note their claim that the paradox can "bias" benchmark tests. To rigorously talk about a measurement being biased, one needs to first clarify which quantity one is aiming to measure. Neil and Winship do not do so. Their implicit belief seems to be that there is some comparison of the stop rates across the populations that would establish whether there is discrimination, and that benchmarks simply provide the wrong comparison. Figuring out which rates are the relevant ones is treated as a statistical problem. But Simpson's paradox reveals that, in general, comparisons of rates or proportions across populations will not be invariant to conditioning on additional variables, absent additional modeling assumptions. And we have seen that causal models serve as an important tool for representing those assumptions. I will now further explore the use of such models to more systematically analyze the scenarios Neil and Winship describe.

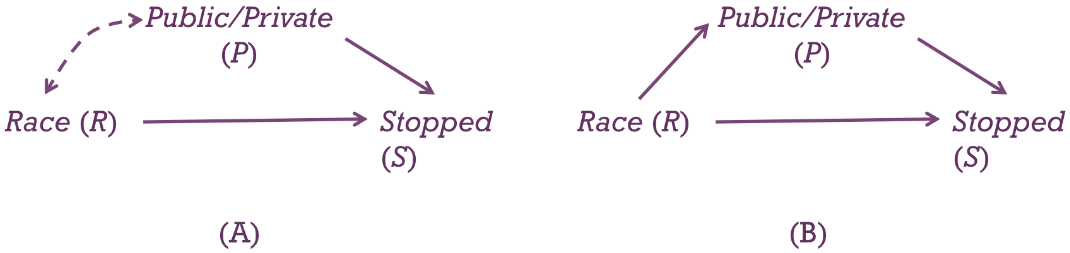


Figure 3: Two possible models for the public space scenario

4 Using Causal Models to Interpret Statistics

Let's begin by considering the scenario in which the difference in stops results from a racial disparity in the numbers of white and black individuals utilizing public spaces. Here we are supposing that the only factor that makes a difference in whether someone is stopped is whether they are in a public or private space. One possible model for this is that given in figure 3(A). This model treats race as a correlated non-cause of being in a public space. If we treat the question of whether there is discrimination here as one of whether race is a cause of being stopped, then the causal model tells us that this effect is only identified when one conditions upon being in a public space.

Now imagine that race is in fact a cause of whether someone is in a public space. Perhaps race made a difference in terms of the neighborhoods in which individuals could get a mortgage, and black families ended up in urban neighborhoods where public space use is more common. If so, then public space usage is now a mediator between race and being stopped, as in figure 3(B). There are two ways one might employ this model for evaluating discrimination. First, one might claim that because race is a cause of public space usage, the decision by police to stop people in public spaces is discriminatory. As such, the correct measure of discrimination is the total effect of race on being stopped and one should not condition upon the mediator in measuring this effect (i.e. if there are no omitted common causes $P(stopped|do(race)) = P(stopped|race)$). Alternatively, if one believes that the racial distribution of public spaces played no role in the policies of where to stop people, one might choose to measure discrimination as the influence of *race* on being *stopped* along the direct path. This is called the *natural direct effect* (Pearl, 2001) and it corresponds to the change in one's probability of being stopped that would result, were they to be of a different race, but where to have the same public space usage they presently do. Concretely, we might ask whether a black individual who is stopped in a public space would have been less likely to be stopped were they white, but still in a public space. If yes, then the natural direct effect of being black (as opposed to white) is negative, and this would on the face of it be good evidence for discrimination.

Which of the two models is appropriate in a particular scenario turns out to be a subtle matter going beyond the debate surrounding causally modeling race (Weinberger, ming). One issue is that even if, as a matter of fact, the correlation between race and public space usage is due to a common cause (e.g. parental race), the explicit use of this correlation in choosing where to patrol could itself be a form of intentional discrimination. In such a case, the model in 3(B) seems to better account for the way in which race influences being stopped via the racial distribution of public spaces, even though we've stipulated that 3(A) gives the data generating process. Sorting this out properly would require building models significantly more complex than the two considered, in order to explicitly represent the influence of the correlation on decision making. Because the model in 3(B) allows for an intuitive separation between the "direct" influence of race via police perception of racial signals and the "indirect" effect of race on who is stopped, it would seem to be more useful. The issue of how to justify its use as an acceptable shorthand for a more accurate model calls for its own paper.

For our purposes, the key point is that once one has linked a discrimination hypothesis to a quantity in a causal model, one then has a theoretical framework for specifying which variables one must (and must not) condition upon order to identify that effect. Additionally, the identification of a causal quantity comes with the theoretical guarantee that there will not be reversals of the sort that arise in

Simpson's paradox. If one believes that race causally raises the probability of being stopped, but it fails to do so once one has partitioned the population into subpopulations, then either the subpopulation-specific probabilistic relationship fail to identify causal effects, or one's initial presupposition about there being an effect in the population is false.

The advantages of the causal modeling approach become especially salient when contrasted with Neil and Winship's analysis. Their analysis of the public-space example is that in choosing a particular benchmark test, one is faced with a "denominator" problem. For instance, if one uses the population-based benchmark, one takes as the denominator the relative proportions of blacks and whites in the population. But because, by hypothesis, the police are only taking into account public-space usage in determining who to stop, one should instead consider these proportions within public and non-public spaces, yielding a different denominator. Neil and Winship criticize this benchmark as employing the "wrong denominator" (2019, p. 79), but provide no normative account of what makes a denominator the right one. Simply saying that individuals must be "similarly situated" does not suffice, since as one conditions on additional "non-race" variables, the denominator will keep on changing, and short of holding fixed all non-race variables they can provide no guarantee that the probabilistic relationships will not keep changing upon conditioning. Their claim that "the denominator that is chosen must accurately reflect the appropriate risk set" (79) is not very illuminating. Alternatively, they suggest that an expert using the population-based benchmark is implicitly assuming that "police are going after people randomly" (79) (as opposed to taking into account location). While it is certainly true that benchmark tests can only provide evidence of discrimination given additional assumptions, the idea that particular choices of denominators can be linked to particular substantive assumptions makes it sound as if the statistical measures used in these tests speak for themselves. Benchmark tests compare proportions across populations, and Simpson's paradox reveals that such proportions will not, in general, be invariant to conditioning on additional variables. As I'll argue in the following section, the task at hand is not simply one of labeling certain statistical quantities as being the right ones to consider, but rather a more general theory of when statistical quantities are appropriately invariant to serve as measures of substantive quantities.

Neil and Winship claim that not only the the population-based benchmark test, but also the criminal-based one faces a "denominator problem". According to them, "using a crime denominator reflects an implicit assumption, likely made unknowingly, that police only stop criminals" (2019, p. 81). Let's reevaluate this from the causal modeling perspective. If one were to treat criminality as a correlated non-effect of race, then the justification of conditioning on criminality is to correct for confounding. A more interesting model is one in which race is modeled as a cause of criminality (the explanation of how prior racial discrimination might promote criminality is undoubtedly complex, but here these details are being black boxed). In such a model, criminality would be a mediator from race to being stopped. If we temporarily assume that our model omits no common causes, then it will be possible to identify the natural direct effect of race on being stopped by conditioning on the mediator (in a particular way). The justification for this conditioning is not an assumption that police only stop criminals, but rather that doing so is relevant for identifying the desired effect. The justification for focusing on the direct effect here is the assumption that even if race does influence criminality, it is legitimate for police to take criminality into account in deciding who to stop, but that race should have no further relevance once criminality status is taken into account.

Although in the absence of confounding, one can identify natural direct effects by conditioning on the mediator, this is only because under such favorable conditions the causal probability distribution $P(stopped|do(criminality))$ equals the conditional probability distribution $P(stopped|criminality)$. If we suppose that there is an unmeasured common cause of the mediator and the outcome – for instance, perhaps socioeconomic status is a common cause of criminality and being stopped – then identifying this effect is much more difficult. One would not be able to infer the effect of criminality on being stopped just by considering the probability of the former conditional on the latter, and any attempt to gain causal information simply by comparing stop rates within the criminal (or the non-criminal) subpopulation will not yield meaningful information. Such confounding can make the desired causal quantities very difficult to identify in practice, but at least causal models provide the tools for specifying the sets of assumptions that license causal identification. While Neil and Winship want to make a direct inference from the use of a benchmark that adjusts for criminality to an implicit assumption that the police are only stopping criminals, determining whether the benchmark is relevant is not so direct. Conditioning on criminality *might* yield a statistical comparison that is relevant to evaluating discrimination, if one makes some extremely favorable assumptions about the modeled scenario. But these assumptions – and more generally the different assumptions one might make about police behavior – are causal assumptions that can be expressed using a causal model. They cannot be read off of the statistical quantities being employed.

Without clarifying the quantity that they are seeking to measure in a particular scenario, Neil and Winship cannot give an account of what it means for a particular benchmark statistic to be "biased". They cannot just say that the target quantity is the probabilistic relationship between race and being stopped among similarly situated individuals, since to render the notion of "similarly situated" to be usable, we need a basis for believing that this probabilistic relationships will not disappear upon partitioning. If the presence of discrimination is taken to correspond to an effect in a causal model, then it is not required that individuals be similar in all respects, but only that they be grouped to be homogeneous with respect to a (de-)confounding variable set. Given such an effect, conditioning on additional variables will still be useful for measuring latent heterogeneity and for deriving more precise effect measurements for individuals or subpopulations, but such conditioning will not make the original effect disappear.

5 Evidence and Assumptions

Pollock (1987) provides a useful distinction between two distinct ways that a new piece of evidence can undermine one's belief in some proposition P. Suppose that proposition P is that it will rain tomorrow, and I believe this based on the testimony of a colleague. A *rebutting defeater* for P is a new piece of evidence for its being false. For instance, perhaps I look at a reliable weather app that says that tomorrow there is no chance of rain. In contrast, an *undercutting defeater* is evidence that my original evidence was unreliable. Suppose I learn that my colleague was conducting a study in which she randomly tells people that it will or will not rain based on the outcome of a coin flip. I am now no longer justified in believing that it will rain tomorrow. The reason for this is not that I now have evidence for the proposition $\neg P$ – I cannot infer that it *won't* rain tomorrow – it's just that my original justification is undermined.

Now consider a case in which one originally observes that police stop minorities at the same rate as non-minorities within a population and infers that there is no discrimination. One then notices that within both the criminal and non-criminal subpopulations, minorities are stopped at a higher rate, and concludes that there is discrimination. One might be inclined to describe this reasoning process as follows. The first observation provided some prima facie evidence for the absence of discrimination, and this evidence was then rebutted by the further information about the subpopulations. I believe we should reject this description. Once one is familiar with Simpson's paradox, it becomes evident that, from the perspective of statistics alone, there is no reason at the outset to assume that the probabilistic relationships among variables in a population will be preserved when one partitions those populations. The new information does not provide counter-evidence to one's initial belief, but rather reveals that one was never justified in holding it in the first place. One should not have taken the statistical evidence by itself to say anything at all about discrimination, except when coupled with further assumptions about the scenario being statistically modeled. As argued, causal models provide one way of systematically representing these assumptions.

Perhaps an analogy will help. Imagine experts are aiming to compare the height of two objects, but the only evidence they have concerns pictures of their shadows. Expert A uses a picture in which the two shadows are of the same length to argue that the two objects are of the same height. Expert B responds by submitting a separate picture in which the lengths of the shadows differ. This certainly casts doubt about the validity of expert A's conclusion. But it would be wrong to respond to this by asking which expert had the correct evidence. Because shadows only provide reliable information about the heights of the objects given information about the position of the light sources relative to the objects, one cannot draw any conclusions from the shadows about the objects absent a hypothesis about the setup. Similarly, once one learns that probabilistic comparisons such as those employed by benchmark tests do not measure invariant quantities, one requires a theory about when such quantities can be meaningfully interpreted. The point is not that benchmark tests are always wrong. In the same way as same-length shadows may indicate same-length objects given the assumption that there is a single light source standing at the same angle in relationship to both objects, it could be the case that the reason police stop minorities at a higher rate than non-minorities is that the police are discriminating. But to say whether this is the case, we need to move beyond the benchmarks themselves to evaluate the hypotheses that would make this so.

These considerations help motivate my objection to Neil and Winship's approach of criticizing particular benchmarks as failing due to their making implicit assumptions that are being violated. At a very general level, they are correct that the validity of benchmark tests depends on assumptions about the relevant case. But these assumptions cannot be read off from the mathematical operations used to derive the benchmarks. Instead, one requires a separate framework for specifying when a comparison of proportions across populations will yield meaningful information that is relevant to evaluating discrimination and when one is simply considering an artifact of the particular variables one chooses to model.

6 "Similarly Situated" Revisited

The question of what counts as evidence for discrimination is not an idle concern. Within the American legal system, a defendant claiming discrimination by police must bring evidence of police discrimination in order to advance to the "discovery" stage that would allow for a more thorough investigation. Following the Supreme Court decision *Armstrong v. United States* (1996), this evidence must involve a "credible showing of different treatment of similarly situated persons" of another race. As Siegler and Admussen (2020) emphasize, this has in practice created an insurmountable barrier towards being granted discovery, since police are not required to disclose their selection criteria and without such a disclosure one cannot determine who would count as "similarly situated".

The idea that one cannot determine who counts as similarly situated without information about the selection criteria used by police reinforces my earlier point that the similarly situated criterion cannot be interpreted purely statistically. What matters is not whether those who are and are not stopped are similar in all respects, but rather whether, within the class of individuals satisfying the purportedly legitimate criteria that the police use in choosing who to stop, race makes any further difference. Yet requiring that defendants present a well-established causal model in order to obtain discovery would clearly be too high a bar, since only through discovery does one have a chance at obtaining the required evidence. To avoid this problem, Siegler and Admussen (2020) defend a statistical basis for granting discovery:

> Courts instead should look to whether the defendant has created a reasonable inference that a disparity exists. If a defendant can show that the police are targeting people of color at a rate greater than their representation in the general population, judges should grant discovery. (1048)

Although I endorse Siegler and Admussen's aim of lowering the evidential standard required for obtaining discovery, a purely statistical standard remains problematic. Disparities by themselves do not necessarily indicate discrimination and, as I've emphasized, one should hesitate to attribute to them any significance absent substantive background assumptions about the modeled scenario. Instead of abandoning causal considerations, a better strategy would be to make it relatively easily for the defense to submit a candidate causal model that favors their claim. The prosecution might then be given the opportunity to submit their own causal model specifying some of the selection criteria, and if they decide not to do so then the defense would be allowed to make their case using their preferred model. Even if, however, the prosecution submits their own model, this by itself is an improvement, since under the status quo they have no incentive to disclose their selection criteria. This, of course, is only a sketch of a procedure. But the main takeaway is that once one acknowledges that statistical evidence by itself cannot differentiate between discriminatory and non-discriminatory explanations of behavior, one should build the use of causal assumptions into every evidentiary stage of the process.

There is an increasingly popular critique of using causal methods for racial discrimination, on which such methods presuppose that one can "isolate" the effects of race understood as a "solid-state" independent of its social role (Kohler-Hausmann, 2018). The preliminary analysis offered here in no way presupposes such an objectionably narrow view of racial discrimination. It is compatible with the analysis that police who do not intentionally track race but instead rely on factors that are correlated with race in determining where and who to search should still count as racially discriminating.[2] Which

[2] Such an analysis would go beyond existing legal doctrine, which requires that one establish both discriminatory intent and discriminatory effect.

factors police may take into account without discriminating is a difficult normative and legal question whose answer may vary across contexts. The analysis here makes only the minimal assumption that among the criteria that police can legitimately use, their legitimacy is not automatically undermined by their being correlated with race. Given this, statistics alone will be unable to detect the presence of discrimination, and further modeling tools are required.

7 Conclusion

Causal models are typically defended on the grounds that they are required in order to predict the outcomes of interventions. The discussion here points to a distinct – and underappreciated – role for these models. Even when one's main interest is statistical interpretation, causal models can be useful for distinguishing between informative and non-informative statistical quantities. The phenomenon of Simpson's paradox is particularly helpful for seeing why this is so. While learning that one variable raises (or lowers, or is irrelevant to) the probability of another in a population might be seen as providing important information about the relationship between these variables, the paradox reveals that such probabilistic relationships will not in general be invariant to partitioning based on additional variables. The lesson I derive from this is that instead of taking facts about probability raising etc. as providing evidence for one's discrimination hypothesis until proven otherwise, one should refuse to draw any conclusions from them until one can provide some basis for believing that the relevant facts are partition invariant. Causal models enable one to do so.

While there exists a growing literature on whether race is a cause (Weinberger, ming), there has been little discussion specifying the advantages of modeling race causally. The discussion in this paper presents a clear advantage. Nothing in this paper proves that one *must* use causal models to properly interpret statistics when evaluating discrimination. But I nevertheless have argued that the statistics themselves are not sufficient. They can only be used for evaluating discrimination given assumptions about the modelled scenario and causal models provide a general, rigorous, and flexible way for modeling these assumptions.

References

Bickel, P. J., E. A. Hammel, and J. W. O'Connell (1975). Sex bias in graduate admissions: Data from berkeley. *Science 187*(4175), 398–404.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 419–437.

Kohler-Hausmann, I. (2018). Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev. 113*, 1163.

Neil, R. and C. Winship (2019). Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology 2*, 73–98.

Pearl, J. (2001). Direct and Indirect Effects. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420.

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive science 11*(4), 481–518.

Ross, C. T., B. Winterhalder, and R. McElreath (2018). Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police. *Palgrave Communications 4*(1), 1–9.

Siegler, A. and W. Admussen (2020). Discovering racial discrimination by the police. *Nw. UL Rev. 115*, 987.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological) 13*(2), 238–241.

Sprenger, J. and N. Weinberger (2021). Simpson's Paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University.

von Kügelgen, J., L. Gresele, and B. Schölkopf (2021). Simpson's paradox in covid-19 case fatality rates: a mediation analysis of age-related causal effects. *IEEE Transactions on Artificial Intelligence 2*(1), 18–27.

Weinberger, N. (Forthcoming). Signal Manipulation and the Causal Analysis of Racial Discrimination. *Ergo*.