


SYMPOSIA PAPER

The Insufficiency of Statistics for Detecting Racial Discrimination by Police

Naftali Weinberger 

Munich Center for Mathematical Philosophy, Ludwig-Maximilian-Universität München, Munich, Germany

Email: naftali.weinberger@gmail.com

(Received 14 April 2023; revised 17 July 2023; accepted 30 August 2023; first published online 20 October 2023)

Abstract

Benchmark tests are employed when testing for racial discrimination by police. Neil and Winship (2019) emphasize that such tests are threatened by Simpson’s paradox, but they avoid analyzing the paradox causally. They consequently cannot elucidate the link between statistical quantities and discrimination hypotheses. Simpson’s paradox reveals that the statistics given by benchmark tests are not invariant to conditioning on additional variables. On this basis, I argue that benchmark statistics should not by themselves be taken to provide any evidence regarding discrimination, absent additional assumptions. Causal models can represent these assumptions.

1 Introduction

Consider a study revealing that police in Pittsburgh, Pennsylvania, search minority drivers at a higher rate than non-minority drivers. This would seemingly provide evidence of discrimination. But there are other possible explanations. Perhaps police make stops based on suspicious activities, and minority drivers disproportionately engage in such activities. Assuming that (a) suspicious activity is a legitimate basis for stops and (b) “suspicious” is not just a covert re-description of the driver’s race, the searches might not be discriminatory. Yet the raw statistics concerning the racial disparity in stops do not differentiate the discriminatory and non-discriminatory explanations. Legal and empirical studies of discrimination therefore employ *benchmark tests*, which involve conditioning on variables differentiating the relevant groups to determine what the group stop-rate disparity would be without discrimination. The idea is that once one accounts for the factors that legitimately could explain the disparity, any remaining disparity is evidence of discrimination.

As Neil and Winship (2019) note, benchmark tests are threatened by Simpson’s paradox (Simpson 1951; Sprenger and Weinberger 2021). To illustrate, it could be that police stop minorities and non-minorities at the same rate in Pittsburgh as a whole, but minorities are stopped at a higher rate within every district. Because probability

theory permits such scenarios, statistical claims involving comparisons of relative rates across populations—including benchmarks—will not be robust to conditioning on additional variables. Although Neil and Winship illustrate 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 discrimination statistics.

Here, I use Simpson's paradox to argue that benchmark statistics by themselves provide no evidence for or against discrimination, absent additional assumptions about the modeled scenarios. Causal models provide a framework for representing these assumptions. Whether they provide the best framework—and, if so, how they should be incorporated into legal practice—are pressing topics for future research.

2 Neil and Winship on benchmark tests

Neil and Winship (2019) illustrate the shortcomings of benchmark tests using a hypothetical population with 100,000 black individuals and 100,000 white individuals, in which the criminality rate is higher among blacks (15,000 vs. 10,000). Police stop 10,000 black individuals and 5,000 white ones, making the stop rate for blacks twice as high. Although one might suggest this reflects the higher criminality rate in the black population, that rate is only 50% higher than the rate in the white population, and thus it cannot account for the difference. This reasoning is captured by the *criminal-based benchmark test*, which statistically adjusts for criminality by comparing stop rates as a proportion of criminality rates in each population. Numerically: $\frac{10,000}{15,000} : \frac{5,000}{10,000} = 1.33:1$. Because this benchmark adjusts for the different criminality rates—revealing that even after adjustment, the black stop rate is higher—it plausibly provides evidence of discrimination. Yet Neil and Winship show that as one specifies further details about this population, there may be either no discrimination or a level of discrimination different from that suggested by this benchmark statistic.

Here's an illustration of why benchmark statistics may mislead (Neil and Winship 2019, 79). Imagine police only stop people in public spaces and that the distribution of public-space users consists of 40,000 blacks and 20,000 whites. Assuming police stop all individuals at a rate of .25, regardless of race, the number of stops will match those in the hypothetical population. But, by stipulation, police are not using race in choosing whom to stop. Neil and Winship analyze discrimination using the "similarly situated" criterion, which entails that when police do not differentiate among individuals who are otherwise similar except for race, they are not discriminating.¹ Accordingly, a benchmark statistic that might suggest discrimination no longer does so after specifying additional information.

Neil and Winship's (2019) analysis of this scenario reflects their general strategy for criticizing benchmarks. After pointing out that the criminal-based benchmark fails to reveal discrimination if black and white individuals are observed by police at different rates, they claim that users of the benchmark falsely presuppose that individuals of different races are observed at the same rates. More generally, they

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

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

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

present scenarios in which the degree of discrimination diverges from that of the benchmark, then fault those using the benchmark for assuming the scenario does not obtain. Although they are correct that benchmarks are only justified given substantive assumptions, their treatment of specific benchmarks as corresponding to particular assumptions suggests that the assumptions in question are statistical. Moreover, they deny that establishing discrimination requires causal inference (2019, 76). In what follows, I 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 in every subpopulation, where subpopulations are derived by conditioning on values of a variable. For example, in winter 2020, the fatality rate among COVID-19 cases 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 supports their being more likely to die, once one learns the person’s age, this relationship reverses (no matter the age). Such cases are consistent with probability theory. They are paradoxical in the sense of being perplexing rather than impossible.

Simpson’s paradox is just one challenge Neil and Winship raise for benchmark tests, but its proper analysis matters for all of them. Because benchmark tests compare proportions within populations partitioned using a set of variables, understanding how such proportions change as one conditions on additional variables is essential to their interpretation. In cases of Simpson’s paradox, the relationships in the population do not match those in the subpopulations. I now present a standard causal analysis of the paradox, highlighting common misunderstandings that also appear in Neil and Winship.

Table 1 compares the success rates of a medical treatment in different populations. In the whole population, the success rate is the same in both the treatment and control groups, so treatment (*X*) and success (*Y*) are uncorrelated. However, the treatment raises the probability of success in both the male (*M*) and female ($\neg M$) subpopulations (modeled as exhaustive). To see what is going on, first note that the



Figure 1. Causal graph for Simpson's (1951) example.

population has more females than males. Second, although the treatment raises the probability of success in both subpopulations, men are more likely to recover than females, even without the treatment.

Figure 1 causally represents the example. The arrows indicate that *gender* and *treatment* are causes of *success* (the quantitative relationship is unspecified). The dashed line between *gender* and *treatment* indicates a correlation, which is presumed not to result from treatment causing gender.

Determining whether the treatment causes success involves disentangling two ways that the former may be evidentially relevant to the latter. First, those receiving the treatment may have a higher success rate because the treatment causes success. Second, it may be merely that learning someone received the treatment provides evidence of their gender (due to the correlation) and that gender predicts one's chance of recovery independent of whether they receive the treatment. Experiments in which one *intervenes* to make the treatment uncorrelated with any potential influence of success (e.g., gender) help disentangle causal from merely evidential relevance. Formally, this distinction is marked by the operator $\text{do}()$, where $P(Y|\text{do}(X))$ 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 bridging causal and statistical assumptions is *identifiability*. Identifiability relates (i) a probability distribution, (ii) a causal graph, and (iii) a causal quantity (e.g., the magnitude of X 's effect on Y). A causal quantity is identifiable if and only if (iff), given a graph, one can uniquely determine its value from the distribution. To illustrate, the effect of treatment on success would *not* be identifiable if these variables had an unmeasured common cause, because even given the probability distribution, one could not determine the extent to which any correlation between treatment and success results from the causal relationship as opposed to the common cause.

In contrast, in figure 1, the effect of *treatment* (X) on *success* (Y) is identified conditional on *gender* (M), using the following formula:

$$P(Y|\text{do}(X)) = \sum_M P(Y|X, M) P(M). \tag{1}$$

This equation derives the effect of X on Y in the population from a weighted average of the conditional probabilities in the male and female subpopulations. Its significance is that although the noncausal conditional probability $P(Y|X)$ may differ arbitrarily from $P(Y|X, M)$ and $P(Y|X, -M)$, the causal conditional probability $P(Y|\text{do}(X))$ averages over conditional probabilities in the subpopulations (which here also correspond to the subpopulation-specific effects). It is therefore impossible for X to causally raise Y 's probability in the population, but not in any subpopulations.

Equation (1) applies only when X does not cause M . For a different case in which X does cause M —perhaps M is a blood chemical via which the treatment is effective—one should *not* condition on M when identifying the effect of X on Y . So whether one should

condition on M —and correspondingly, whether to consult the population or the M -partitioned subpopulations—depends on the causal model. Note that the causal model in figure 1 is statistically indistinguishable from one in which X causes M : any data generated by one model could have been generated by the other. This highlights the importance of causal information in determining whether to consider populations or subpopulations.

As we will see, although one should not condition on intermediate variables (“mediators”) in evaluating the *total* effect of X on Y along all paths, one does condition on mediators (in particular ways) when evaluating X ’s influence on Y along particular paths.

After presenting examples involving Simpson’s paradox, Neil and Winship explain:

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 . . . this problem can bias benchmark tests, as well as the outcome test, for stops and searches. (2019, 85)

This comment reveals two common misconceptions about the paradox (Sprengrer and Weinberger 2021, §3.2). First, the paradox does not show that aggregation over heterogeneous populations is problematic. Aggregation is not a problem when it results in an average effect, and the solution to the paradox is not always to *disaggregate* because one should not condition on mediators. Second, their comment regarding statistics that are “inconsistent with reality” suggests a concern about objectivity. But if one relies on conditional probabilities identifying genuine effects, Simpson’s paradox does not threaten the objectivity of causation. The subpopulation effects derived from conditioning can differ from the population effect, but this is because the subpopulations have different distributions of background factors, and the effects are therefore objectively different.

Finally, note the claim that the paradox can “bias” benchmark tests. To talk about a measurement being biased, one needs to specify which quantity one aims to measure. Neil and Winship (2019) do not do so. They write as if there is some comparison of stop rates across the populations that would establish discrimination, and 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 comparisons of rates or proportions across populations are not, in general, invariant to conditioning on new variables, absent further assumptions. I will now show how causal models represent those assumptions in discrimination contexts.

4 Using causal models to interpret statistics

Causal models should not be conceived as a substitute for normative theorizing about which criteria police may use. Nevertheless, given an account specifying legitimate search criteria, linking discrimination claims to effects in particular models enables one to define one’s measurement target, as I now illustrate.

Let’s begin with the scenario in which whether someone is stopped only depends on whether they are in public or in private. Panel A of figure 2 presents a plausible

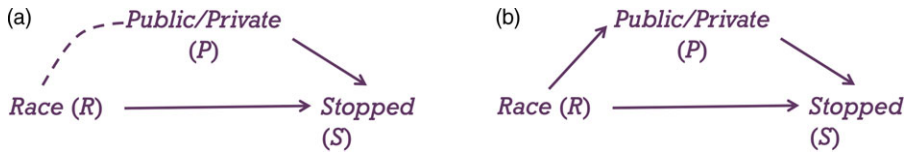


Figure 2. Two possible models for the public-space scenario.

model in which race is a correlated non-cause of public-space usage. If we link discrimination here to whether race causes being stopped, the model indicates that this effect is identified only when one conditions upon public-space usage.

Now suppose race does cause public-space usage. For example, perhaps race influenced where individuals could get mortgages, and black families ended up in urban neighborhoods where public-space use is more common. Public-space usage is then a mediator between race and being stopped, as in panel B of figure 2. There are two possible uses of this model for evaluating discrimination. One possibility is that because race causes public-space usage, deciding to stop people in public spaces is discriminatory. The correct measure of discrimination is then the total effect of race on being stopped, and one should not condition upon the mediator (i.e., absent common causes, $P(\text{stopped}|\text{do}(\text{race})) = P(\text{stopped}|\text{race})$). Alternatively, if the racial distribution of public spaces did not influence policies regarding where to stop people, one might link discrimination to the influence of race on being stopped along the direct path. This is the *natural direct effect* (Pearl 2001), which is the change in the probability of being stopped that would result if one were of a different race, but still had the same public-space usage. For instance, would a black individual stopped in public have been less likely to be stopped if they were white but still in public? If so, there is a natural direct effect, providing evidence of discrimination.

Which of these models is appropriate is a subtle matter. Even if the correlation between race and public-space usage results from a common cause (e.g., parental race), using this correlation in choosing where to patrol could itself be discriminatory. If so, the model in panel B of figure 2 seems to better account for the influence of race on being stopped via public-space usage, even though panel A of figure 2 gives the data-generating process. This tension between understanding causal models as representing data-generating processes and as modeling decision making calls for its own article.

Here, the key point is that linking discrimination hypotheses to quantities in a causal model enables one to specify which variables must (and must not) be conditioned upon. Moreover, causal quantities come with a theoretical guarantee against Simpson's paradox. If one believes that race causally promotes being stopped in a population, but it lowers the probability of being stopped in all subpopulations, then either the subpopulation-specific probabilistic relationships do not identify causal effects, or one erred in positing an effect in the population. Although conditioning can certainly yield subpopulation-specific effects differing from the population average, it cannot make a genuine effect disappear.

I have emphasized confounding due to common causes, but conditioning on a common effect of two variables can also hinder identifiability. This phenomenon of *endogenous selection bias* (Elwert and Winship 2014) is especially relevant to

discrimination contexts because sampling from a population that is homogeneous with respect to a variable (e.g., public-space usage, being stopped by police [Knox et al. 2020]) amounts to conditioning on it, and can thus bias an effect.² Causal assumptions are thus essential for evaluating the processes influencing one's statistical sample (Knox and Mummolo 2020).

Because Neil and Winship (2019) lack a framework for specifying which quantities measure discrimination, they cannot spell out what makes a benchmark "biased." They are correct that benchmarks are only justified given additional assumptions about the scenario, but their criticism of particular benchmarks as employing the "wrong denominator" (2019, 79) in deciding what to condition upon suggests that the assumptions can be read from the mathematical derivation of the statistical quantities. One might suppose that Neil and Winship do give an account of the measurement target via the "similarly situated" criterion, which says that there is discrimination when individuals are treated differently despite being similar in all respects except for race. Statistically, this corresponds to grouping people based on a list of non-race variables, then determining whether, within a particular homogeneous group, race influences the probability of being stopped. But which non-race variables must be included? Without additional assumptions, anything short of a maximal attribute set leaves open the possibility that probabilistic relationships may disappear upon further partitioning. Thus, the similarly situated criterion cannot be of practical use if interpreted statistically; it further requires the assumptions supplied by causal models.

5 Evidence and assumptions

Pollock (1987) distinguishes between two ways evidence can undermine one's belief in some proposition P . Suppose that P is that it will rain tomorrow, which I believe based on a colleague's testimony. A *rebutting defeater* for P is evidence of its falsity. For example, perhaps my weather app predicts no chance of rain. In contrast, an *undercutting defeater* is evidence that my original evidence was unreliable. For example, suppose I learn that my colleague was conducting a study in which she randomly told people it would rain. I have no new evidence *against* P —I cannot infer that it *won't* rain—it's just that my original justification is undermined.

Now imagine one observes that police stop minorities at the same rate as nonminorities 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 thus concludes that there is discrimination. One might be inclined to describe this reasoning as follows. The first observation provided *prima facie* evidence against discrimination, which was then rebutted by the further information. I submit that we should reject this description. Given Simpson's paradox, statistics alone provide no reason to assume that the probabilistic relationships among variables in a population will be preserved upon partitioning. The subpopulation information does not rebut one's initial belief but rather reveals that one was never justified in holding it. Statistics only provide evidence regarding discrimination when coupled with further assumptions, such as those embedded in causal models.

² Thanks to Clark Glymour for raising this point.

The question of what counts as evidence for discrimination is not an idle concern. Within the American legal system, a defendant claiming police discrimination must bring evidence to advance to the “discovery” stage, enabling a more thorough investigation. Following the U.S. Supreme Court decision *United States v. Armstrong* (1996), this evidence must involve a “credible showing of different treatment of similarly situated persons” (470) of another race. Siegler and Admussen (2020) argue that this creates an insurmountable barrier to being granted discovery because police need not disclose their selection criteria, and without these criteria, one cannot determine who counts as “similarly situated.”

That the similarly situated criterion cannot be applied without knowing the police’s selection criteria reinforces my point that it 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, among those satisfying the purportedly legitimate criteria employed by police, race makes any further difference. Yet requiring defendants to present a well-established causal model to obtain discovery would be too high a bar, because only through discovery can one obtain the required evidence. Siegler and Admussen (2020) therefore 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 reveal discrimination, and one should avoid making inferences from them absent background assumptions about the modeled scenario. Rather than abandoning causal considerations, it would be better to allow the defense to propose a causal model favoring their claim. The prosecution might then be allowed to submit their own model specifying some of the selection criteria, and, if they abstain, the defense would be entitled to their preferred model. Even if, however, the prosecution submits their own model, this by itself is an improvement, because currently, they have no incentive to disclose their selection criteria. This, of course, is only a sketch of a procedure. The main takeaway is that because statistical evidence alone cannot differentiate between discriminatory and nondiscriminatory explanations, one should build causal assumptions into every evidentiary stage of the process.

6 Conclusion

Causal models are typically promoted as enabling one to predict the outcomes of interventions. The discussion here highlights a distinct role. Even when interpreting statistics, causal models help disentangle informative from non-informative statistical quantities. Simpson’s paradox illuminates why this is so. Although a correlation between two variables might be taken to suggest a substantive relationship, the paradox reveals that probabilistic relationships will not in general be invariant to partitioning based on additional variables. Instead of taking probabilities as providing evidence regarding discrimination until proven otherwise,

one should not draw any conclusions without some basis for believing that the relationships are partition invariant. Causal models provide such a basis. Nothing in this article proves that one *must* use causal models when interpreting discrimination statistics. But statistics alone are not sufficient. They are only reliable given assumptions about the scenario, and causal models provide a general, rigorous, and flexible method for modeling these assumptions.

Acknowledgments. Thanks to John Jackson, Dean Knox, Julian Schüssler, and Chris Winship for feedback on earlier drafts. This article benefited from feedback from audiences at Cambridge, Turin, the 28th Biennial Philosophy of Science Association (PSA) Meeting, the 2022 Centre for the Experimental-Philosophical Study of Discrimination (CEPDISC) Conference on Discrimination, and online at the Penn/Rutgers Philosophy of Race Reading Group. Thanks to Alex Tolbert and Kareem Khalifa for organizing the symposium. This research was funded by a generous grant from the German Research Foundation (DFG) - Project number 441311834.

References

- Bickel, Peter J., Eugene A. Hammel, and J. W. O'Connell. 1975. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187 (4175):398–404. <https://doi.org/10.1126/science.187.4175.398>
- Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40 (1):31–53. <https://doi.org/10.1146/annurev-soc-071913-043455>
- Knox, Dean, Will Lowe, and Jonathan Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* 114 (3):619–37. <https://doi.org/10.1017/S0003055420000039>
- Knox, Dean, and Jonathan Mummolo. 2020. "Toward a General Causal Framework for the Study of Racial Bias in Policing." *Journal of Political Institutions and Political Economy* 1 (3):341–78. <https://doi.org/10.1561/113.00000018>
- Kohler-Hausmann, Issa. 2018. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination." *Northwestern University Law Review* 113 (5):1163–228.
- Neil, Roland, and Christopher Winship. 2019. "Methodological Challenges and Opportunities in Testing for Racial Discrimination in Policing." *Annual Review of Criminology* 2 (1):73–98. <https://doi.org/10.1146/annurev-criminol-011518-024731>
- Pearl, Judea. 2001. "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, edited by Jack Breese and Daphne Koller, 411–20. Burlington, MA: Morgan Kaufmann.
- Pollock, John L. 1987. "Defeasible Reasoning." *Cognitive Science* 11 (4):481–518. [https://doi.org/10.1016/S0364-0213\(87\)80017-4](https://doi.org/10.1016/S0364-0213(87)80017-4)
- Ross, Cody T., Bruce Winterhalder, and Richard McElreath. 2018. "Resolution of Apparent Paradoxes in the Race-Specific Frequency of Use-of-Force by Police." *Palgrave Communications* 4 (1):1–9. <https://doi.org/10.1057/s41599-018-0110-z>
- Siegler, Alison, and William Admussen. 2020. "Discovering Racial Discrimination by the Police." *Northwestern University Law Review* 115 (4):987.
- Simpson, Edward H. 1951. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 13 (2):238–41. <https://doi.org/10.1111/j.2517-6161.1951.tb00088.x>
- Sprenger, Jan, and Naftali Weinberger. 2021. "Simpson's Paradox." In *The Stanford Encyclopedia of Philosophy, Summer 2021 ed.*, edited by Edward N. Zalta. Stanford: Stanford University Press.
- U.S. Supreme Court. 1996. *United States v. Armstrong*, 517 U.S. 456.
- von Kügelgen, Julius, Luigi Gresele, and Bernhard 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. <https://doi.org/10.1109/TAI.2021.3073088>