A Reasoned Basis for Inference: Randomization and Design Justifying Estimation and Testing

Jake Bowers * Costas Panagopoulos†

July 19, 2009

Abstract

How should one estimate and test comparative effects from a field experiment of only 8 units (i.e. where consistency as a property of estimators offers little comfort)? What about from a large observational study with an outcome variable that has too many zeros to be usefully modeled as Poisson, too many large values to be Negative-Binomial, and may be some other mixture of unknown or hesitantly guessed at probability distributions? Must the lack of a correct probability model for outcomes or paucity of data in general lead to incorrect confidence intervals?

In this paper we show that statistical inference is still possible without either models of outcomes or large samples. We show how one can base testing and estimation on *models of the design of the study*, and specifically, on the process by which the values of the explanatory variable were produced. These *models of assignment* form a basis for valid hypothesis tests, confidence intervals with correct coverage, and point estimates. This framework, called by us, "randomization inference" or more generally "design-based inference", is not meant to replace inference based on models of the data generating process or inference based on both models of the data generating process and models of the parameters of that process, but it is meant to help those scholars whose research designs do not allow comfortable justification of the model-based techniques to continue their work while providing persuasive answers within a well-established framework for statistical inference.

As an example, we show how one may use design-based inference to make credible tests using a unique field experiment of the effect of newspaper advertising on aggregate turnout with only 8 observations and with an observational study of drug safety with a very skewed and strangely distributed count outcome. In addition, we present some innovations in the use of linear models (both frequentist and Bayesian) as a way to allow outcome- and parameter-models to assist the design-based inference without requiring commitments to the usual assumptions that would be required for direct causal inferences using those methods.

*NOT FOR CIRCULATION. Assistant Professor, Dept of Political Science, University of Illinois @ Urbana-Champaign *Corresponding Author Contact Information:* 702 South Wright Street, 361 Lincoln Hall, Urbana, IL 61801 — 217.333.1203 — jwbowers@illinois.edu. *Acknowledgements:* Thanks to Dan Carpenter, Ben Hansen, Mark Fredrickson, Tommy Engstrom and Joe Bowers.

†Assistant Professor, Dept of Political Science, Fordham University

1 Statistical Inference in the Land of Make Believe

Statistics as currently practiced in political science requires a lot of imagination: we imagine that nature produced our outcome variable according to some known probability distribution (i.e. we posit a data generating process); we imagine that our sample size is close to infinite; we pretend that our data arrived via a known sampling mechanism from a known population[1]; and we fantasize that we know the mathematical formula which relates control variables with both the variable(s) of causal interest, called here the "explanatory variable", and the outcome. Sometimes we add to this list of stories the name of the probability distribution(s) which describes our uncertainty about the effects of our covariate and explanatory variables prior to estimation.

Data generating process (DGP) models require at least some such stories.[2] And these stories in turn require a set of explanations about how the values of the causally important, or explanatory, variable appeared in our data (i.e. about the research design producing our data): were they randomly assigned? did units (e.g. people) consciously select which values to take on? or was the process otherwise haphazard (or unknown but not random)? or perhaps a more or less known function of other variables? And, of course, the meaningfulness of our results depends crucially on how these values (and the values of the outcome variable) map onto the concepts with which we explain how the world works (i.e. about concepts and measurement). We often call these stories, "assumptions", and a large part of statistics as applied in the social sciences focuses on helping articulate exactly how worried we ought to be when our assumptions are approximations.

Of course, assumptions help us simplify the world enough to ask and answer specific questions. And the best storytellers compel us by carefully and clearly justifying each required story. For a simple example, consider how one might justify the assumption that our outcome was produced by nature according to a Poisson process. A weak justification of this assumption might be to say, "I observe counts. Counts are often conveniently modeled as a Poisson DGP." A strong justification might be to notice that, at the micro-level, the values of the outcome emerge from a process in which events happen in time independent of one another.[3] Given this micro-foundation, one can logically deduce that a Poisson DGP for the counts of events occuring in a given unit of time follows.[4] If, however, the values of the outcome are observed to be counts, but the analyst has no detailed story about how those counts came to be, then she would be wise to feel uncomfortable telling the Poisson story (even if there is no other easily available story for her to tell). That is, DGP models (among the other justifications for a given bit of data analysis) can be made more or less plausible depending on the persuasive and logical skills of the story-writer.

[1] Perhaps one which makes our analysis units independent of one another, or perhaps which produces some known dependence among them

[2] Note that least squares and all Bayesian models imply a DGP story, although only Bayesian models must be explicit about prior distributional stories.

[3] Perhaps this process is itself the result of some strategic interaction of units, some cooperative interaction, some random interaction, or no interaction. We avoid providing a substantive theoretical story here in order to keep attention on the statistics.

[4] Equivalently, one can deduce that the amount of time between the occurrence of such events follows an exponential distribution (See, *inter alia*, Ramanathan (1993, page 64), King (1989, page 48-50), Grimmett and Stirzaker (1992, page 229).). Derivations of both are available in most texts on probability and an expanded example is available online at `http://jakebowers.org/PAPERS/poissonproof2.pdf`

Although artful and careful persuasion is possible, the creation and justification of these stories, or "commitments" (Berk, 2004), can often seem very burdensome to scholars who have compelling political and social and economic theory to engage, but who don't have much to say about the statistical stories that justify and make meaningful common data analytic practice. In this article we propose to make the justification of statistical inference less burdensome for at least some political scientists. In so doing, we hope to help scholars return and maintain their focus on scientific inference [cite Heckman], concepts, measurement, and theory rather than on half-believed rationalizations of conveniently chosen models.

We do this by introducing, explaining, and demonstrating a mode of statistical inference which is frequentist but which does not require the stories demanded by DGP based approaches.[5] Of course, it, like all methods, requires its own fictions and fantasies. Specifically, "randomization inference", as developed by Neyman (1990) and Fisher (1935) does not require a model of outcomes but it does require a model of assignment.[6] Thus, since it requires its own stories, it is not uniformly better than extant approaches. We will show, however, that the pretense required of this mode can be easily assessed — an assessment of a kind that would be quite difficult for DGP based approaches.[7] Basically, since inference in this mode is based primarily on stories about the design of the study rather than on outcomes, there is more possibility for an analyst to provide strong and credible justifications for their statistical inference in so far as the design of their study is under their control. In contrast, inference requiring models of outcomes is almost never based on parts of the study under the control of the researcher, and thus, such stories are harder to believe and require more work to be credible.

This paper proceeds first by introducing two substantive political science problems, each of which might give pause to someone whose only tool is based on stories about how the outcomes were produced. Then, using the example of an 8 city randomized study of newspaper advertisements and turnout we delve into the details of randomization inference and illustrate, in fact, how regression analysis can aid such inference without requiring the kinds of commitments that usually burden analysts. Next we consider a quasi-experimental design of drug policy evaluation allows us to illustrate the large-sample properties of, now, design-based rather than randomization-based tests. And the

[5] And as a frequentist approach does not require explicit justifications of prior distributions.

[6] This body of techniques closely related to "exact inference" when asymptotic approximations are not used and to "permutation inference" — a term which highlights one of the mechanisms for generating estimates. We use "randomization inference" here and elsewhere to emphasize connections of the technique with *design*. Although these terms are very closely related, they are not identical. For example, Neyman (1990); Imai (2008); Hansen and Bowers (2009) demonstrate randomization based inference that is neither exact nor permutation based while Keele et al. (2008) show an exact and permutation based approach in randomized experiments. We will demonstrate both exact and approximate versions of this mode of inference in this paper.

[7] Bayesian approaches also require a statement about a data generating process. Assessment of the assumptions required of such models is also possible, of course, although such assessments are less informative than those used to check approximations used in randomization inference, the kind of inference introduced in this paper. However, it is worth noting that, in addition to the kind of formal linking of data-generating-process claims with real-world processes sketched in the case of the Poisson distribution, one may use research design to rule out certain kinds of linear models and selection processes. And a mode of assessing the full model that has emerged from Bayesian scholars, "inspection of the predictive posterior" has great utility for both Bayesian and frequentist approaches which require DGP models and related parameterizations (See Gelman et al. (2004) and Box (1980) for elaboration and explanation of these kinds of useful ideas for model checking and fit for Bayesian models.)

paper concludes with some discussion of some potential and common concerns about this mode of inference, highlighting especially the cases in which inference based on stories about a design of a study may be worse than inference based on stories about a data-generating process.

1.1 Example 1: Can newspaper advertisements enhance turnout in low salience elections? The case of the 8 City Newspapers Randomized Field Experiment.

In the days just before the November 2005 elections, C. Panagopoulos fielded an experiment to assess the effects of non-partisan newspaper ads on turnout in low salience elections. This was, to our knowledge, the first experiment to investigate the impact of newspaper ads on turnout, and as a pilot study, it was small, involving only eight cities, matched into pairs on the turnout in the previous election. [8] Within each of the 4 pairs, one city was assigned at random to receive black-and-white newspaper ads in local newspapers encouraging citizens to vote. Panagopoulos (2006) provides more detail on the design of the experiment and detailed analysis of the conclusions. Table 1 shows all of the observations in the study with associated design and outcome features.

				Turnout	
City	State	Pair	Treatment	Baseline	Outcome
Saginaw	MI	1	0	17	16
Sioux City	IA	1	1	21	22
Battle Creek	MI	2	0	13	14
Midland	MI	2	1	12	7
Oxford	OH	3	0	26	23
Lowell	MA	3	1	25	27
Yakima	WA	4	0	48	58
Richland	WA	4	1	41	61

Table 1: Design and outcomes in the Newspapers Experiment. Treatment with the newspaper ads is coded as 1 and lack of treatment is coded as 0 in the 'Treatment' column.

We see here Saginaw (control) and Sioux City (treated) in one pair. Turnout in Sioux City increased from before the treatment to after the treatment by 1 point (22 vs. 21) and was higher than its control unit (Saginaw) after treatment (22 vs. 16). In addition, turnout in Saginaw (which was not exposed to the experimental newspaper ads) decreased by 1 point (16 vs. 17) from the election before the treatment to the election after the treatment. Those three different pieces of information suggest that the treatment had a positive effect on Sioux City. However, a thoughtful reader will have, by now, realized that each of those three comparisons have some flaws: something other than the treatment could have caused turnout to increase over time in Sioux City, baseline turnout was higher in Sioux City than Saginaw, and it is easy to wonder whether, in the absence of treatment, Sioux City would also have had a slight decline

[8] About 281 cities with populations over 30,000 held a mayoral election in 2005. The Newspapers study focused especially on the roughly 40 cities with indirect election of mayors (i.e. where mayors are elected by city councils, not directly by the public). And, among these cities, only those cities in which the city council had been unanimous in their election of the mayor in the previous election were considered potential experimental units. After collection of covariates (such as vote turnout in the previous municipal election and partisanship of the election) roughly 9 cities had complete data to allow matching into pairs, and roughly 1 city was discarded as not easily matchable with any of the others on the basis of turnout in the previous election. [These are rough numbers. Check.]

in turnout in the same way that occured in Saginaw. In some senses we might think that the simple control versus treated comparison of 22-16=6 percentage points of turnout would be the right estimate of a treatment effect for Sioux City here. Yet, since Sioux City only increased by 1 percentage point from baseline, we might wonder if somehow 6 pct pts overstates the effect. That is, in this dataset, one challenge will be to use information other than mere assignment to treatment and control to produce compelling estimates of treatment effects. This study also is very small: there are only 8 cities grouped into 4 pairs. The small sample size raises doubts about estimation strategies grounded in arguments about consistency (of either points or intervals). The availability of baseline outcomes (as well as other covariates) suggests that, since we know more about these units, we should use this information to enhance the precision of our estimates if not also the persuasiveness of our comparison. This example will allow a very clear and easy exposition of randomization inference, and the complications regarding use of baseline and covariate information offer an opportunity to show how randomization inference allows for more than just simple comparisons of treated and control units.

1.2 Example 2: Did the regulatees capture the regulator? The case of a policy change at the FDA.

Reacting to public concern at extended waiting times for potentially life-saving drugs to be approved by the FDA, on October 29, 1992 Congress implemented the Prescription Drug User Fee Act (PDUFA) which required the FDA to act on 90% of standard drugs within 12 months. In order to fund the increased work, PDUFA required drug companies to pay a fee for each drug they submit — currently around $200,000. The new funds and policy regime enabled the FDA to nearly double the number of reviewers and other workers it hired, and it successfully met and even surpassed its deadline speed goals [cite to Carpenter]. Never before had the FDA faced deadlines. Never before (as far as we know) had those regulated had such direct control over the budget of a regulator.[9]

The tenor of criticism from those fearing that capture and speed would endanger the public is well represented by the following quote from a report by Public Citizen:

> Since the passage of PDUFA in 1992 there have been an unprecedented number of drugs approved and then withdrawn for safety reasons. Nine of the drugs ... approved ... from 1993 to 2000 have been withdrawn because of safety concerns. By comparison, only five of the drugs approved ... from 1985 through 1992 later had to be withdrawn. ... At least five of these nine drugs ... were approved despite known safety problems and the availability of multiple treatment options in other, older (and safer) drugs approved for the same medical uses. (Lurie and Sasich, N.d., from Public Citizen)

Of more direct medical concern were reports that Adverse Events Reports (AER)s had increased on average per drug from around 430 from 1962 to 1992 to 720 after 1992 until 2000. When medical professionals diagnose a patient as having a negative reaction to some drug, the medical professionals seeing the patient may report this problem to the FDA or to the drug manufacturer (which is then required to forward this information to the FDA). The system is voluntary and is meant to alert the FDA

[9] We think that the pharmaceutical companies actually have some more fine-grained control over how their money is spent, but we don't have details.

for the need to engage in further study rather than to be the final word in drug safety problems. [10] That said, these reports are one of the best extant measures of drug safety — especially from the perspective of political scientists since they are an important part of the post-approval information stream arriving at the FDA.[11]

The questions of interest to social scientists in this case concern capture of an agency [cites], and the effects of deadlines on decision making, especially high stakes technical decision making by government actors [cite Carpenter]. Here we take on the simplest implication from such concerns for the sake of methodological exposition: that the policy change decreased the ability of the FDA to adequately control the safety of the drugs they approved.

This study includes many drugs — in fact it includes the entire population of drugs approved by the FDA since 1962.[12] The outcome of interest, Adverse Events, is a very skewed, overdispersed, count that ranges from 0 to 18,000 (50% of the drugs had between 1 and 360 events reported).

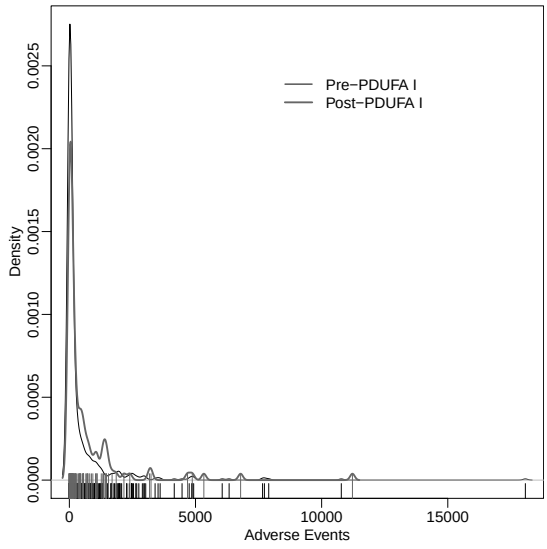


Figure 1: The distribution of adverse events reports for drugs approved by the FDA between 1962 and 2000. Black lines show drugs submitted Jan 1, 1962–Oct 29, 1992. Gray lines show drugs submitted Oct 29, 1992–Jan 1, 2000. PDUFA I was implemented Oct 29, 1992. The longer lines in the rug-plot represent PDUFA I drugs.

Any method of estimation of policy effects here requires a comparison of drugs submitted and approved under the PDUFA I regime to drugs submitted and approved before the change. We will engage with this question briefly when we analyze these data. Of more importance for this paper is the question of statistical inference. What is the data generating process producing adverse events? What theory might we have to justify a

[10] See `http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm` for more on this system.

[11] The data from the FDA study have been collected and checked by Dan Carpenter and his colleagues. For other analyses of these data and more in-depth discussion see Bowers et al. (2007); Carpenter et al. (2007); Carpenter (2004); Carpenter et al. (2009); Nardinelli et al. (2008)

[12] Some submissions and approvals were recorded as far back as 1943, but the regulatory regime in the FDA changed in 1962 to emphasize safety and to become a gatekeeper for the market for drugs and so we limit our population to the post-1962 drugs.

zero-inflated negative-binomial? A mixture of integer truncated Gammas? Or something else entirely?[13] Most analysts at this point would commit to some model of outcomes, pretending as if they had a persuasive stochastic process model grounded in social science theory to justify the model (both the probability model and the parameterization). For some of these analysts such a story might be very important to tell given their intellectual agenda. For others, with other intellectual agendas, perhaps deriving and justifying a model of outcomes would be an unnecessary burden.

In this paper, we present an alternative way of story-telling: we can focus our tall-tale telling on a model of the design rather than on the outcomes, using linear models to enhance precision without committing to auxiliary assumptions of such model, and maintain some integrity regarding estimation of effects. In this application to observational data, the model of design is not inherently more plausible than the model of outcomes (in contrast to the randomized study where the model of design is based in the physical act of randomizing and the, one assumes, faithful reports of the human doing the randomizing). Yet, since "design" does not stop once data are collected, a mode of inference based on design may well offer the analyst more control over the raw materials of the story told than a mode of inference based on outcomes. "Design" in this context means any data analysis or data collection or planning thereof done before estimation of treatment effects. In this understanding we agree with Rubin when he says, "by 'design' we mean all contemplating, collecting, organizing, and analyzing of data that takes place prior to seeing any outcome data." (Rubin, 2008, page 810).

1.3 Plan of the Paper

The paper shows how to estimate effects from these datasets without large-sample assumptions or data generating process models. When we do introduce the convenience of large-sample approximations or the precision enhancement arising from a model of outcomes, the statistical inference will still be based entirely on the design of the studies.[14] The linear models used here will not require commitments to the standard assumptions of these models: they will be used to reduce noise in the outcomes, not to directly estimate treatment effects. The confidence intervals we produce will include hypotheses that reject a true null no more than the pre-specified $1 - \alpha$ of the time: that is a 95% CI will be guaranteed to contain the true estimate at least 95% of the time if the experimental manipulation/policy intervention was re-assigned. In contrast, extant methods relying on large samples will produce confidence intervals labeled as, say, 95% CIs but which will in fact contain the true estimate less often than specified (and less often is as precise a statement as we can make about the failure of these large sample CIs to have correct coverage without simulation). [simulations comparing coverage to add empirical confirmation to this theoretical claim are planned but not completed]

First, we will demonstrate randomization inference on the Newspapers study. Since that study is small, it allows us to lay bare the details of the method to (1) aid comprehension by newcomers and (2) contrast with other modes of inference which would require many more pages and much more background to fully describe. Second, the Drugs study will allow us to use large-sample approximations and to engage with an observational study

[13] Perhaps a Glockenspiel estimator? (Achen, 2002, page 440)

[14] In this paper we use "model of outcomes" and "data generating process model" interchangeably. Both imply a probability distribution governing the values of the outcome and some parametrization of that distribution.

— an engagement which will require much more work to generate and justify a model of assignment than the randomized Newspapers study. Finally, we will discuss and re-emphasize the limitations of this method and the trade-offs required of those who choose to use it for a particular problem over the tools currently common in political science. We do not desire to replace those tools. For those who already masterfully justify the stories required of data generating process models, randomization inference may be merely convenient in certain circumstances. For those who find the stories required by our common toolkit a source of burden and worry, we offer an attractive alternative.

2 The Newspapers Study

What is the effect of newspaper advertisements on aggregate vote turnout in the Newspapers dataset (shown in Table 1)? By "effect" here we refer to a counterfactual comparison. The advertisments can be said to have an effect if the turnout of cities i treated with the advertisements ($Z = 1$), $r_{Z=1,i}$, would have been different in the absence of advertisements ($Z = 0$). We can write the potential outcome to control as $r_{Z=0,i}$ or more simply r_{0i} to denote the response of city i without advertisements, and $r_{Z=1,i} \equiv r_{1i}$ for the response of city treated with advertisements.[15] By "causal effect", τ, we refer to a comparison of potential outcomes such as $\tau_i = r_{1i} - r_{0i}$. Notice that this framework is a conceptual heuristic: we cannot actually ever observe both r_{1i} and r_{0i}.[16] We could represent these potential outcomes in the Newspapers design as follows in Table 2.

i	b_i	Z_i	R_i	r_{1i}	r_{0i}
Saginaw	1	0	16	?	16
Sioux City	1	1	22	22	?
Battle Creek	2	0	14	?	14
Midland	2	1	7	7	?
Oxford	3	0	23	?	23
Lowell	3	1	27	27	?
Yakima	4	0	58	?	58
Richland	4	1	61	61	?

Table 2: Treatment (Z), Observed outcomes (R), and potential outcomes (r_1, r_0) for Cities (i) within Blocked Pairs (b_i) in the Newspapers Experiment.

For example, we observe that turnout was 16% in Saginaw. We take this to mean turnout in the absence of treatment (r_{0i}) is 16% in Saginaw. We don't know, without

[15] We can write r_{0i} because we also assume that the potential turnout in city i is unaffected by treatment to other cities. If the treatment giving to one city, j, influenced outcomes in another city i, , we would have to define the potential response of city i to control in terms of both the treatment assigned to it and also to city j: perhaps $r_{Z=\{0,0\},i}$ where $Z = \{0, 0\}$ would mean that both units received control rather than treatment. This assumption is reasonable in this dataset, but by no means is a trivial assumption to maintain in many political science studies (Brady, 2008).

[16] The idea that one must compare possible outcomes, or "potential outcomes" to make causal effects meaningful was introduced in the 1920s by Neyman (1990) and most prominently elaborated and developed by Rubin (1974, 2005). For more on the intellectual history of this idea and spirited arguments in its favor see Holland (1986); Sekhon (2008). For commentary and criticism of the potential outcomes framework (also often known as the Neyman-Rubin conceptualization of causal effects) (Brady, 2008). And also see Rosenbaum (1999) for practical strategies using this framework in the context of observational studies.

further information and/or assumptions, how the turnout in Saginaw would have been had Saginaw instead of Sioux City been exposed to newspapers advertisements in the 2–3 days before the election. Clearly, the act of making causal inferences requires replacing the "?"s in Table 2 with meaningful numbers. How can we get them?

Let us recall our definition of a causal effect as a comparison of potential outcomes: $\tau_i = r_{1i} - r_{0i}$. If there were no effect for, say, Sioux City, $\tau_{i=\text{Sioux City}} = 0$ implying that $r_{1,i=\text{Sioux City}} = r_{0,i=\text{Sioux City}}$. That is, if there were no effect, turnout in Sioux City without advertisements would be the same as turnout with advertisements. We know that turnout in Sioux City in the presence of advertisements was 22%. Thus, if advertisements had no effect on turnout, turnout in Sioux City in the control condition would have had to be 22%. Notice that positing, or hypothesizing, that treatment had no effect and then representing "no effect" in terms of our definition of a causal effect allows us to fill in the missing data. This way of thinking about what "no effect" means is very clear: "no effect" means that we would observe the same outcomes for each unit regardless of experimental condition assigned. This kind of hypothesis implies something specific about each and every unit in the data. Often called a "sharp null hypothesis", this idea and randomization-based causal inferences based on them (but not on potential outcomes) was first proposed and developed by in Fisher (1935).[17]

So, a strict null hypothesis of "no effect" implies that, for all i, $r_{1i} = r_{0i}$. If the hypothesis were true, the missing potential outcomes in Table 2 would be known. Most importantly, we would know the potential outcomes under control for the treated observations. Combined with the observed outcomes for the control group, a given sharp null hypothesis specifies how the outcomes would look in the absence of treatment. Given a vector of potential outcomes in the absence of treatment generated by a sharp null hypothesis, we can now generate a distribution representing the possible treatment effects that could arise from the experiment even if there were no effect. To do so, we will need (1) a test statistic summarizing the relationship between treatment assignment and observed outcomes, and (2) a model of the design of the study (i.e. a model of how treatment was assigned), and eventually, in order to calculate a confidence interval (3) a model of the effects of the study.

2.1 Test statistics summarize treated-versus-control differences

Random assignment requires us to consider comparing cities i within pair b across treatment assignment as the basis for our test statistic. We know that cities within pairs received newspaper ads without regard to any other characteristic of those cities. Thus, comparing 2005 turnout in cities receiving the newspapers (the treated cities) to 2005 turnout in cities which did not have those particular newspapers ads (the control cities) is, in expectation, a comparison that reflects only the manipulation and not any other characteristic of the cities.

Write R_{bi} for the observed percent of registered voters who voted in 2005 in each city

[17] The example of the Lady Tasting Tea given in Fisher (1935, Chapter 2) is perhaps the most famous application of the sharp null hypothesis and randomization inference. The Newspapers example used here is more interesting in that it is not a toy example, has to do with voting turnout rather than tea-tasting, and has other features (like stratified treatment assignment and a continuous outcome) that make it a more useful example for political scientists. For more on the Lady Tasting Tea and its usefulness as an example in comparative politics see Sekhon (2005). See also Agresti and Min (2005).

$i \in \{1, 2\}$ in pair $b \in \{1, 2, 3, 4\}$. The observed outcomes are a function of potential outcomes: $R_{bi} = Z_{bi} r_{1bi} + (1 - Z_{bi}) r_{0bi}$. We use capital letters to refer to random quantities and lowercase letters to refer to fixed quantities. In this experiment we know that treatment was assigned at random, thus we know that Z is random. Potential outcomes are assumed to be fixed characteristics of units which can be revealed by the experiment. We can write a simple difference of mean turnout within pairs as:

$$\begin{aligned} t(Z, R)_b =& \frac{\sum_i Z_{bi}^T R_{bi}}{\sum_i Z_{bi}} - \frac{\sum_i (1 - Z_{bi}^T) R_{bi}}{\sum_i (1 - Z_{bi})} \\ =& \text{Mean Turnout Among Treated Cities in Pair } b - \\ & \text{Mean Turnout Among Control Cities in Pair } b. \end{aligned} \tag{1}$$

And we can calculate the overall difference as the average of the within pair averages from equation 1.

$$t(\mathbf{Z}, \mathbf{R}) = \frac{\sum_{b=1}^{B} t(Z, R)_b}{B}. \tag{2}$$

The within-block formula written in equation 1 is generalizable to situations where blocks contain more than one treated unit and/or more than one control unit. The formula for calculating the overall effect as a weighted average of within-block averages is not generalizable to different sized sets because it assumes that each block ought to contribute equally to the overall average. When we have one treated and one control unit per block, one could argue that each block contains the same amount of information about the overall treatment effect as any other. When the block sizes vary, then blocks with very lopsided treated-to-control ratios provide less information than blocks with more equal ratios and weights ought to reflect this difference.[18]

A cleaner way to combine those formulas would be to use directly use vector notation where $\mathbf{Z}$ collects the Z_{bi} and $\mathbf{R}$ contains the R_{bi} , and the notation $|\mathbf{Z}|$ means the "size" of $\mathbf{Z}$ or $|\mathbf{Z}| = \mathbf{Z}^T \mathbf{1}$:

$$t(\mathbf{Z}, \mathbf{R}) = \frac{\mathbf{Z}^T \mathbf{R}}{|\mathbf{Z}|} - \frac{(\mathbf{1} - \mathbf{Z})^T \mathbf{R}}{|(\mathbf{1} - \mathbf{Z})|} \tag{3}$$

Applying equation 1 or 3 to the Newspapers dataset gives an average difference of treated and control units within pairs in turnout after treatment of 1.5 percentage points. Is 1.5 percentage points a "real" result? Say, for the sake of argument, that there were no effects of newspaper advertisements: how much more (or less) plausible does this observed result make the that hypothesis? If we want to use probability to answer this question, we would need to compare the fixed quantity that we observe ($t(\mathbf{Z}, \mathbf{R}) = 1.5$) to the distribution of such statistics defined by the null hypothesis $H_0 : \tau_i = 0$.

[18] See Hansen and Bowers (2008) for generalizations of these formulas to unequal sized blocks and clustered random assignment and for formal arguments about how to weight the contributions from different blocks. See Imai (2008) for an argument in favor of weighting unequal sized blocks equally.

2.2 Simple Inference for a Null of No Effect

The probability distribution we require represents the probability of observing all of the possible outcomes of the experiment in the case that the treatment had no effect. That is, it would tell us the probability of observing $t(Z, R) = 0$ percentage points, and the probability of observing $t(Z, R) = 1.5$ percentage points (and any other reasonable effect) under the hypothesis of no effects. So, first we need to know the range of possible effects and then we need to assign some probability to each of these effects. We will see here that the design of this experiment allows us to accomplish both tasks with no other assumptions. We use this method (often called a "permutation" method or an "enumeration" method) in this section of the paper to show the simplicity of randomization based tests.

We begin with the task of delineating the domain of this distribution — the values of the treatment effect for which we require probability statements. Here we have four blocks, each with two units, and each with one and only one treated unit. So, within block we have $\binom{2}{1} = 2$ ways to assign treatment. And across the four pairs, we have $\prod_{s=1}^{4} \binom{2}{1}_s = 2^4 = 16$ ways to assign treatment.[19]

Table 3 shows the 16 different ways to assign treatment to cities, respecting the blocking structure of the design. This matrix is often labeled Ω and it contains all of the possible assignment vectors $\mathbf{z}$ of which $\mathbf{Z}$ is the one we observed. For example, the first possible assignment vector $\mathbf{z}$ from the set of all possible such vectors Ω would assign treatment to Saginaw but control to Sioux City, while the 2nd $\mathbf{z}$ would assign control to Saginaw and treatment to Sioux City.

city	pair	$\mathbf{Z}$	$\mathbf{z} \in \Omega$															
Saginaw	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
Sioux City	1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
Battle Creek	2	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
Midland	2	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Oxford	3	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
Lowell	3	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Yakima	4	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
Richland	4	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1

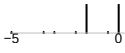
Table 3: Possible treatment assignments for the Newspapers study. Pairs of cities define the columns. Permutations of the assignment mechanism (within pair fixed random assignment) define the rows.

This table is the model of assignment of the study.[20] For each $\mathbf{z} \in \Omega$ in Table 3 we can calculate a test statistic $t(\mathbf{Z}, \mathbf{R}|\mathbf{Z} = \mathbf{z})$ representing the differences between treated and control units for each of the possible ways that assignment to treatment could have happened. If the sharp null hypothesis of no effect were true then $\mathbf{r}_0 = \mathbf{r}_1$ and then $\mathbf{R} = \mathbf{Z}\mathbf{r}_1 + (1 - \mathbf{Z})\mathbf{r}_0 = \mathbf{Z}\mathbf{r}_0 + (1 - \mathbf{Z})\mathbf{r}_0 = \mathbf{Z}\mathbf{r}_0 + \mathbf{r}_0 - \mathbf{Z}\mathbf{r}_0 = \mathbf{r}_0$. That is, if our null were true, what we observe in $\mathbf{R}$ is what we would observe under the control condition. Thus, the collection of $t(\mathbf{Z}, \mathbf{R}|\mathbf{Z} = \mathbf{z} \in \Omega)$ summarizes all of possible test statistics if the

[19] Although we have eight units, four of which were treated, the number of possible assignments is not $\binom{8}{4} = 70$ because of the blocking into pairs. In this way the Newspapers dataset is even a bit simpler than the classic Lady Tasting Tea example.

[20] One could write this model more parsimoniously as is often done, or merely list all of the possible ways for treatment to be assigned.

null were true. Since we know the probability of seeing each $\mathbf{z} \in \Omega$, we now have the distribution of the test statistic under the null hypothesis. In this case, the probability of seeing each $\mathbf{z}$ happens to be uniform with probability 1/16.

Now we are ready to answer our question about how strange it would be observe a mean difference as far away from 0 as 1.5 if, in fact, there were no effect. The probability of observing a result as large or larger than 1.5 is the sum of the probabilities of at 1.5 and larger. The probability of observing a result as small or smaller than -1.5 can be calculated in the same manner. And the two-sided p-value is the sum of those calculations. The full probability distribution, 

, is very simple: each vertical line represents the amount of probability for each possible outcome — most of the mass is concentrated at 0, -1.5 and 1.5, with small amounts of mass scattered across the line from -5 to 5. More detail is shown in Table 4.

$t(\mathbf{Z}, \mathbf{R})$	-5	-3.5	-3	-2	-1.5	-0.5	0	0.5	1.5	2	3	3.5	5
$p(t(\mathbf{Z}, \mathbf{R}))$	0.06	0.06	0.06	0.06	0.12	0.06	0.12	0.06	0.12	0.06	0.06	0.06	0.06

Table 4: Randomization based probability distribution of paired mean treatment–control differences) under the sharp null hypothesis of no effect.

A common two-sided p-value is defined as the sum of the absolute value of the mass at or greater than the observed value.[21] In this case the there are 12 ways out of 16 to observe a value as far from zero as 1.5 (or -1.5). Thus, the weight of evidence that our observed value of 1.5 places against $H_0 : \tau = 0$ is summarized by $p = 12/16 = 0.75$.

Other test statistics Notice that nothing about this procedure requires a comparison of simple means. We could produce the same result with a standardized version (i.e. the test statistic of the simple t-test). Or we could use a sum of ranks among the treated: For example, if $\mathbf{q} = \text{rank}(\mathbf{R})$, then we could define the rank sum test statistic: $t(\mathbf{Z}, \mathbf{q}) = \mathbf{q}^T\mathbf{Z}$. Notice that since the ranks are a function of observed responses, they are also functions of potential outcomes. Different test statistics might have different statistical power or might otherwise summarize results in more or less substantively meaningful ways.[22] Lehmann (1998); Keele et al. (2008) among others suggests rank based tests for their insensitivity to outliers and enhanced power compared to mean-based tests. Table 5 shows the distribution of the paired-rank sum test statistic under the sharp null hypothesis of no effect. The observed rank sum is 6, and the two-sided p-value is $2 \times 0.4375 = 0.875$. Since rank sums are always positive, the two-sided p-value must be defined as twice the one-sided p-value rather than using the absolute values of the domain of the randomization distribution.

$t(\mathbf{Z}, \mathbf{R})$	0	1	2	3	4	5	6	7	8	9	10
$p(t(\mathbf{Z}, \mathbf{q}))$	0.06	0.06	0.06	0.12	0.12	0.12	0.12	0.12	0.06	0.06	0.06

Table 5: Randomization based probability distribution of the paired rank sum test statistic under the sharp null hypothesis of no effect.

[21] Another common definition is twice the one-sided p-value (or even twice the smaller of the two one-sided p-values) [cites].

[22] For example, Hansen and Bowers (2009) used "number of votes" rather than "probability of voting".

2.3 Summary of the Test of the Sharp Null of No Effects

Specify a hypothesis about τ $H_0 : \tau = \tau_0$, where τ is some function of potential outcomes to control r_{0i} or treatment r_{1i}.

Specify a test statistic $t(\mathbf{Z}, \mathbf{R})$. For example, if $\mathbf{q} = \text{rank}(\mathbf{R})$, then $\mathbf{q}^T\mathbf{Z}$ (rank sum). Or $\frac{\mathbf{Z}^T\mathbf{R}}{\mathbf{Z}^T\mathbf{1}} - \frac{(1-\mathbf{Z})^T\mathbf{R}}{(1-\mathbf{Z})^T\mathbf{1}}$ (diff of means).[23]

Compute the test statistic as implied by H_0 $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z})$. Under the sharp null of no effect $\mathbf{R} = \mathbf{r}_0$.

Compare $t(\mathbf{Z}, \mathbf{R})$ *to* $t(\mathbf{z}, \mathbf{R})$ *for all possible* $\mathbf{z} \in \Omega$ Equation 4 summarizes the doubt cast by our observed test statistic against the null hypothesis:

$$\Pr(t(\mathbf{z}, \mathbf{R}) \geq t(\mathbf{Z}, \mathbf{R}) | \tau = \tau_0) = \frac{\sum_{\mathbf{z} \in \Omega} \mathbf{1}\{t(\mathbf{z}, \mathbf{R}) \geq t(\mathbf{Z}, \mathbf{R})\}}{K} \tag{4}$$

where Ω is the matrix of all possible treatment assignments, and K is the total number of possible assignments in Ω, in the case of independent assignment across strata, $K = \prod_b \binom{n_b}{\sum_i^{n_b} Z_i}$. In our case $K = \prod_{s=1}^{4} 2 = (2)^4 = 16$.

Notice that this test did not require any model about how turnout occurs in cities, nor did it require assumptions about large sample sizes, or any particular functional form relating assignments and outcome (recall that the test of the sharp null is represented by the same pattern of $\mathbf{r}_0$ regardless of model of effects). The mechanism by which the p-values were produced is completely transparent (especially so in the case with $K = 16$!).

Yet, we are left with some discomfort: First, political science actually knows something about turnout. That is, although precise models of outcomes may be a burden (and more of a burden with outcomes like the Adverse Events data), we would like to use what we know. We will show later how models of outcomes may be used within this framework without requiring inference itself to be based on models of the data generating process. Second, knowing that the null of no effect is not implausible is not the same as producing a range of values that are plausible. We will next demonstrate how such tests as we have executed here may be "inverted" (cite to page in Lehman among many other books) to estimate a range of plausible values for the effect of newspaper advertisements on vote turnout.

2.4 Confidence Interval: Assessing Hypotheses about effects

So far we have assessed a hypothesis about no effects. If we want to talk about plausible effects, we must assess hypotheses about some non-zero effects. Recall that a confidence interval is *defined* as the range of hypotheses that would be accepted at some α level denoting the risk of falsely rejecting a true hypothesis. That is, given a choice of acceptable Type-I error rate, one can create a confidence interval out of hypothesis tests. This method, called "inverting a hypothesis test" is a well known procedure and is not specific to randomization inference [cite a couple of textbooks like Rice (1995) or

[23] Here I have used paired versions of these test statistics because of the random assignment within pairs. I am using matrix notation $\mathbf{R}$ rather than R_{bi} for clarity on the page.

others perhaps Hodges and Lehmann (1964)?]. For example, we have so far assessed $H_0 : \tau = \tau_0, \tau_0 = 0$ and have p-values of 0.75 and 0.875 for the mean and rank-based test statistics respectively. Regardless of test statistic, $\tau_0 = 0$ must be within any reasonable confidence interval (say, of $\alpha = .025$ for a 95% CI let alone $\alpha = .12$ for a 88% CI). That is, $\tau_0 = 0$ is a plausible value for the effect of advertisements on turnout. What about other values?

In order to assess hypothesis about other values we must now add some structure to the problem. Namely we must posit a model of effects. First, we'll explain what we mean by model of effects, and second demonstrate how such a model allows us to assess hypothesis about effects and thereby to produce a confidence interval.

2.4.1 Models of Effects

For the rest of this paper we will be working with a very simple model of effects, namely the constant, additive effect model:

$$\tau = r_{1i} - r_{0i} \tag{5}$$

which implies that $r_{1i} = r_{0i} + \tau$: meaning, the potential outcomes under treatment are merely the potential outcomes under control plus some constant τ which is the same for all cities. I make this assumption both for convenience and also to make these analyses parallel those that others might do using models of outcomes (e.g. linear model based analyses which wrap models of effects into models of outcomes). There is nothing about this mode of inference, however, that requires such a model of effects.

Recall our notation:

Treatment $Z_{ib} \in \{0, 1\}$ for unit i in strata/block/pair b. $\mathbf{Z}$ collects all of Z_{ib}.

Observed Outcomes $R_{ib} = Z_{ib} r_{1ib} + (1 - Z_{ib}) r_{0ib}$ are a function of random assignment and fixed potential outcomes.

Covariates $\mathbf{x}$ is a matrix of fixed covariates (variables uninfluenced by treatment). Strata indicators ($b = 1 \ldots B$) are covariates.

Recall also that we define a treatment effect as a function of potential outcomes. Here are merely a few of the possible models of treatment effects that one may use.[24]

Constant, additive effects $r_{1ib} = r_{0ib} + \tau$ (like t-tests — very common, implied by most linear model based analyses)

Varying, additive effects $r_{1ib} = r_{0ib} + \tau_{ib}$ (especially useful for binary outcomes (Hansen and Bowers, 2009; Rosenbaum, 2002a, 2001, See, e.g.))

Effect proportional to dose $r_{1ib} = r_{0ib} + \beta(d_{1ib} - d_{0ib})$ (Z changes D — instrumental variables based approaches imply this model. Hansen and Bowers (2009); Bowers and Hansen (2006) combine this model with the model of varying, additive effects.)

Dilated effects $r_{1ib} = r_{0ib} + \delta(r_{0ib})$ (Effect of the treatment larger among those with larger r_{0ib})

[24] Rosenbaum (2002c, Chapter 5) elaborates these models of effects among others. [ToDo: Find articles which apply each one of these models.]

Displacement effects $r_{1ib} > \theta > r_{0ib}$ where θ is some value of the order statistics of r_{Cib}. (the effect of the treatment strongest/evident at the 80th percentile).

Recall that to test the sharp null of no effects, no particular model of effects is required. It turns out that the sharp null of no effects implies the same pattern of r_{0i} for all such models of effects. However, to test hypotheses about effects (which is required to produce confidence intervals), a model of effects is necessary and different models will have different confidence intervals. In fact rejection of a null hypothesis could either be said to tell us something about the particular value of τ posited or that our model of effects is wrong (Rubin, 1991, § 4.1)

Consider $H_0 : \tau = \tau_0$ to generalize from the simple hypothesis of no effect. If $\tau = \tau_0$ and assuming a model of constant, additive effects from equation 5 where $r_{Tsi} = r_{Csi} + \tau_0$, then:

$$\begin{aligned} R_{bi} &= Z_{bi} r_{1bi} + (1 - Z_{bi}) r_{0bi} \\ &= Z_{bi} r_{0bi} + Z_{bi}\tau + r_{0bi} - Z_{bi} r_{0bi} \\ &= r_{0bi} + Z_{bi}\tau_0 \end{aligned} \tag{6}$$

or

$$r_{0bi} = R_{bi} - Z_{bi}\tau_0 \tag{7}$$

So, our null hypothesis again tells us what our potential outcomes would be as a function of the hypothesis, assignment, and observed outcomes. But this time rather than showing that $H_0 : \tau = \tau_0 \;\Rightarrow\; R_{ib} = r_{0ib}$ our model of effects means that $H_0 : \tau = \tau_0 \;\Rightarrow\; R_{ib} = r_{0ib} + Z_{bi}\tau_0$. Notice that the test of a constant additive effect of 0 and as a test of the varying additive effect of $\tau_{0i} = 0$ imply the same pattern of potential outcomes and observed outcomes if $\tau_0 = \tau_{0i} = 0$

Now consider a model of effects in which each unit may have a different effect but the effect is still additive: $r_{Tsi} = r_{Csi} + \tau_i$. Now, $H_0 : \tau_i = \tau_{0i}$ and since a hypothesis must specify a pattern of responses across all units, we state H_0 as a comparison of $N \times 1$ vectors: $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0$ where $\boldsymbol{\tau}_0$ contains a pattern of τ_{0i} which may differ across the units of the study. For a null of no effect $H_0 : \boldsymbol{\tau} = \boldsymbol{\tau}_0 = \mathbf{0}$. However, any $\boldsymbol{\tau}_0$ is testable by forming the hypothesized $\mathbf{r}_C = \mathbf{R} - \mathbf{Z}\boldsymbol{\tau}_0$ and using $t(\mathbf{Z}, \mathbf{R} - \mathbf{Z}\boldsymbol{\tau}_0)$ as the test statistic to summarize the comparisons of potential outcomes under the null. In this case of varying effects, the posited model of effects also restricts the range of possible $\boldsymbol{\tau}_0$. For example, if we assume a simple model of effects $r_{Tsi} \geq r_{Csi}$, we would not consider hypotheses which would contradict this statement. For these rest of this paper, we use a model of constant, additive effects so as to make what we are doing here as similar to the familiar data generating process world as possible.

2.4.2 A Confidence Interval by Inverting A Set of Hypotheses

The logic of § 2.2 can apply directly here. Now the test statistic is $t(\mathbf{Z}, \mathbf{R} - \mathbf{Z}\boldsymbol{\tau}_0)$ rather than $t(\mathbf{Z}, \mathbf{R})$, but Table 2 still represents the ways that the experiment could have occurred. Thus, for a given hypothesized value of τ, τ_0 where the "0" stands for null

hypothesis, we can calculate $t(\mathbf{z}, \mathbf{R} - \mathbf{z}\boldsymbol{\tau}_0)$ for all of the $\mathbf{z} \in \Omega$ and refer to equation 4 for a p-value to summarize the implausibility of this τ_0. If the p-value is greater than or equal to our α value, τ_0 is inside the CI, otherwise it is excluded from the CI as implausible.

Say we want to test $H_0 : \tau = 1$. Our model of effects and the logic of equation 7 says that, if our null hypothesis were true, potential outcomes to control among the treated would be potential outcomes to treatment minus $\tau_0 = 1$: $r_{0bi} = R_{bi} - Z_{bi}\tau_0$. This operation of removing the hypothesized effect from the treated group is the second addition to the simple test of the sharp null of no effects. However, once we have specified a model of effects and adjusted outcomes according to a given hypothesis, we can evaluate the evidence against this hypothesis using the same procedure as above: for each of the possible $\mathbf{z} \in \Omega$ calculate the test statistic, now using the adjusted outcomes, $t(\mathbf{z}, \mathbf{R} - \mathbf{z}\boldsymbol{\tau}_0)$, and refer to equation 4 for a p-value.

Using the paired Wilcox rank sum statistic and applying it to all 16 of the $\mathbf{z} \in \Omega$ we discovered that $\tau_0 \in \{-7, 6\}$ formed the boundaries of a confidence set within which the two-sided p-values were all greater than or equal to .25 and outside of which the p-values were smaller than or equal to .125 — an 88 % CI (recall that a $100(1-\alpha)$ CI contains hypotheses not-rejected at level α and excludes hypotheses rejected at level α or less). A 66% CI (which approximates $\pm$ 1 standard error for Normal variates) is [-2,5] percentage points of turnout change. We can not calculate a 95% CI from these data because the atom of the probability distribution is of size 1/16: we could, in principle have a $100(1-(1/16)) \approx 94\%$ CI but in practice such a CI would be incredibly wide.[25]

We could produce a point estimate here by shrinking this confidence interval. Yet, for this application, the confidence intervals themselves summarize the evidence adequately. The plausible effects of newspaper advertisements on turnout in these 8 cities ranges from -7 percentage points of turnout to 6 percentage points of turnout.[26]

2.4.3 Fundamental Assumptions, Discreteness, Flexible Hypothesis Testing, and Simple DGP-Based CIs

This section includes some discussions which might well end up as footnotes, but which we thought might deserve more space in this working paper.

Fundamental Assumptions: Comparability and Non-interference In producing these confidence intervals we added a model of effects to our story about random assignment. It turns out the random assignment story implies a particular, and more general assumption that is required for causal inference and there is another general assumption that we also justify by appealing to the design, without which we would not be able to conceive of let alone estimate effects as we have done.

What do we know with confidence about this design? We know that treatment was

[25] Hypotheses evaluated here ran from -20 to 20 percentage points of turnout. Even at -20 and 20, the two-sided p-values for the paired rank sum test were never less than .125.

[26] Evaluating $t(\mathbf{Z}, rank(\mathbf{R} - \tau_0\mathbf{Z}))$ for each of the 16 $\mathbf{z} \in \Omega$ is, in fact, not necessary. One can produce a $1-\alpha$ confidence interval using the paired rank sum test with one command in `R`: if $R_b = R_{ib} - R_{jb}, i \neq j$ (i.e the responses within pair are summarized by their differences), then `wilcox.exact(R,alternative="two.sided",conf.int=TRUE,conf.level=.88)` produces an 88% confidence interval that is the same as the one arrived at above by direct enumeration.

randomly assigned within pair. We know that treatment was assigned in such a way as to prevent advertisements given to one city to have any effect on the turnout of any other city (within or across pairs). We know that each city had some probability of receiving the treatment before it was actually assigned, and that assignment was governed only by a random number generator, and thus not related to any other feature, observed, or unobserved, of the cities. These sentences paraphrase two of the technical requirements for filling in the missing data in Table 2 on page 7 by *any method*. These requirements are often known as "ignorability" and "non-interference"; "non-interference" is itself related to a broader set of assumptions collectively known as the "stable unit value assumption" or SUTVA. [cites on ignorability and SUTVA from Brady, Sekhon, Rubin and Rosenbaum].

The ignorability assumption is important for everything we have done so far. In this particular case, the ignorability assumption amounts to us believing that, within matched pair, each city had an equal probability of receiving the treatment. Thus, any other differences between cities will merely add the same constant to each spike in the randomization distribution and will not change our p-values. If we are wrong about this assumption, then our randomization distribution will not reflect the posited null hypothesis, and it will be hard to know what our p-values mean.

The non-interference part of the SUTVA assumption is important for the confidence intervals but not for testing the sharp null of no effects (Rosenbaum, 2007). This assumption does allow us to write r_{1i} (potential outcome to treatment for unit i) versus $r_{11111111,i}$ or $r_{11111110,i}$ (potential outcome to treatment for unit i given some combination of treatment assignment in the rest of the study). That is, we can define models of effects by writing simple functions of $r_{1i}, r_{0,i}, Z$ because we have assumed that the potential outcomes of each unit are independent of each other.

Discreteness of Enumerated CIs Above we defined the 88% CI as $[-7, 6]$ but also noted that the boundary p-values inside the interval were .25 and those right outside the interval were .125. In most large sample hypothesis testing regimes, the p-value just inside the boundary of the interval are only a tiny bit larger than those outside it. In this case, our 88% CI actually could encompass an 80% CI or even a 75+ε % CI (where ε means "just a little bit") since the p-values we observe just inside the boundary are .25. Notice one feature of confidence intervals created using randomization inference on display here: The probability that a confidence interval constructed in this way contains the true value of τ is *at least* 1-α (Rosenbaum, 2002c, page 45). In this way, confidence intervals created using randomization inference are guaranteed to have correct coverage, and will be conservative if their significance level (88% or $\alpha = (1/8)$) is not exactly the same as their size. Rosenbaum (2002c, Chapter 2) also proves that these intervals are unbiased and consistent (such that more information is better — produces smaller intervals).

Composite Confidence Intervals Although we have created confidence intervals by testing null hypotheses, we have not discussed alternative hypotheses. Yet, we know that specification of the alternative hypothesis will matter a great deal for the ability of a given test to reject the null. A two-sided alternative of $\tau \neq \tau_0$ requires twice as much evidence against the null to reject it as a one-sided alternative such as $\tau > \tau_0$.

Since we create the confidence intervals by testing many hypotheses, we have many opportunities to reject or not-reject. So far, we have compared every null against the two-sided alternative. This alternative makes sense especially for the sharp null of no effects: The null of $\tau_0 = 0$ could imply no extant guesses about τ (and thus $H_A : \tau \neq \tau_0$). Yet, testing for effects implies that we are willing to consider non-zero values for τ — at least provisionally. If we posit a negative value, say, $\tau_0 = -1$, then what is the alternative? It could still be $\tau \neq \tau_0$, but, notice that we already know (from our first test of no effects) that τ could be 0, and thus we have a sense that $\tau > -1$. Entertaining $\tau_0 = -1$ after having tested $\tau_0 = 0$ implies that we are not really interested in the two-sided alternative, but in the one-sided alternative that $\tau < \tau_0$. A similar logic can apply to a hypothesis of $\tau_0 = 1$ (i.e. that this null can be thought of as really implying an alternative of $\tau > \tau_0$.) A confidence interval for $\tau = \mathbf{r}_1 - \mathbf{r}_0$ could thus be constructed as a two-sided interval for nulls where $\tau_0 = 0$ and two one-sided interval for nulls where $\tau_0 \neq 0$. Notice, this proposal involves specifying the set of alternatives at the same time as specifying the set of nulls — the alternatives would be defined only by the nulls, not by the data. Most analytic constructions of confidence intervals do not have any easy way to stitch together a confidence interval out of independent parts in this way. Rosenbaum (2008) provides more general theory that encompasses the simple example described above. In this section of the paper, we will continue to build two-sided confidence intervals, although we could make smaller confidence intervals using the approach outlined here; confidence intervals that would have the same operating characteristics of the simpler intervals.

WWLRD (What would linear regression do?) A t-test from a linear regression with dummy variables representing pairs in this case report a 95% CI of [-7.73,10.73]. Ignoring the paired structure of the data yields [-36.05,39.05]. Both of these tests require either large samples of independent, homoskedastic observations or an assumption about an independent, homoskedastic, Normal data-generating process. Freedman (2008) points out that, even in large samples, or even if the Normal DGP is roughly correct, treatment and control groups almost always have different variation on the outcome variable and thus linear regression in the case of randomized experiments does not produce correct confidence intervals. Green (2009) suggests that Freedman's concerns may have little effect on much of common large-sample and/or DGP model-based analysis of political science experiments. What we have shown in this paper so far is that a DGP process model is not necessary for statistical inference from an experiment.

Ought one to prefer a DGP process model when analyzing the results of an experiment? We have shown that it is unnecessary. It is also worth noting here it has long been known that that the validity of a DGP process model also depends on the validity of an assignment model [cite to Heckman, Achen, Rubin]. Thus, adding a DGP model to the analysis of an experiment does not imply that one can therefore spend less time on the model of assignment. In the case of a randomized, controlled trial, a DGP seems (we hope after reading this far into this paper) to be more of a burden than a help since it requires its own set of justifications and work.

2.5 Bringing the DGP back in: Model-assisted, Randomization-justified Inference

Yet, we must recall that political scientists are not ignorant about turnout. That is, although the DGP ought to seem like a burden, and the clarity and simplicity of the

randomization-based interval ought to appeal, knowledge of outcomes ought to help us somehow. More information should be better. The only question is how to use the additional information while maintaining our focus on models of assignment as the basis for statistical inference.

We introduced the rank-based transformation of R above for three reasons: First, we noted that rank-based tests tend to have more power than mean-based tests with non-normal outcomes. Second, we wanted to depart from the common use of "average treatment effects" in the causal inference literature — not because such usage is a problem, but, to show that the average is merely one possible summary of effect and thus, we hope, expanding the possibilities for applied analysts.[27] Third, we wanted to plant the idea that one could be creative about test statistics to foreshadow this section.

Recall that the procedure for randomization inference depends testing hypotheses about potential outcomes and a model of effects. Together, $H_0 : \tau = \tau_0$ and $\tau_0 = \mathbf{r}_1 - \mathbf{r}_0$ imply a particular pattern of $\mathbf{r}_0$ that we can observe by adjusting observed responses such that $\mathbf{r}_0 = \mathbf{R} - \tau_0\mathbf{Z}$. We want to know something about $t(\mathbf{Z}, \mathbf{r}_0)$ and the hypothesis and the model of effects provides $t(\mathbf{Z}, \mathbf{R} - \tau_0\mathbf{Z}) = t(\mathbf{Z}, \mathbf{r}_0)$ since we cannot observe $\mathbf{r}_0$ for all cities directly. The variance in the distribution of $t(\mathbf{Z}, \mathbf{r}_0)$ depends in part on differences in potential outcomes given different treatment assignments (i.e. a difference between treated and control subjects) but part of this variation within treated and control observations is due to covariates (observed or unobserved). Noisy outcomes will make it harder to distguish control from treated observations. Imagine that we could regress $\mathbf{r}_0$ on some set of covariates on the matrix $\mathbf{x}$ but not $\mathbf{Z}$; say these covariates are known from previous literature to predict aggregate turnout. The residuals from such a regression, $\mathbf{e}$, should be less variable than $\mathbf{r}_0$ and uncorrelated with $\mathbf{x}$ (meaning that a regression of $\mathbf{e}$ on $\mathbf{x}$ would produce an $R^2 \approx 0$.) Such a regression does not involve looking at effects of treatment, thus, protecting our inferences from concerns about data mining. But such a regression is impossible since we do not observe $\mathbf{r}_0$ for cities were $Z = 1$.

Of course, the same logic of replacing $\mathbf{r}_0$ with $\mathbf{R} - \tau_0\mathbf{Z}$ can be used to test a hypothesis about a particular configuration of potential responses to control. Rosenbaum (2002b, §4) shows us that we can define a "residual producing function" or perhaps a "de-noising function" (our terms, his idea) $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x}) = \mathbf{e}$ and, given some model of effects, one can test hypotheses $H_0 : \tau = \tau_0$ using $t(\mathbf{Z}, \mathbf{e})$. To summarize, a linear model can aid the production of randomization justified confidence intervals via the following steps:

Define a function to produce residuals $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x}) = \mathbf{e}$. This could be OLS, influential point resistent regression, or a smoother. The residuals, $\mathbf{e}$ will be calculated from fixed quantities $\mathbf{r}_0$ and $\mathbf{x}$ and so will be fixed.

Compute adjusted outcomes based on a $H_0 : \tau = \tau_0$ Since $\mathbf{r}_0 = \mathbf{r}_1 - \tau$, we can calculate $\mathbf{e}_0 = \tilde{\varepsilon}(\mathbf{R} - \tau_0\mathbf{Z}, \mathbf{x})$ where $\mathbf{x}$ is a matrix of covariates predicting $\mathbf{R}$.

[27] For those readers in the know, this paper is not a Fisherian manifesto contra Neyman. However, we are taking a Fisherian line here because we feel it is (a) particularly clear and (b) allows the Neyman work emphasizing average effects to be understood and broadened: For example, Hansen and Bowers (2009) show a Neyman-style approach to estimating causal effects using a test statistic that is not a constant average effect and that is defined from a Fisherian point of view. [cite to pages in Lehmann and Cox showing that Neyman and Fisher are both doing pretty much the same randomization inference.]

Compute $t(\mathbf{Z}, \mathbf{e}_0)$ and compare to $t(\mathbf{z}, \mathbf{e}_0; \mathbf{z} \in \Omega)$ p-values and CIs follow directly.

Although the step of using a model of outcomes to enhance the randomization inference does add more stories that require their own justification, the process of justification ought to be less burdensome than it would otherwise be if the inference itself were based on a DGP model. First, notice that one would not make this step to using a model of outcomes if one did not know something about the outcomes. That is, in the absence of knowledge about outcomes, the confidence intervals already produced suffice.

Second, this outcome model is meant only to reduce noise in the outcomes. A correct specification is not required. Incorrect specifications will add noise and will thus make the CI wider but will not change the coverage of the CI. We emphasized $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x})$ above — a noise-reduction or residual-producing function — rather than $R = \mathbf{x}\boldsymbol{\beta}$ in part because we never need to examine the coefficients let alone assess uncertainty about them in this model. Recall that the only source of randomness in this framework about which we have confidence is $\mathbf{Z}$, and $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x})$ does not include $\mathbf{Z}$. Thus, it has the status of a smoother or other data summary/description, not of a model of a data generating process. Thus, there is less to justify from the perspecitive of the technical features of the model. This is not to say that one must not be thoughtful in choosing such a model — a bad choice could inflate the confidence interval.[28]

Figure 2 shows the 88% CIs that arise from different covariance adjustment specifications. Each confidence interval is labeled with the linear model formula from our noise-reduction function, and the intervals are plotted in order of width (widest on top, narrowest on bottom). We also include three models which use draws from the normal distribution as covariates — i.e. merely including noise as a covariate. The confidence interval without covariance adjustment [-6–7] is labeled r_0 — it is the 4th line from the top of the plot. The top two lines represent the 88% confidence intervals from the two noisiest noise models (one with a draw from a N(10,5) and the other with a draw from N(0,5)). In fact, these two intervals are truncated — we did not test hypotheses beyond -10, 10. So, adding noise expands the intervals.

Conversely, covariates which predict the outcome reduces the interval: the bottom line is an 88% CI running from -1 to 5 percentage points of turnout after removing the linear and additive relationships between median age and percent black (both 2000 census figures for the cities) and post-treatment turnout.

Covariance Adjustment for Imbalance Recall that we worried in § 1.1 about about the fact that, within matched set, previous turnout was not exactly same between the two cities — and that, perhaps simple treatment-minus-control differences might overstate the treatment effect. One could also a define a test statistic which would remove the effect of baseline turnout from post-treatment turnout. For example, Table 6 replicates and extends Table 1 including data on other covariates. We might calculate the treatment effect for Sioux City of 22-16=6 given the comparison city of Saginaw. Of that 6, however, one might imagine that at least 4 points of that are due to the baseline

[28] What if we did have good information about the prior distributions of $\boldsymbol{\beta}$ from $\tilde{\varepsilon}()$ and/or we knew how the units were sampled from a larger, well-defined population (or we knew the DGP)? Perhaps then we could imagine a posterior distribution of $\mathbf{E}$ (no longer lowercase) which would itself generate a distribution over the CIs and allow for the kinds of model comparisons at which Bayesian methods often excel.

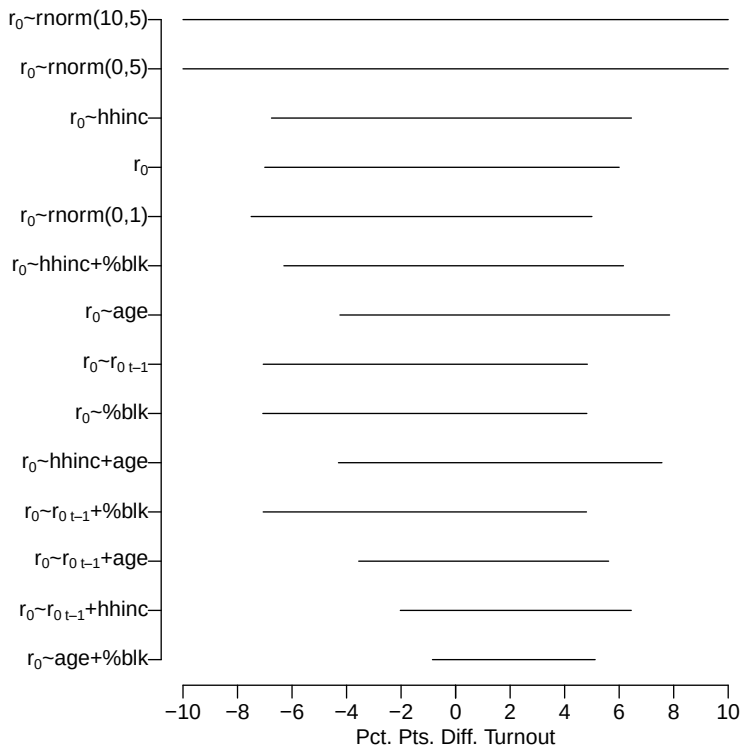


Figure 2: 88% Confidence Intervals for effects of Newspapers advertisements on Turnount (Percentage Points of Turnout). All models run with pair-aligned data [i.e. treated-control outcomes and covariates] (equiv. fixed effects for pair). Here $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x})$ uses OLS. Turnout in the previous election is labeled $r_{0\,t-1}$. Covariates "hhinc"="Household Income" and "age" are median 2000 census figures for the city, "% blk" is percent black as of 2000 census. Noise only models have covariates labeled "rnorm" and represent random draws from Normal distributions N(0,1), N(0,5) and N(10,5). The 88% CIs for the N(0,5) and N(10,5) noise models are wider than -10 to 10 but are truncated here.

difference in turnout between the cities, and so the "true" treatment effect ought to be 6-4=2. Thus, one could imagine a simple generalization of our mean differences test statistic (equation 1) which would write the baseline-adjusted effect for pair b using $\mathbf{r}_{\text{pre}}$ to mean baseline turnout as:

$$\begin{aligned} t(\mathbf{Z}, \mathbf{R}, \mathbf{r}_{\text{pre}})_b &= \left(\frac{\sum_i \mathbf{Z}_{bi}^T \mathbf{R}_{bi}}{\sum_i \mathbf{Z}_{bi}} - \frac{\sum_i (1 - \mathbf{Z}_{bi}^T) \mathbf{R}_{bi}}{\sum_i (1 - \mathbf{Z}_{bi})} \right) - \left(\frac{\sum_i \mathbf{Z}_{bi}^T \mathbf{r}_{\text{pre},b,i}}{\sum_i \mathbf{Z}_{bi}} - \frac{\sum_i (1 - \mathbf{Z}_{bi}^T) \mathbf{r}_{\text{pre},bi}}{\sum_i (1 - \mathbf{Z}_{bi})} \right) \\ &= \text{Mean Treated–Control Post-Treatment Turnout Difference in Pair } b - \\ &\quad \text{Mean Treated–Control Baseline Turnout in Pair } b. \end{aligned} \tag{8}$$

For those comfortable with linear models for estimating effects, equation 8 represents a change score or "gain score" model (i.e. $R_t - R_{t-1} = \beta_0 + \beta_1 Z$); a model which is equivalent to $R_t = \beta_0 + \beta_1 Z + R_{t-1}$.[29] And it is common to re-express such models

[29] Another strategy to account for these baseline differences would be to use percentage change as the test statistic. Consider the Sioux City versus Saginaw pair: our current adjustment would calculate

				Turnout				
City	State	Pair	Treatment	Baseline	Outcome	% Black	Median Age	Median HH Income
Saginaw	MI	1	0	17	16	43	31	26485
Sioux City	IA	1	1	21	22	2	33	37429
Battle Creek	MI	2	0	13	14	18	35	35491
Midland	MI	2	1	12	7	2	36	48444
Oxford	OH	3	0	26	23	4	21	25164
Lowell	MA	3	1	25	27	4	31	39192
Yakima	WA	4	0	48	58	2	31	29475
Richland	WA	4	1	41	61	1	38	53092

Table 6: Design,covariates and outcomes in the Newspapers Experiment. Treatment with the newspaper ads is coded as 1 and lack of treatment is coded as 0 in the 'Treatment' column. % Black, Median Age, and Median HouseHold Income from the 2000 Census. Numbers are rounded integers.

without requiring that the baseline outcome enter with a fixed coefficient of 1 so that we might write $R_t = \beta_0 + \beta_1 Z + \beta_2 R_{t-1}$. The lines in Figure 2 with the covariate $r_{0\,t-1}$ reflect exactly this kind of baseline adjustment — only without basing inference for the effect of treatment on post-treatment outcomes on the β_1 in a linear model.

Notice also in Table 6 that other covariates display some imbalance: especially notice that the percent black in the treated city is higher than the percent black in the control city in each pair. Such a pattern might make us worry about the randomization procedure applied here. It turns out that one can use exactly the same procedures shown earlier to test the sharp null hypothesis of no relationship between percent black and treatment assignment. The p-value for this test is .125: the null of no effect would fall just outside an 88% confidence interval as implausible (the 88% CI based on the same exact rank-based test as used here runs from -41.9 to -.1). What this really means is that differences in black percent are not large enough to greatly confuse confidence intervals based on treatment assignment. Yet, the relationship between treatment and percent black and turnout is strong enough to cause adjustment for it to be even more powerful than adjustment for baseline outcomes. In general, this framework for statistical inference can be used for placebo tests or randomization procedure assessments just as it can be used to produce plausible ranges of treatment effects [cite to Abadie and Sekhon and Titiunik on placebo tests]

2.5.1 Discussion

Rosenbaum reminds us that:

> Although randomization inference using the responses themselves and ignoring the covariates yields tests with the correct level and confidence intervals with the correct coverage rate, more precise inference might have been possible if the variation in fitted values had been removed. (Rosenbaum, 2002b, page 290)

(22-16) - (21-17)= (22-21)-(16-17)=2 as the adjusted effect. We could also use percentage changes: for example, if a treated city had a turnout of 2 at baseline but a turnout of 4 after treatment we might say that turnout doubled, or that the percentage change was 4/2=200%. So, in our case we would have (22/21) - (16/17) = .1 or a positive change from baseline of 10%. I don't pursue this strategy here because it is harder to decode, for example: 10% change in this case is the same as an adjusted turnout change of 2 percentage points — which is the more substantively meaningful metric in any case.

We have shown how to make our inference more precise using substantive knowledge. Although, in exchange for methodological discussion, we did not carefully argue for our covariance adjustment model using substantive knowledge here. And, we displayed a variety of such models to illustrate how the confidence intervals react to such adjustment, whereas a substantive analyst would, presumeably just show one such interval in addition to the simpler, unadjusted interval.

Notice that we cannot include all of our covariates in this model: after accounting for pairs we have at most 3 degrees of freedom. And using all 3 reduces the variability in the residuals to nearly zero — thus making the null of no differences between treated and controls on the residuals always very plausible. That is, we speculate that, in this small experiment, part of the variance in the outcome that is being removed as extraneous to the treatment effect is in fact related to the treatment effect — and that, in the extreme, one can remove so much noise from the outcomes as to make any differences between treated and control disappear. Thus, we do not adjust for baseline-outcomes *and* % black *and* median age, but only for subsets of 2 of these variables at a time here.

In his article introducing these ideas, Rosenbaum (2002b), eschews OLS and instead uses Huber and Ronchetti (2009)'s robust regression to down-weight high leverage observations. The Newspapers study does not display particularly high leverage points in any of the regressions run above, thus, when we replicated the OLS models using the M-estimator, we saw no changes. [30]

The potential for covariance adjustment to overfit the data and the potential for a few observations to unduly influence the fit suggest some limits to the application of this approach to small datasets like the one used here. The imbalance on some of the covariates was adjustable, but the adjustment also caused the CIs to "bounce" or not to remain exactly centered over the same central range of values. Yet, the imbalance was not severe enough (and the information in the data was small enough) such that the narrower plausible ranges of treatment effects tended to also to be plausible in the lights of the wider intervals.

These limitations do suggest some avenues for elaboration and extension of these procedures. Bayesian covariance adjustment would answer many of the concerns about overfitting and might allow more substantive information to be brought to bear on the problems of influential points.[31]

2.6 Credible Inference is Possible without a DGP in a Small Randomized Field Experiment

We have shown that statistical inference does not require a model of outcomes or a model of effects as long as (1) one restricts attention to tests of the sharp null of no effect and (2) one believes one's model of assignment. Inference about effects (rather than no effect) requires adding substantive knowledge about the process relating treatment

[30] We use the term "leverage" here to refer to the fact that we are worried about observations that are very different from others on the scale of the covariates, not necessarily highly "influential" observations which also might have a large effect β on a linear model. When we replicated our analyses we followed Rosenbaum (2002b, § 2.4) in using Huber's M-estimator as implemented in the `rlm` function packaged with `R`.

[31] For binary outcomes, see especially the promising ideas in (Gelman et al., 2008) or alternatively the frequentist development of shrinkage models as described Zorn (2005).

to potential outcomes (i.e. a model of effects). Rejecting null hypotheses about effects may either indicate that $\tau \neq \tau_0$ or that the model of effects is wrong. Inference can be made more precise if the analyst knows something about the outcomes and uses this information for noise-reduction. At no point did we rest the validity of the coverage of the confidence interval on a DGP process model or asymptotics, although we did add more and more structure to the inference. What is nice about this, we think, is that one can base statistical inference in experiments on some very credible stories. Although most confident with tests of the sharp null of no effects, one is not restricted to such tests, and, in fact, DGP-models can be harnessed so that inference about effects is targeted as precisely as possible on those parts of the outcome not determined by well-known covariates.

Did newspaper advertisements matter for vote turnout in the cities studied? The narrowest range of plausible values — created using a composite interval following § 2.4.3 — yielded a plausible range of effects of [-1,4]. The two-sided 88% interval ranged from -1 to 5.[32]

This investigation of the Newspapers experiment also yields a few implications for design of studies. First, small samples do not require analysts to eschew probabilistic inference. We have long known that Bayesian inference is possible in small samples. And this paper shows another mode of inference that may also be useful.[33] Second, this study re-emphasizes the importance of substantive information about the outcome of interest in helping "permitting the phenomenon under test to manifest itself" (Fisher, 1935, page 24). Assigning treatment within sets of units which are homogeneous on the outcome enhances the precision of estimation.[34] Thus, in study with few units it is wise to stratify before assigning treatment. Substantive information is also important in allowing for effective covariance adjustment and in justifying the model of effects that is maintained in order to test effects.

In a sense, randomization inference allows scientific attention to return back to the political phenomena of interest: what causes what? what counter-factuals ought we to entertain? what units received the treatment, in what way? That more information is better within this framework is a good thing. In the next section we engage with the question of applying this framework to an observational study. There especially we will see that inference based on a model of treatment assignment (there, more like treatment selection than assignment), requires yet more social scientific justification and elaboration. In that case, although we will have dramatically less confidence in our model of treatment than we do here, in a randomized study, we will suggest that inference in that case based on a model of treatment may still be more compelling and scientifically useful than a inference based on a model of outcomes.

[32] These intervals were actually tested with hypotheses ranging from -10 to 10 in intervals of .1 but the results reported here are rounded to integers for ease of reading. The actual ranges were: two-sided interval = [-.8, 5.1], composite interval=[-.7, 3.7].

[33] There is also some interesting recent research on higher order asymptotics which appears to allow DGP based inference to proceed in small samples as well Brazzale et al. (2006); Strawderman (2000).

[34] Any textbook's treatment of the paired versus unpaired t-test will attest to this fact dramatically (See, for example Rice, 2007, § 11.3). Rosenbaum (2005) shows this nicely, but we have long known about this from at least Kish (1965), [find part of Campbell, Stanley, let alone Blalock, etc..], and also have recent advice re-emphasizing this fact from Imai et al. (2008).

3 Did the regulatees capture the regulator? The effect of PDUFA on Drug Safety

[Incomplete section]

3.1 Building and Justifying a Model of Selection

On what grounds might we claim that we have more confidence in a model of selection into the PDFUA regime rather than a model of drug safety outcomes? A common approach to a model of drug safety outcomes would to inspect Figure 1 (or some summary statistics otherwise describing the empirical distribution of the outcome variable) and then to choose from some list of probability distributions a candidate data generating process. This process would be parameterized with some linear function ($\boldsymbol{\beta}\mathbf{x}$) suitably transformed. And inference would be justified by the correctness of those two choices. The FDA dataset is not a sample of a population, but rather a census of drugs, so in this case appeals to sampling from populations would not be reasonable justifications for inference.

What about basing our inference on a model of selection? First let us posit a simple model: $\Pr(Z_{is} = 1|\mathbf{x}) = 1/n_s$. That is, we might claim that conditional on some covariates $\mathbf{x}$, the probability of a company deciding to submit a drug before versus after the PDUFA act is constant. Say, these covariates included something like "date range" of drug submission. Could we claim that in some narrow range of dates around the policy change, drugs submitted just before the change were only submitted then due to chance — that their probability of being submitted after the date was the same as their probability of being submitted before the date?

On its face, this model of selection doesn't make sense. Drug companies have a lot (really really a lot) of money at stake in their drugs. If any actor has an interest in acting strategically and not randomly, even in small windows of dates, it would be drug companies interacting with the government and each other. What will be very interesting here, is that we will claim that, in fact, it is precisely these kinds of constraints which make the "as if random in a small window" model of selection a good description of what was happening in and around the time of PDUFA. In a more substantively oriented article we might develop a formal theory to make clear how our model of selection maps onto the actors and their environment. Here, we merely provide a few pieces of evidence to gesture at how we might imagine a substantive research might go about building a model of selection on firmer ground than a model of a data generating process.

3.1.1 Justifying the posited model of selection

Informants We had two in-depth interviews with decisionmakers at a major pharmaceutical company. One of them had been an FDA reviewer during roughly 1995-2000, so she could both about her company's current practices regarding the FDA (she, like many ex-FDA reviewers, is in charge of the relations between this company and the FDA) as well as talk about what she experienced as an FDA reviewer during the first PDUFA policy regime. The other informant (at this same company, unfortunately) has been part of teams making decisions about when to go forward with drug development, trials, and FDA submission as well as post-submission marketing. She was able to discuss the point of view of the business decision makers regarding FDA submission timing.

Basic results: (1) The FDA submission process involves what are essentially contracts written between the companies and the FDA regarding the details of the designs of the clinical trials. That is, the FDA and the drug companies agree many months in advance how long a trial will last, among many other aspects of the pre-submission aspects of the drug studies. There is no way to end a trial early or late other than for emergencies (ending it early because of unforeseen safety problems, for example). Thus, the drug companies do not have fine-grained control over the timing of their drug submissions. (2) The amount of money earned by the companies in the first few months after a drug is approved is so huge (so the informants tell me) that it is a strictly dominant strategy to submit as fast as possible: they have their own formal theorists, and their own formal theorists are telling them this (or at least, this is what my informants say). (3) Even if they thought they could get an unsafe drug through the FDA process by delaying it, they wouldn't because (a) they think of themselves as saving lives not pushing dangerous snake-oil and (b) the money lost would be immense.

Media Coverage in the Year(s)/Month(s) Before PDUFA As far as we can tell from Lexis-Nexis searches, the legislative history of this act begins just about 2–3 months before the act is voted on in Congress. That accords with anecdotal evidence from Carpenter [cite] that the bill appeared in the limelight fairly suddenly and gained support from the drug companies quickly. All of this occuring much too quickly for any company to change submission or clinical trial plans even if they could. [Add data from Lexis-Nexis searches.]

Sudden changes in submission activity before/after PDUFA We observe the timing of all submissions and approvals. If drug companies thought it was in their interest to delay submission until after the act (or speed up submission before the act), and they somehow, contra other information, did have fine grained control over their submission process, then we ought to see in the data large changes in the months before and/or after the act. [TO DO]

These pieces of information make us comfortable (at least for a methods paper) moving ahead with the presumption of a very simple model of selection. Notice that this model is based on social science.[35] Here, the social science is quick, but it does involve large-N analysis, rough content analysis, and in-depth interviews. That is, the model of selection is something posited and then explored and justified. Next we will bolster this model with more research design — post-stratification.

3.1.2 Seeking a latent experiment by matching

Comparability in the Newspapers experiment was justified by the design of the experiment. Here we aim to justify a comparability argument, but we will impose a design on data already collected rather than collecting the data according to a pre-specified design. This problem, of facing a set of data that history has generated, and with history out of the control of the analyst, is common in political science. Design, for such observational studies, involves exercising "choice as an alternative to control" (Rosenbaum, 1999). Here, we use tools such as matching and propensity score models to instantiate the choices that we would like to make based on our model of selection.

[35] Thanks much to Walter Mebane for helping us realize this.

Instantiating our model of assignment: Propensity Score Models

Matching

Balance Testing

3.2 Estimating Treatment Effects using the Design implied by the Model of Selection and Resulting Matched Sets

3.3 Sensitivity Analysis

3.4 Summary

4 Discussion

We have shown that statistical inference is possible in the absence of a data-generating process model and in the absence of large-samples. Some very basic form of statistical inference is possible even without a model of effects. We have also shown that this form of inference can be made more precise with linear models. Those who know the history of statistics will not find this surprising, although they might be happy to see how the work of Neyman and Fisher can be extended to handle modern data analysis problems. Others, whose only training in statistics has occurred within departments of political science, sociology, or economics, will be, we hope, pleasantly surprised.

Randomized experiments make models of assignment more credible (in general) than observational studies. And randomized experiments are a natural place to apply these methods.[36] Observational studies do not allow control over assignment, although they offer opportunities to choose observations and contexts so as to make relationships between selection (or assignment or occurrence) and outcomes clear. Yet, we do not think that observational studies make data-generating process models more credible than models of selection a priori. We have tried to demonstrate, with the FDA data, an example in which a data generating process model might be hard to construct and justify and in which a model of selection was comparatively easier.

In cases where physical randomization has occurred, then it is certain that the scholar knows a lot more about the assignment mechanism than anything else. In that case, there are few arguments against using randomization inference. In observational studies, we could be wrong about the assignment/selection mechanism. That is, we might say that, within matched set, we cannot distinguish the data observed from that which would have been produced from an experiment with fixed assignment probabilities (conditional on numbers so assigned). But, we might have other assignment mechanisms which would be similarly plausible — if we represented the strict nulls implied by other assignment mechanisms, it is possible that we would not reject those either. What does this mean for the confidence we would have in our results in such cases?

To answer this question we have to ask, "Compared to what?" Imagine a case in which the scholar had a formal derivation of the stochastic process producing the outcomes, and a formal derivation with very strong empirical grounding, but with little knowledge about how units selected values of the causally important variable. In that case, in some sense, a data generating process based approach would dominate a

[36] Keele et al. (2008) provide a detailed argument in favor of applying these kinds of techniques to laboratory experiments.

(latent) randomization based approach. I say, "in some sense", because ignorance about a selection process is sure to produce biased estimates of causal effects in any mode of inference (cite) and so that correct data generating process model must include a credible selection process model too.

If the analyst knows enough about the selection process to feel confident in treatment effect estimates arising from a data-generating process based approach, then, of course, the question is raised whether that analyst would prefer to use that knowledge in a randomization based approach (using fewer other assumptions) or to use that knowledge in a more model-based approach (i.e. to have a selection equation in addition to an outcome equation).

Perhaps there are intermediate situations where confidence in assumptions about the probability distribution of the outcomes outstrips confidence in assumptions about the probability distribution of the selection process. In those case, we would advocate the DGP based approach.

The general point is *scholars should base their estimates of causal effects on the assumptions about which they have the most confidence; and the assumptions that are the easiest to more compellingly justify.*

To this end, we think offering scholars an alternative mode of inference can only improve data analysis and knowledge communication and cumulation. We do not require all people all the time to use these techniques. And we do not think that data generating process based methods lack utility in general. The question that we encourage data analysts to ask themselves, however, is which method offers assumptions which are the easiest to justify in a particular research design and theoretical context. In the end, we think that greater diversity of modes of inference will help the whole discipline and will help raise standards — if one does not have options then one not might feel compelled to justify one's choices.

This then, is our argument for using randomization inference in observational studies: (1) if one post-stratifies, one can, in theory, design a configuration of data that is statistically indistinguishable from an experiment and just as a physical experiment provides the "reasoned basis" for inference, a latent experiment created using matching provides another such reasoned basis (if also a basis that deserves more checking); (2) if one is not post-stratifying, but has more worries and uncertainty about the choices of likelihood, linear model, and/or prior, required of a data generating process then making assumptions about the selection process may allow more confidence than making the other kinds of assumptions [here, it is a choice among assumptions all of which the analyst has reason to doubt, just some more than others]. In addition, we offer the advice about matching and balance assessment which offers a way to check and strengthen arguments for the assumptions about the selection process where the other modes of inference do not offer as direct and clear diagnostic tools about their central assumptions.[37]

[37]The best extant methods, for example inspecting the predictive posterior distribution [a technique accessible to both Bayesians and model-based frequentists] is a kind of rough tool. Lack of fit can be due to either an incorrect parameterization of the likelihood, an incorrect prior, and incorrect likelihood, or even problems of convergence and computation.

4.1 Common Questions and Concerns

4.1.1 What about Time-series? Or Factor Analysis?

Past work shows how these basic ideas can be extended: to instrumental variables (Imbens and Rosenbaum, 2005), to clustered assignment within pairs (Imai et al., 2008), to clustered assignment with instrumental variables and a binary outcome (Hansen and Bowers, 2009), to other multilevel designs (Braun, 2003; Small et al., 2008) to longitudinal data (Haviland et al., 2007), and to cases with interference between units (thus violating a part of the SUTVA assumption) (Rosenbaum, 2007).

We have not seen randomization inference used in situations with extensive non-random missingness in the outcomes in addition to non-random selection of units into complex intermediate outcome classes (labeled by Frangakis and Rubin (2002) as "principal strata"). Nor do we know about extensive applications to complex structural models of causality or measurement (i.e. no time-series, no structural equation models, no IRT/factor analysis). There is nothing inherent about randomization inference which would prevent application elsewhere, but it has just not received the kinds of sustained attention that the linear model has received over the past 50 years.

4.1.2 How can we learn about the population by focusing on the sample?

By conditioning our statistical inference on the sample at hand, one might worry that we compromise the scientific importance of conclusions emerging from the analysis. That is, one might ask, "Who cares that the effect of some causal process was τ in your sample? I care about the population, or about future time points, or about other possible samples." We think that external validity is as important as internal validty in the eventual judgement of the scholarly community about the value of a study or theory or finding.[38] We don't believe, however, that one is required to answer the question, "What happened in your study?" in terms of a (possibly unknown and undefined) population let alone other samples (across time and/or space). In fact, we feel that randomization inference *improves* the ability of researchers to talk about the general from the specific. By offering researchers the ability to have great confidence that what they are seeing in their specific sample is actually what was happening in that sample (or that their sample does not provide enough information to offer much information about their chosen test statistic), researchers have a much more solid sense of the specific from which to address the general.

We do believe that statements about causality as regards a social scientific theory must involve consideration of the general (as long as the general is defined clearly). That is, a clear effect of a treatment in only one sample, once, and that is never replicated or replicable, does not tell us as much about a general causal theory as such an effect which can be replicated and is replicated. (And perhaps, more strongly and also more realistically, any given causal theory ought to imply a host of observational phenomenon and effects — some of them are easier to see in certain situations and samples than others — and what happens in one sample, in fact, then, ought to reflect directly on the theory without any reference to external validity [cite Rosenbaum].) That is, imagine we were skeptical about the equation governing the velocity of a ball dropped from a height. That equation (an operationalization of a theory) ought to imply something about *this ball* from *this height* and if we don't observe what we'd expect to see in *this*

[38] For discussions of "external" and "internal" validity see (cite Campbell or others).

specific instance then the theory is cast (perhaps weakly) into doubt. And the precision and confidence with which we can talk about this specific instance adds to the strenght of this doubt. Of course, replication would also add to the doubt. But, it is our hunch that replication would strengthen the doubt less so on an occasion by occasion basis than the research design and analysis of a particular critical implication of the theory.

In general, thinking that a linear specification or a likelihood or a prior offers the analyst any inherent confidence in generalizability is, we think, wrong headed. Well justified likelihoods and smooth specifications and priors may well offer us more confidence that our findings in a *given study* are the result of occurrences in the world and the research design rather than an accident of a series of modelling choices. But again, the confidence we desire is about the effects in a given study. And the more confidence we have in this study's findings, the more confidence we will have when we turn to thinking about generalizing these findings; when we turn to asking what these particular findings mean for some well-defined general.

One could imagine a workflow which, given a general theory and empirical implications of it, would (1) assess the implications using a particular research design and data analysis where both design and analytic methods are chosen so as to maximize the clarity and confidence of the assessment, and (2) then the scholar would ask, "What do these findings mean for our doubt about this theory as it would apply elsewhere in the domain of application?" and (3) perhaps this question would spawn other implications which would themselves be testable in the specific and particular, and the question about the theory and general upon confrontation with the next set of confident and clear results would again raise the next set of questions and perhaps new theory.

To imagine that a single test or a single moment of data analysis can answer both "What happened in this instance?" and "What does this finding mean for other instances (places, times, circumstances)?" seems to us to ask too much of weak tools let alone weak theory (and weak humans).

What this all means is that (1) by recommending randomization inference we are deeply concerned about external validity, (2) and that we think in fact that claims about external validity and theoretical importance of specific findings are enhanced by careful attention to the specific and particular first and foremost (after careful thought about the theory and implications).

4.1.3 Testing the many hypotheses that comprise a confidence interval sounds like a big multiple comparisons problem.

Although we are testing many hypotheses to generate a confidence interval, we are only testing each hypothesis once. We do not have to adjust the p-values for these tests to account for multiple testing (the way one might want to do so in the case of running many regressions with different predictors, hunting for one with a p-value of less than .05 upon which to build a post-hoc story about how the outcome comes about). The CI creation is driven by pre-specification of a Type I error rate and of null hypotheses *not* by exploration of the data. This pre-committment protects CIs created from such direct inversion of hypothesis tests from multiple testing problems.

4.1.4 How should one choose a test statistic?

Randomization inference offers the advantage that an analyst may choose a test statistic that ties tightly with the substantive concerns facing the analyst and also among test statistics of equivalent substantive meaning, can choose among them for the one which will produce the tightest confidence intervals. This flexibility can be see as a disadvantage or an element of arbitrariness:

> Such [randomization-based] P-values can be calculated for: any sharp null hypothesis ..., any test statistic (e.g., the median difference in matched-pair differences), and any a priori definition of unusualness (e.g., squared value greater than observed squared value). Consequently, there is a great deal of arbitrariness (flexibility in the hands of a knowledgeable and wise analyst) except by convention (Rubin, 1991, page 1222)

But flexibility is also an advantage. Whereas in linear models the analyst is not asked "why a slope or a predicted probability?", analysts using randomization inference bear the burden of explaining of why they chose their test statistic. In exchange for justifying their choices they may choose any function of treatment assignment/selection and outcomes that is most meaningful to their work and easiest to communicate to their audiences. Of course, and perhaps most importantly, different test statistics will have different power in different designs. Thus, a common convention is to use multiple different test statistics, reporting them all.

4.1.5 Is this useful for small-N studies?

If an analyst has few observations, perhaps with non-numerical values assigned to variables, that analyst may feel like (1) such small-N, qualitative data are not amenable to probability statements and (2) even if the image/text/audio/video observations were recoded into some numeric form and made into a dataset, the paucity of rows in the matrix would make such probability statements meaningless. Of course, neither of these intuitions is correct in general. Although some may prefer not to make formal probability statements for other reasons, such statements are possible to make within either the randomization inference frequentist framework or the Bayesian framework. Neither requires many observations, although both will provide wide intervals in the absence of much information. And neither requires explicitly numeric coding (although in software, numbers may well be assigned to values like "oranges", "apples" or "revolution" as temporary placeholders). The fact that Bayesian inference does not require large samples is well known. With randomization inference, scholars with small number of observations have yet another tool-box that they could use to make probability statements should those statements be desirable.

Randomization inference, in addition, does not require "covariation" or associated concepts as a driving metaphor. Since any function of what we've called "treatment assignment" and observed outcome can produce a test statistic, scholars who feel uncomfortable with linear models or who do not desire to specify complex probability models a priori, can engage in hypothesis testing and produce confidence intervals based mainly on their knowledge of the "selection" or "assignment" process. Thus, for example, if a scholar of in a comparative historical study were comfortable listing all of the ways that the causally important set of variables could have occurred, then that scholar has some of the basic ingredients for statistical inference based on this model

of assignment/selection/occurrence (even if the model is same as Ω rather than some more pithy statement).

4.2 Conclusion

[Ringing conclusion]

References

Achen, C. H. (2002). Toward A New Political Methodology: Microfoundations and Art. *Annual Review of Political Science*, 5:423–450.

Agresti, A. and Min, Y. (2005). Frequentist Performance of Bayesian Confidence Intervals for Comparing Proportions in 2× 2 Contingency Tables. *Biometrics*, 61(2):515–523.

Berk, R. (2004). *Regression Analysis: A Constructive Critique*. Sage.

Bowers, J., Carpenter, D., and Hansen, B. (2007). Drug safety and deadlines: Causal inference about a regulatory regime change at the fda. Presentation at the Annual Meeting of the Robert Wood Johnson Scholars in Health Policy Research.

Bowers, J. and Hansen, B. B. (2006). Attributing effects to a cluster randomized get-out-the-vote campaign. Technical Report 448, Statistics Department, University of Michigan.

Box, G. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430.

Brady, H. (2008). Causation and explanation in social science. *Oxford handbook of political methodology*, pages 217–270.

Braun, T. M. (2003). A mixed model formulation for designing cluster randomized trials with binary outcomes. *Statistical Modelling: An International Journal*, 3(3):p233 –.

Brazzale, A. R., Davison, A. C., and Reid, N. (2006). *Applied Asymptotics*. Cambridge University Press.

Carpenter, D. (2004). The political economy of fda drug review: processing, politics, and lessons for policy. *Health Affairs*, 23(1):52–63.

Carpenter, D., Bowers, J., Grimmer, J., Moffitt, S., Nall, C., and Zucker, E. J. (2007). Deadline effects in drug regulatory review. Robert Wood Johson Health Policy Scholars Working Paper # 35.

Carpenter, D., Moffitt, S., Moore, C., Rynbrandt, R., Ting, M., Yohai, I., and Zucker, E. (2009). Early entrant protection in approval regulation: Theory and evidence from fda drug review. *Journal of Law, Economics, and Organization*.

Fisher, R. (1935). *The design of experiments. 1935*. Oliver and Boyd, Edinburgh.

Frangakis, C. and Rubin, D. (2002). Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29.

Freedman, D. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed. edition.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383.

Green, D. P. (2009). Regression adjustments to experimental data: Do david freedman's concerns apply to political science? Unpublished Manuscript.

Grimmett, G. and Stirzaker, D. (1992). *Probability and Random Process*. Oxford University Press, New York, 2nd edition.

Hansen, B. B. and Bowers, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236.

Hansen, B. B. and Bowers, J. (2009). Attributing effects to a cluster randomized get-out-the-vote campaign. *Journal of the American Statistical Association*, to appear.

Haviland, A., Nagin, D., and Rosenbaum, P. (2007). Combining Propensity Score Matching and Group-Based Trajectory Analysis in an Observational Study. *PSYCHOLOGICAL METHODS*, 12(3):247.

Hodges, J. L. and Lehmann, E. L. (1964). *Basic Concepts of Probability and Statistics*. Holden-Day, San Francisco.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, 81:945–970.

Huber, P. and Ronchetti, E. (2009). *Robust statistics*. Wiley-Blackwell.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, 27(24).

Imai, K., King, G., and Nall, C. (2008). The essential role of pair matching in cluster-randomized experiments, with application to the mexican universal health insurance evaluation. *Unpublished manuscript, submitted to Statistical Science. http://gking. harvard. edu/files/abs/cluster-abs. shtml*.

Imbens, G. W. and Rosenbaum, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):109+.

Keele, L., McConnaughy, C., and White, I. (2008). Statistical inference for experiments. Unpublished manuscript.

King, G. (1989). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Cambridge University Press, New York.

Kish, L. (1965). *Survey Sampling*. John Wiley and Sons, New York, NY.

Lehmann, E. (1998). *Nonparametrics*. Springer, revised first edition.

Nardinelli, C., Lanthier, M., Temple, R., and Carpenter, D. (2008). Drug-review deadlines and safety problems. *The New England Journal of Medicine*, 359(1):95.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9 (1923). *Statistical Science*, 5:463–480. reprint. Transl. by Dabrowska and Speed.

Panagopoulos, C. (2006). The impact of newspaper advertising on voter turnout: Evidence from a field experiment. Paper presented at the MPSA 2006.

Ramanathan, R. (1993). *Statistical Methods in Econometrics*. Academic Press, San Diego.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 2nd edition.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, 3rd edition.

Rosenbaum, P. (2007). Interference Between Units in Randomized Experiments. *Journal of the American Statistical Association*, 102(477):191–200.

Rosenbaum, P. (2008). Testing hypotheses in order. *Biometrika*, 95(1):248–252.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies (with discussion). *Statistical Science*, 14(3):259–304.

Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88:219–231.

Rosenbaum, P. R. (2002a). Attributing effects to treatment in matched observational Studies. *Journal of the American Statistical Association*, 97:183–192.

Rosenbaum, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.

Rosenbaum, P. R. (2002c). *Observational Studies*. Springer.

Rosenbaum, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician*, 59(2):147–152.

Rubin, D. (1974). Estimating the causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.*, 66:688–701.

Rubin, D. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3):808–840.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47:1213–1234.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–331.

Sekhon, J. (2005). Making Inferences from 2× 2 Tables: The Inadequacy of the Fisher Exact Test for Observational Data and a Bayesian Alternative.

Sekhon, J. S. (2008). Opiates for the matches: Matching methods for causal inference. Unpublished manuscript.

Small, D., Ten Have, T., and Rosenbaum, P. (2008). Randomization Inference in a GroupRandomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance, and Quantile Effects. *Journal of the American Statistical Association*, 103(481):271–279.

Strawderman, R. (2000). Higher-Order Asymptotic Approximation: Laplace, Saddlepoint, and Related Methods. *Journal of the American Statistical Association*, 95(452):1358–1364.

Zorn, C. (2005). A Solution to Separation in Binary Response Models. *Political Analysis*, 13(2):157–170.