OUTCOME TESTS FOR POLICIES

CHANGHWA LEE[†], MALLESH M. PAI[‡], AND RAKESH VOHRA[†]

THIS VERSION: OCTOBER 30, 2021

– Click here for the latest version –

ABSTRACT: The marginal outcomes test (Becker (2010)) has become a 'go-to test' of (un-)fairness/ disparate impact in classification or allocation settings. We consider settings with two key properties: (1) the underlying attribute of the agent being classified is strategically chosen by the agent, and (2) the adjudicator/ institution commits to a rule/ *policy*, taking into account strategizing by the agent. In this setting we show the outcome test is misspecified: the optimal rule will result in different marginal outcomes across demographics, even in the absence of any discriminatory motive for the principal. We derive a correctly specified test in such a setting. The test statistic requires estimation of both marginal and average outcomes—the latter portion captures the effect on agents' incentives. Under additional assumptions we identify the direction of misspecification for the classical marginal outcomes test.

KEYWORDS: marginal outcome tests, discrimination.

JEL CLASSIFICATION: D63, D82, K40

Extended Abstract

There is much interest in evaluating the "fairness" of various socioeconomic institutions, e.g. criminal justice, access to employment/credit/education etc. In practice, this often boils down to focusing on a specific binary decision,[1] and comparing if this differs across various demographics e.g. black vs white defendants, male vs female job applicants. Within the economics literature, and more generally, the "gold standard" is the *marginal outcome test*, originally due to Becker (2010). A failure of this test is interpreted as evidence of discrimination by the decision maker (see e.g. Hull (2021), Bohren, Haggag,

[†] DEPARTMENT OF ECONOMICS AND DEPARTMENT OF ELECTRICAL & SYSTEMS ENGINEERING, UNIVERSITY OF PENNSYLVANIA

[‡] DEPARTMENT OF ECONOMICS, RICE UNIVERSITY

[1] For example a judge choosing whether to acquit/ convict a defendant, a bank choosing whether or not to extend a loan to a loan applicant, an employer deciding whether or not to employ a job candidate.

Imas, and Pope (2019)). It has been applied to a wide variety of settings.[2] In this paper, we revisit the question: when is the marginal outcome test valid? We identify a natural class of settings of interest where a "fair" principal would choose a rule that fails the marginal outcome test, and identify the correct test for such settings. Specifically, these are settings where the principal is choosing a *policy*, and agents are responding *strategically* to the chosen policy.

To fix ideas, let us first outline the model that the marginal outcome test implicitly assumes, focusing on the example of checking for racial bias in traffic stops by state troopers for contraband. A set of motorists each has a payoff-relevant attribute that is not directly observed by the decision maker (whether or not they are carrying contraband). The decision maker observes information about the motorist (including their race) and makes a binary decision on whether to interdict. Once the decision is made, this attribute is observed (i.e. upon conducting a traffic stop, the trooper learns whether the motorist was carrying contraband). The null hypothesis of no discrimination is that conditioned on being *marginal*, i.e. conditioned on the decision maker being indifferent, the distribution of outcomes should be similar across races—after all, ceteris paribus, a decision maker should be indifferent at roughly the same rate of successful interdiction. Differences are either the result of a preference by the decision maker to pull over e.g. black drivers at a higher rate ("taste-based discrimination") or of an incorrect statistical model that causes the decision maker to over-estimate the risk of (marginal) black drivers ("incorrect statistical discrimination"). The underlying economic logic of the test is clear and uncontroversial, and therefore seemingly universally applicable.[3]

Formally, we show that marginal outcome tests may fail when the outcome of the agent is not exogenously determined, but instead depends on a strategic choice made by the agent (e.g. in our running example, the agents choose whether or not to carry contraband). In particular, suppose the decision maker chooses and commits to a decision policy *a priori*, and the agent understands this policy at the time of their own choice. In the language of Game Theory, the decision maker is a Stackelberg leader, or, equivalently, in the language of mechanism design, the decision maker has commitment. The agent's choice is thus based on a cost-benefit calculation given decision maker's policy (e.g. both the benefits of carrying contraband, and the associated risk of being apprehended). Therefore, the decision maker announces a policy that optimizes an objective function, taking into account that agents will respond to the underlying policy.

Our main positive result shows how to test for discrimination in such settings. In particular, the principal in such settings will design a policy that accounts for how it affects

[2] Some notable examples include: in the context of lending (Ferguson and Peters, 1995), judicial decision making (Arnold, Dobbie, and Yang (2018), Alesina and La Ferrara (2014)), traffic stop/ search decisions (Knowles, Persico, and Todd, 2001; Anwar and Fang, 2006; Antonovics and Knight, 2009) etc.

[3] Operationally, one still needs to (correctly) identify the marginal agent which can be difficult in practice. The marginal outcomes test may also fail in richer models, see e.g. Canay, Mogstad, and Mountjoy (2020) which we discuss in further detail below.

the choices of the agents. If the different groups have differences in how they respond to policies, then the optimal policy must account for this.

Settings that satisfy the desiderata we describe are easily motivated in practice. The idea that agents' relevant choices may be strategic and may respond to policy choices of the decision maker is of course standard in economics, and has long been considered in related settings (e.g. the design of affirmative action policy, see e.g. Coate and Loury (1993), Foster and Vohra (1992) or Fryer Jr and Loury (2013)) but largely absent from the literature on evaluating fairness. That the decision maker is a Stackelberg leader, as hinted at above, can be thought of as a policy choice by the underlying institution. For example, in the case of traffic stops it may amount to guidance issued by the leadership directing troopers on whom to stop. Similarly, as decision-making gets increasingly automated by the use of computers/ machine learning/ AI, it may amount to a choice of algorithm by the institution (e.g. the use of automated rules to determine who gets issued a loan in a banking setting, or the use of resume scanning software by an employer to determine which applicants get called back for an interview).

References

ALESINA, A., AND E. LA FERRARA (2014): "A test of racial bias in capital sentencing," *American Economic Review*, 104(11), 3397–3433.

ANTONOVICS, K., AND B. G. KNIGHT (2009): "A new look at racial profiling: Evidence from the Boston Police Department," *The Review of Economics and Statistics*, 91(1), 163–177.

ANWAR, S., AND H. FANG (2006): "An alternative test of racial prejudice in motor vehicle searches: Theory and evidence," *American Economic Review*, 96(1), 127–151.

ARNOLD, D., W. DOBBIE, AND C. S. YANG (2018): "Racial bias in bail decisions," *The Quarterly Journal of Economics*, 133(4), 1885–1932.

BECKER, G. S. (2010): *The economics of discrimination*. University of Chicago press.

BOHREN, J. A., K. HAGGAG, A. IMAS, AND D. G. POPE (2019): "Inaccurate Statistical Discrimination: An Identification Problem," Working Paper 25935, National Bureau of Economic Research.

CANAY, I. A., M. MOGSTAD, AND J. MOUNTJOY (2020): "On the use of outcome tests for detecting bias in decision making," Discussion paper, National Bureau of Economic Research.

COATE, S., AND G. C. LOURY (1993): "Will affirmative-action policies eliminate negative stereotypes?," *The American Economic Review*, pp. 1220–1240.

FERGUSON, M. F., AND S. R. PETERS (1995): "What constitutes evidence of discrimination in lending?," *The Journal of Finance*, 50(2), 739–748.

FOSTER, D. P., AND R. V. VOHRA (1992): "An economic argument for affirmative action," *Rationality and Society*, 4(2), 176–188.

FRYER JR, R. G., AND G. C. LOURY (2013): "Valuing diversity," *Journal of political Economy*, 121(4), 747–774.

HULL, P. (2021): "What Marginal Outcome Tests Can Tell Us About Racially Biased Decision-Making," Discussion paper, National Bureau of Economic Research.

KNOWLES, J., N. PERSICO, AND P. TODD (2001): "Racial bias in motor vehicle searches: Theory and evidence," *Journal of Political Economy*, 109(1), 203–229.