PA

The Essential Role of Statistical Inference in Evaluating Electoral Systems: A Response to DeFord *et al*.

Jonathan N. Katz[1], **Gary King**[2] **and Elizabeth Rosenblatt**[3]

[1] *Kay Sugahara Professor of Social Sciences and Statistics, California Institute of Technology, DHSS 228-77, 1200 East California Boulevard, Pasadena, CA 91125, USA. Email: jkatz@caltech.edu, URL: jkatz.caltech.edu*
[2] *Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA. Email: King@Harvard.edu, URL: GaryKing.org*
[3] *Post-BA Affiliate, Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA. Email: ERosenblatt@alumni.harvard.edu, URL: elizabethrosenblatt.com*

Abstract

Katz, King, and Rosenblatt (2020, *American Political Science Review* 114, 164–178) introduces a theoretical framework for understanding redistricting and electoral systems, built on basic statistical and social science principles of inference. DeFord *et al*. (2021, *Political Analysis*, this issue) instead focuses solely on descriptive measures, which lead to the problems identified in our article. In this article, we illustrate the essential role of these basic principles and then offer statistical, mathematical, and substantive corrections required to apply DeFord *et al*.'s calculations to social science questions of interest, while also showing how to easily resolve all claimed paradoxes and problems. We are grateful to the authors for their interest in our work and for this opportunity to clarify these principles and our theoretical framework.

Keywords: redistricting, representation, fairness, statistical inference

1 Overview

The goal of Katz, King, and Rosenblatt (2020) is to "deploy a crucial principle of statistics that is often ignored in this literature—defining the quantities of interest rigorously and separately from the measures used to estimate them" (p. 2). Only by separating quantities of interest, can one meaningfully evaluate their empirical measures, make claims vulnerable to being proven wrong, or offer appropriate measures of uncertainty. Statements that do not adhere to this fundamental statistical principle may still offer some descriptive uses, but are not of direct use for the inferential, causal, or applied goals of the social sciences. Valid scientific inference requires well-defined quantities of interest, estimators with known statistical properties, clear assumptions, and accurate uncertainty estimates.

Our article applies this statistical principle to the social science concept of an *electoral system*—a set of rules that allocates legislative seats among candidates on the basis of citizen votes. Examples of such rules include plurality voting within a single-member district, the absence or presence of voter fraud, district boundary lines, rules for drawing the boundary lines, registration requirements, etc. Because the importance of a new electoral system rests solely on the consequences it may have for *future* elections to be held under its rules, past (and thus observed) election results may help with estimation but cannot define a reasonable notion of an electoral system's overall fairness.

DeFord *et al.* (2021) does not separate quantities of interest from empirical measures, and either ignores future elections or assumes that past election outcomes are exactly equal to future results. The article includes no definitions of quantities of interest, separation of these quantities from empirical measures, uncertainty estimates, estimators, or formal statistical properties of proposed measures. The article even includes claims referring to the lack of need for statistical

Political Analysis (2023)
vol. 31: 325–331
DOI: 10.1017/pan.2021.46

Published
2 December 2021

Corresponding author
Gary King

Edited by
Lonna Atkeson

inference, uncertainty, estimators, understanding of future elections held under existing redistricting plans, and assumptions essential to inference. For example, the article is excited to note "the standard mean–median score…relies on no swing assumption at all!" Indeed, no assumptions are needed to make certain descriptive calculations about previous elections, but inferential assumptions (such as the nature of partisan swing, among others) are essential for estimating features of electoral systems of interest to social scientists or of importance to the general public. The article's key "characterization theorem" discretizes a result first given in Rosenblatt (2017, pp. 15–24), that the integral of a vote distribution equals the seats–votes curve (a cumulative density function), and then repeats the result from the literature that "skewness in the [vote] distribution becomes partisan bias in the seats–votes curve" (King 1989); it does not reference the future or counterfactuals.

Choosing this noninferential path enables DeFord *et al.* (2021) to calculate many interesting descriptive statistics but rules out learning from the resulting calculations about the fairness of electoral systems that by definition depend on unobserved characteristics of the future. In this regard, every such deterministic claim in the article—using jarring terms to a social scientist or statistician about "mathematical guarantees," "deterministic" results, "locked out of Congressional representation," or "elementary mathematical manipulation[s]"—has no bearing on a measure's potential usefulness in evaluating the fairness of electoral systems or redistricting plans. As Wasserman (2012) put it in the context of a similar situation, "I don't know of a single statistician in the world who would analyze data this way."

We first discuss the ideas in DeFord *et al.* (2021) in hypothetical data and then offer corrections to their approach to provide valid analyses of real elections and redistricting plans.

2 Resolving Apparent Paradoxes in Hypothetical Data

The central quantity of interest in the literature and in Katz *et al*. (2020) is the seats–votes curve $S(V)$, the main features of which are partisan bias (defined as the deviation from "partisan symmetry," which simply formalizes the concept of treating others as you want to be treated) and electoral responsiveness, both defined for all possible values of the average district vote V in future elections held under the same electoral system.[1] Partisan bias is defined (in Definition 2, p. 3) from the seats–votes curve as $\beta(V) = \{S(V) - [1 - S(1 - V)]\}/2$, which is the proportion of the seats the Democrats receive more than the Republicans if (in different future elections under the same electoral system) they each had received the same $V \in [0, 1]$ proportion of votes. Our article also discusses the statistical properties of various proposed estimators of $\beta(V)$, and the weaknesses of more limited quantities, such as $\beta(0.5)$, and their estimators.

DeFord *et al.* (2021) considers two patterns that could occur in historical data to be "paradoxes." From the perspective of the social sciences, this view results from three methodological mistakes: (1) not defining a quantity of interest separately from its measures, (2) not providing statistical estimators with known properties or uncertainty indicators, and (3) examining only limited descriptive measures rather than the full seats–votes curve. These mistakes have two major theoretical consequences, which we address first before discussing each pattern.

2.1 Theoretical Consequences

To illustrate the first consequence of these mistakes, consider the widely discussed complaints about bias in the electoral college: Democrats claim that they need more votes than Republicans

1 Define the seats–votes function as $S(V \mid \mathbb{P}, \mathbb{E}, X) \equiv S(V)$ (the expected proportion of seats for the Democrats given the average district vote V, a populace $\mathbb{P}$, electoral system $\mathbb{E}$, and other fixed characteristics X), and the *seats–votes curve*, as the seats–votes function defined for all possible values of the average district vote $V \in [0, 1]$ in future (actual or hypothetical) elections to be held under the same electoral system. As Assumption 1 of Katz *et al*. (2020) clarifies, a coherent seats–votes curve must be defined independently of any observed election outcome.

to win the White House; in other words, this electoral system is claimed unfair, because it is not symmetric. To evaluate this claim requires estimating the extent of any deviation from symmetry in the set of future elections held under the current electoral system at issue, along with proper uncertainty estimates over future votes and electoral college delegates. In other words, to evaluate the electoral system—the partitioning of America into states and the winner-take-all rule within each in the contest for delegates—we must treat it as fixed, whereas votes by Americans in future elections are unknown. This of course follows the standard Bayesian paradigm, conditioning on what we know (the electoral system and prior votes) and modeling probabilistically what remains unknown (the outcome of future elections). In contrast, in DeFord *et al.* (2021), the analysis is backward: treating future votes (which are unknown) as if they were fixed at past votes and the electoral system (which is known) as random. How much solace would it provide to those who view the electoral college as unfair to explain that (for example) the vast majority of other possible ways of districting America into states would produce bigger bias? This might be computationally interesting, but it is irrelevant to claims about the fairness of the electoral college. The same goes for DeFord *et al*.'s analyses of redistricting, to which we now turn.

The second consequence of the three mistakes is the article's implicit and unjustified normative argument that electoral system fairness depends solely on the characteristics of elections held *prior* to the implementation of a redistricting plan, rather than its future consequences. We give an example of this in the context of discussing the claimed paradoxes.

Pattern 1

DeFord *et al.* (2021) assumes the same hypothetical data generation process to study both patterns identified: a state heavily favoring the Republicans, four seats, an average district vote of $V = 0.25$, a fixed and equal turnout probability for every citizen in the state, orthogonality of district lines and citizen votes, and uniform partisan swing applying exactly (i.e., with no random error). We illustrate these two patterns with two corresponding redistricting plans. In redistricting *Plan 1*, the Democratic vote proportions are $V_i = \{0.10, 0.15, 0.20, 0.55\}$. Because the Democrats win one seat in this election under this plan, the largest number possible with $V = 0.25$, DeFord *et al*. make the normative assumption that this plan is (at least) fair to the Democrats. However, this observation about a set of past election results guarantees nothing about future elections to be held under this plan: In any future election, the Democrats could win 1, 2, 3, or all 4 seats. The claim that another redistricting plan (that was not implemented) would have been less favorable to the Democrats in a past election says nothing about what may happen when elections are run under the plan being evaluated. Fairness in an electoral system involves what may happen in (hypothetical or real future) elections held under the plan at issue.

DeFord *et al*. then use this normative standard for past elections and judges the empirical pattern to be paradoxical because, in the same data, $\beta(0.5) < 0$, a bias favoring Republicans.[2] We show the fallacy of this reasoning by rendering the seats–votes curve implied by the assumed data generation process (see Figure 1, top left panel). The observed average district vote V is represented by a black diamond at $V = 0.25$ horizontally and seat share $S(0.25) = 0.25$, which is one seat, vertically. The bottom-left panel gives the corresponding partisan bias for each possible value of the average district vote, $\beta(V)$. And indeed, we can see Republican bias at $V = 0.5$, $\beta(0.5) = -0.25$, which is indicated by the blue line dropping below the dotted line at zero.

DeFord *et al*. treats measures of $\beta(0.5)$ as fixed features of the observed data. In contrast, to understand what $\beta(0.5)$ means to social scientists, consider many elections in which the

2 DeFord *et al.* (2021) discusses three measures corresponding to $\beta(0.5)$, two with the weaknesses described in our article and the third, what they call the "partisan Gini score," which has the same weaknesses as the other two, in addition to being offered without either an interpretable metric (i.e., in terms of the number of legislative seats unfairly gained) or known statistical properties.

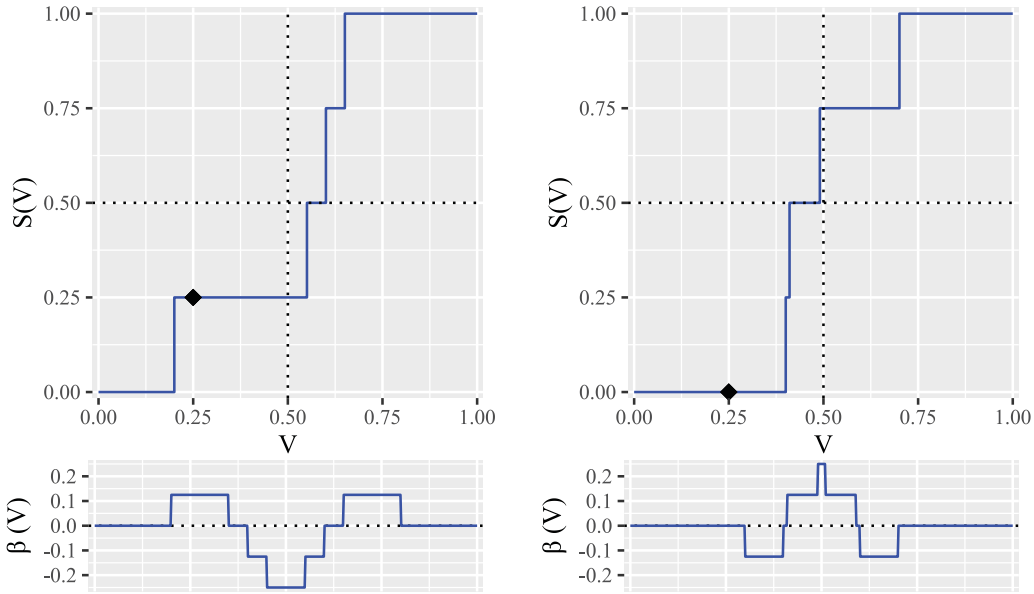

Figure 1. Plan 1 (left panel) and Plan 2 (right panel).

Republicans under Plan 1 receive half the votes. According to the DeFord *et al*. data generation process, if the parties split the votes equally in many future elections held under this redistricting plan, the Republicans would (unfairly) receive, on average, three seats and the Democrats would receive one. Thus, if the parties knew they would be splitting the vote equally, the Democrats would oppose Plan 1, because it is heavily biased toward the Republicans. Of course, from a political science perspective, it makes little sense to assume that the parties would split the vote equally when, in the one observed set of district election results, the Democrats only received a quarter of the average district vote. As such, this is one of the examples like that in our article of $\beta(0.5)$ being a highly limited measure.

To see the full pattern, look again to the bottom-left panel in Figure 1, but do not only look only at the $V = 0.5$ point. This particular seats–votes curve thus shows bias in favor of the Republicans in some ranges and the Democrats in others. The relevant range for a party is where that party expects to be in future elections. With average district vote observed only at 0.25, we would expect to see similar election results in future elections held under the same redistricting plan (i.e., assuming the electoral system, populace, and other election characteristics remain constant). And, as the bias graph shows, near $V \approx 0.25$, bias is above the line indicating Democratic bias.

This completely resolves the first claimed paradox: A Democratic gerrymander in a heavily Republican state would prefer to redistrict in a way that favors them in the previous (observed) election only as an indicator of an electoral system that favors them in future elections. If the gerrymander predicts V correctly, Plan 1 does exactly this (unless they misjudge the electorate; see Grofman and Brunell 2005).

Pattern 2

Consider now a second pattern, which DeFord *et al*. considers a paradox, because the Republicans win all four seats even though $\beta(0.5) > 0$, indicating bias toward the Democrats. This observation entails the normative assumption that the Republicans winning all four seats in a previous election indicates that the redistricting plan yet to be implemented is biased in favor of the Republican Party. In contrast, as above, the fairness of a particular redistricting plan requires information about what would happen in future elections held under this plan, not what prior election results might have been like under plans that were not used.

To be specific, this example has the same setup with the same data generation process as the first pattern, except that *Plan 2* has Democratic election proportions $V_i = \{0.05, 0.26, 0.34, 0.35\}$ with $\beta(0.5) = 0.25$. The mistaken interpretation here is the same as for Plan 1: under the DeFord *et al*. assumptions, all elections held under this plan in which the parties split the votes equally would in fact be unfair, because the Democrats always receive three seats and Republicans always receive one. If, instead, we expect future election results to be near the observed average, $V \approx 0.25$, the result would be approximately fair. If V were slightly larger, the electoral system would unfairly favor the Republicans. This also explains why the Republican decision-makers in Utah, one of the most Republican states in the nation, favors our "Symmetric Democracy" model of electoral systems rather than the "Symmetric Democracy with Minority Party Protection" model. Endogenous lawmaking by partisans seeking advantage is also not a good measure of fairness.[3]

As both analyses demonstrate, the fairness of a fixed electoral system with a given redistricting plan involves statistical inference about the appropriate quantity of interest (in this case, the full seats–votes curve) in future, as yet unknown, election results.

3 Principles for Evaluating Actual Electoral Systems

Researchers have leeway in choosing computationally convenient descriptive statistics for exploring observed data, as DeFord *et al.* (2021) ably demonstrates. However, for social scientists and others, learning about the consequences of redistricting plans (i.e., in future elections) requires explicitly defined quantities of interest, measures with known properties, and accurate uncertainty estimates. Thus, to reduce the disconnect between fields, we now offer six comments designed to help social scientists and those involved in redistricting to leverage this work for broader goals.

First, DeFord *et al*. finds "uncontested seats and variable incumbency effects" computationally inconvenient and so replace their U.S. House election data, in a study of House redistricting, with data from statewide U.S. Senate elections. Unfortunately, the high prevalence of split-ticket voting, differential candidate effects, the impact of redistricting on incumbents and uncontestedness, and district turnout differences make this kind of "data bait and switch" empirically unreasonable. Election returns for statewide offices can provide useful measures of underlying party support as an input to statistical analyses, but not as a replacement for direct observations on the election under study (Gelman and King 1994).

Second, DeFord *et al.* (2021, pp. 6 and 10 and footnote 16) makes an incorrect mathematical claim that the average district vote equals the statewide vote share for a party only in the highly restrictive and "idealized scenario that districts have equal numbers of votes cast (i.e., equal turnout)." In fact, ensuring the two are equal only requires the much more empirically reasonable assumption that turnout and vote share are uncorrelated (see Katz *et al.* 2020, Appendix A). An assumption of constant turnout across districts rarely fits real election data and should not and need not be assumed.

Third, the "seats–votes curve" is defined coherently only for all districts in an entire legislature. For example, Katz *et al*. (2020) study redistricting conducted by each state of all districts within its state house or senate. Proper computation or estimation of the seats–votes curve for the U.S. House, as attempted in DeFord *et al.* (2021) at the state level, should instead be performed

3 As a reminder, our article formalized two definitions of electoral system fairness widely accepted by scholars, adversaries on all sides of most redistricting litigation, and those who write legislation and constitutions: (1) For *competitive electoral systems* (i.e., where each party has a reasonable chance of winning a majority of votes statewide in future elections), our article defines "Symmetric Democracy," which requires partisan symmetry, nonnegative electoral responsiveness, and unanimity (see Katz *et al*. 2020, Definition 4, p. 4, for formal definitions), (2) For *noncompetitive electoral systems* (i.e., where one party is "confident of a statewide majority"), it offers "Symmetric Democracy with Minority Party Protection," which requires partisan symmetry, nonnegative responsiveness, and minority protection (see Katz *et al*. 2020, Definition 1 and Appendix).

nationally, even if their goal is to estimate the effects of congressional redistricting conducted in any one state on Congress as a whole. Conducting analyses under the implied assumption that House districts within one state somehow constitute a "legislature" is not reasonable.

Fourth, the idea of an electoral system in the literature and the law in almost every jurisdiction is that, once the rules are set, voters are able to cast their ballots however they please—even if the votes make a gerrymanderer's predictions wrong. Thus, the fairness of a redistricting plan cannot be judged solely on the basis of one election outcome without inferences about the future. When DeFord *et al.* (2021, p. 5) regards "as premises" that certain election outcomes indicate redistricting is biased toward certain parties, the results may be of descriptive value, but more work is required to use it to evaluate an electoral system.

Fifth, DeFord *et al*. focus on legislatures with small numbers of districts, using deterministic (uniform partisan swing) calculations and no uncertainty estimates. We show above that in these situations, the observed patterns are not paradoxes. Moreover, the patterns themselves will almost always disappear well within properly calculated confidence intervals when switching to more empirically appropriate stochastic uniform partisan swing calculations (see our Assumption 4, p. 9). Moreover, methods discussed in our article and commonly used in the literature and in litigation easily estimate all relevant quantities and uncertainty estimates from a single year of district-level election data. The claim in DeFord *et al.* (2021, footnote 12) that JudgeIt and other commonly used redistricting software packages require multiple elections is incorrect.

Finally, many of the issues in DeFord *et al.* (2021) result from its goal of a single, quantitative bright line rule for detecting gerrymandering, which is unusual in academia or the courts. In the literature on electoral systems, as in most academic fields, scholars avoid drawing conclusions from single sources of evidence or knife-edged quantitative thresholds and instead seek broader understanding from all available observable implications of a theory (King, Keohane, and Verba 1995, p. 28ff). Similarly, few legal tests adopted by the Courts employ bright line rules based on quantitative measures alone. Instead, quantitative tests are typically employed as part of multi-pronged factor tests. Examples include the use of the Herfindahl–Hirschman Index in judicial determinations on horizontal mergers, price inquiries in corporate fiduciary duty cases, evaluations of interest rates in evaluating debt contracts as unconscionable, inquiries in patent law (Olson and Fusco 2012), the calculus of negligence under Learned Hand's theorem (*U.S. v. Carroll Towing*, 159 F.2d 169 [2d Cir. 1947]), and the Gingles three-pronged test for the Voting Rights Act (*Thornburg v. Gingles*, 478 U.S. 30 [1986]). In each of these legal fields, quantitative measures are employed as one element of a holistic evaluation. In the many situations where partisan symmetry has been employed by courts, it is as one substantive prong in evaluating the fairness of districting plans alongside an evaluation of procedural fairness and other concerns.

4 Concluding Remarks

Learning about the empirical world requires inference—using facts you know to learn about facts you do not know. This is familiar to social scientists in making causal inferences, where the facts we do not know are the potential outcomes, such as what would happen if a given redistricting plan is or is not implemented (which reveals the problem with DeFord *et al*.'s claim that redistricting does not "require the reader to commit to this or any particular choice of nongerrymandered baseline"). It is also true in evaluating the fairness of electoral systems, where the facts we do not know are about features of future elections presently unknown but to be held under a fixed redistricting plan, and most other concerns to social scientists or public policymakers.

Acknowledgment

Our thanks to Moon Duchin and her coauthors for thoughtful comments on the first draft of this article.

References

DeFord, D., et al. 2021. "Implementing Partisan Symmetry: Problems and Paradoxes." *Political Analysis*, forthcoming.

Gelman, A., and G. King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38 (2): 514–554. https://j.mp/unifiedEc.

Grofman, B., and T. L. Brunell. 2005. "The Art of the Dummymander: The Impact of Recent Redistrictings on the Partisan Makeup of Southern House Seats." In *Redistricting in the New Millennium*, edited by I. M. Schwartz, 183–199. Lanham, MD: Lexington Books.

Katz, J. N., G. King, and E. Rosenblatt. 2020. "Theoretical Foundations and Empirical Evaluations of Partisan Fairness in District-Based Democracies." *American Political Science Review* 114 (1): 164–178. https://GaryKing.org/symmetry.

King, G. 1989. "Representation through Legislative Redistricting: A Stochastic Model." *American Journal of Political Science* 33 (4): 787–824. https://j.mp/2o46Gkk.

King, G., R. O. Keohane, and S. Verba. 1995. "The Importance of Research Design in Political Science." *American Political Science Review* 89 (2): 454–481. http://gking.harvard.edu/files/abs/kkvresp-abs.shtml.

Olson, D., and S. Fusco. 2012. "Rules versus Standards: Competing Notions of Inconsistency Robustness in Patent Law." *Alabama Law Review* 64 (3): 647–696.

Rosenblatt, E. M. 2017. "Judging Gerrymandering: Improving Methods for Measuring Partisan Distortion and Its Component Parts." Bachelor's Thesis, Harvard College (Presented at the Tufts Redistricting Conference May 2017). https://nrs.harvard.edu/urn-3:HUL.InstRepos:38986910.

Wasserman, L. 2012. "Minimaxity, Statistical Thinking and Differential Privacy." *Journal of Privacy and Confidentiality* 4 (1). https://doi.org/10.29012/jpc.v4i1.611.