Fisher's randomization mode of statistical inference, then and now.

Jake Bowers [*] Costas Panagopoulos[†]

May 14, 2011

Abstract

How is statistical inference possible when $n = 8$? How can we infer without a sample from a population? How should we choose methods for assessing causal claims when we have low information (like a small sample, a binary outcome, a multilevel design with few clusters, or a weak instrument)? R. Fisher answered these questions in 1935 showing that valid small sample hypothesis tests are possible, inference does not require a population, and choices about assessing causal effects can arise from design.

This paper reframes and extends Fisher's method, showing that it is a practical alternative for political scientists. As an example, we show how to assess treatment effects using a field experiment of the effect of newspaper advertising on aggregate turnout with only eight observations. In the end, we produce confidence intervals using linear models, but requiring none of the standard assumptions of linear models to guarantee valid statistical inferences.

keywords: randomization inference; analysis of experimental data; covariance adjustment; small sample statistical inference

[*] Assistant Professor, Dept of Political Science, University of Illinois @ Urbana-Champaign (jwbowers@illinois.edu). *Acknowledgements:* Thanks to Chris Achen, Joe Bowers, Dan Carpenter, Wendy Tam Cho, Tommy Engstrom, Mark Fredrickson, Don Green, Ben Hansen, Jim Kuklinski, Cara Wong. Part of this work was funded by NSF Grants SES-0753168 and SES-0753164 (collaborative grants with Ben Hansen). Thanks are also due to participants in seminars, talks, and workshops at the Center for Political Studies at the University of Michigan, the Experiments in Governance and Politics network at Columbia University, the American Sociological Association Methodology Section and the Institute for Government and Public Administration at the University of Illinois at Urbana-Champaign.

[†] Assistant Professor, Dept of Political Science, Fordham University

1 Introduction

> [I]t may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged. (Fisher 1935, 21)

In 1935 R. A. Fisher presented a general method for assessing scientific claims using significance tests. The tests themselves were to be justified based only on the physical act of randomization rather than sampling from a population, required no large sample assumptions (they were demonstrated in the context of 8 cups of tea), involved no functional model linking treatment to outcomes, nor a probability model of outcomes.

While appreciation for Fisher's design insight — that randomization produces comparable groups across which unconfounded causal inferences may be made — has spurred randomized studies throughout political science (Druckman et al. 2006), few have made use of his closely linked statistical insight. In part this lack of popularity arises from Fisher's own exposition of the method in an era of slow and scarce computing resources. In part our failure to take advantage of the promises of his method arises from a lack of fit between our current conceptual apparatus relating causal to statistical inference and Fisher's ideas. This paper aims to make Fisher's randomization inference conceptually and computationally available to political scientists. We first explain, briefly, the nature of Fisher's randomization inference as invented in 1935 to emphasize the promise of this approach. We then show that what might appear to be an overly simple idea ending with the test of a single null hypothesis can be elaborated to enable political scientists to execute nearly any kind of analysis, but with the added confidence and benefits that flow from using randomization not only for causal inference but also for statistical inference. By the end of the paper, we will show that valid hypothesis tests and confidence intervals can arise from regression analysis using small samples and without the kinds of nuisance assumptions that otherwise tend to be required by linear models.

2 The Nature and Potential of Fisher's Randomization Inference as of 1935

To explain Fisher's randomization-based statistical inference we use a randomized field experiment of eight cities.

2.1 The Nature of Fisher's Randomization Inference as Applied to a Small 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 (Panagopoulos 2006). This was, to our knowledge, the first experiment to investigate the impact of newspaper ads on turnout. This small pilot study involved eight cities, matched into pairs based on similarity of proportion turning out to vote in the previous election and a few other covariates. One city in each pair was assigned at random to receive newspaper ads in local newspapers encouraging citizens to vote. All cities assigned to receive advertisements had advertisements run in the local newspapers. Table 1 shows all of the observations in the study with key design and outcome features. In three of the four pairs turnout after treatment was higher in the treated city than the control city.

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

Table 1: Design and outcomes in the Newspapers Experiment. The Treatment column shows treatment with the newspaper ads as 1 and lack of treatment as 0. Panagopoulos (2006) provides more detail on the design of the experiment.

Was the effect of the advertisements manifest or negligible? Manifest or negligible compared to what? Fisher's answer was, in essence, "Compared to theory." He suggested the use of a substantively motivated claim about a state of the world as an object against which the nature of our observed effect might be judged. In particular, he focused on one such claim, the hypothesis that the treatment had no effects. In other words, if we compare our observed state of the world with a theoretically motivated claim about way our manipulation ought to work in the world, we will have a way to talk about how surprised we should be upon observing extant differences between treated and control cities. In Fisher's language, an effect is manifest if a hypothesis of no effects makes the effect surprising. There are three parts to Fisher's hypothesis test: a hypothesis, a summary of an observed relationship (often called a "test statistic"), and a probability distribution. The hypothesis and design imply a distribution for the test statistic. And the probability distribution quantifies how surprising it would be to observe the test statistic if the null were true. Let us explain these three parts briefly.

A hypothesis provides a conceptual standard for assessing the 'no effects' question: 'No effects' means no effects. What does 'no effects' mean? Fisher (1935, Chap 2) suggests that a treatment has no effect when units would display the same outcome regardless of treatment condition: for example, 16% of Saginaw citizens would vote if given advertising and Sioux City would show 22% turnout absent advertising.

A test statistic summarizes observed data. And we might summarize what we observe about the overall relationship between advertising and turnout with the difference in mean turnout between the treatment and control groups — 29.25 − 27.75 = 1.5 — as our test statistic.

The counter-factual depends on treatment assignment. How surprising is a 1.5 percentage point difference from the point of view of strictly no effects? A phenomenon is surprising when it is not common or expected. So, what kinds of mean differences in turnout ought we to expect if the null were true? What generates variation against which surprise can be assessed? Although there are many ways to conceptualize this variation (if we observed new samples of cities during this election, observed different elections for these same cities, etc...), the one variation producing operation that we can easily formalize in this study is the act of assigning treatment to one city within each pair. As is not uncommon in political science applications, other sources of variation are more vague and thus more difficult to formalize enough to generate a probability distribution — for example, neither the one election nor the eight cities are a sample drawn with some known sampling plan or known probabilistic mechanism from some clear population.

Repeating the experiment produces a distribution of test statistics under the null. So, let us consider what would happen if we were to repeat this experiment, randomly assigning a different set of treatments within pairs. If, by chance, we swapped the observed assignments in all pairs, and represented the null hypothesis of no effects by leaving the outcomes unchanged regardless of treatment assignment, the difference in mean turnout would be -1.5. If we kept the same assignments, but merely switched the first pair, the test statistic would be 0. If we continued to repeatedly reassign treatment differently, calculating the test statistic each time, we would build up a picture of what kinds of mean differences are more or less common in the world of the null hypothesis. This picture is the distribution of the test statistic under the hypothesis of no effects. This "no effect"-distribution (more commonly called the "null-" or "randomization-" distribution) for the mean-difference in turnout between treated and control cities ranges from -5 to 5. About 0.38 of these differences are greater than or equal to 1.5. Convention in the social sciences would be to call the observed number surprising from the perspective of the hypothesis if, say, only 1 in 20 replications of the "no effects"-thought experiment produced values as large as or larger than 1.5. Fisher in 1935 would say that we can never know whether a treatment had an effect but that, in this study, we can't easily rule out the idea that treatment had no effect — the observed mean difference is just not that surprising from the perspective of the null hypothesis.[1]

The nature of Fisher's statistical inference is that simple. Following Fisher, our investigation centered around converting a vague but substantively interesting question about effects into a precise hypothesis about no effects. The computational thought-experiment produced a quantitative description of the hypothesis in action. Our observed value is not very surprising from the perspective of the hypothesis of no effects.

Randomization-based tests fulfill their promises. The quote from Fisher at the start of this paper made a claim that randomization will "suffice to guarantee the validity of the test of significance". We have so far shown that randomization can "suffice", but what about "validity"? As Fisher meant it, it appears to have encompassed both the design idea that comparisons were unconfounded by extraneous factors but also the statistical idea that the hypothesis test fulfilled its promises. We already know well how randomization helps us make unconfounded comparisons. What does it

[1] All computations reported in this paper (like the p-value of 0.38 reported above) are available for reproduction and exploration by interested readers from `http://` This paper is written in the mixture of R and $\LaTeX$ known as Sweave (Leisch 2002, 2005). Those interested in learning more may download the source code of the paper and apply themselves to adapting it for their own purposes.

mean for a statistical test to fulfill a promise?[2] One promise that tests make is to rarely reject a correct null hypothesis. Imagine constructing a treatment assignment that has no relationship with outcomes. In this situation, a test of the hypothesis of no effects ought to make the test statistic appear unsurprising if our test is working well. Of course, just by chance we might find a few cases in which a test statistic appearing from a simulation study like this would appear surprising and we would reject the correct null incorrectly. Imagine setting a rejection threshold with the *p*-value of .05: One branch of statistical theory suggests that a well-operating test ought to have a false positive rate no greater than the *p*-value rejection threshold: so we should see no more than 5% of true nulls rejected by a test which is "valid" on the criteria of false positive rate. That is, across repeated realizations of the same design with no effect, no more than 5% of realizations should have *p*-values less than .05. In fact, at any level of "surprisingness" or test rejection threshold, α, a valid hypothesis test should reject the null or encourage us to categorize the test statistic as overly surprising no more than $100\alpha\%$ of the time when the hypothesis under consideration is correct. Fisher's test guarantees this kind of behavior (Rosenbaum 2002*c*, Chap 2).

For example, a paired *t*-test applied to our 8 city data and design, but using a simulation in which no systematic relationship is allowed between treatment and turnout, would encourage us to reject the true null of no effects 6% of the time when α is between .01 and .1 and 12.5% of the time when α is between .11 and .15. Sometimes the *t*-test applied to these data would be too quick to suggest we claim that the observed relationship is surprising from the perspective of the null of no effects (in standard parlance, to claim the relationship is "Significant") and sometimes the *t*-test would be too conservative. The Fisher's approach, in contrast, never rejects the null of no effects more than the α value. It is always correct or conservative — never leading us to claim "significance" precipitously.[3]

2.2 The Promise of Fisher's Method as of 1935

Fisher's insight in 1935 allows the precise assessment of a substantively meaningful hypothesis with a probability statement justified by the research design. Fisher's approach is promising at first

[2] I am borrowing the idea that a level α statistical test makes a promise to reject a true hypothesis no more than 100α% of the time from Rosenbaum (2009, p. 365).

[3] See Appendix A for some simple R code assessing false rejection rates of the *t*-test and Fisher's randomization-based test (both using a *t*-statistic).

glance because of what it does not require. The validity and meaning of the p-value reported above did not require knowledge of a sampling plan or a population from which sampling occurred. In fact, these cities are not a random sample of a well delineated population of cities. Nor did this p-value require us to turn our knowledge of the stochastic process producing aggregate turnout in US cities into a likelihood function and associated parameterization. Nor did we make asymptotic arguments. The eight cities support a valid statistical inference here with no apologies necessary for a small sample.

Small samples are only one kind of design for which Fisher's method appears promising. In general, Fisher's method recommends itself in low information circumstances or when assumptions like likelihood functions for outcomes and functional forms are hard to justify and/or distract from the social scientific task at hand.

The clearest case in which information is low is when we have a small sample, like the 8 city study we analyze in this paper. However, a Normal or t-distribution can fail to well approximate the distribution of our test statistics under the null even when samples are not small. The three most common examples of such circumstances in the social sciences involve binary outcomes, clustered or aggregate units, and weak instruments. For example, in a study of $n = 77$ independent observations (with 44 successes) and $p = 7$ covariates, Brazzale, Davison, and Reid (2006, §4.2) summarize their simulation study by stating that "the fit of the logistic regression model to the 77 binary responses appears to give about the same information as the fit of the same $p = 7$ covariates to $n = 10$ continuous responses from the logistic distribution." (p. 46) (see also Davison (2003, ex. 10.17)). Harrell (2001, § 4.4) suggests a rule of thumb of $\min(n_1, n_2)/\#\text{ covariates} = n_{\text{continuous}}$ for quickly assessing the equivalence in information provided by a binary outcome compared to a continuous outcome (n_1, n_2 are the numbers of responses in the two categories of the outcome). Applying Harrell's rule to the Brazzale et al example would suggest that the $n = 77$ and $p = 7$ study is roughly equivalent a study with $n = \min(44, 33)/7 \approx 6$ continuous outcomes.[4] Of course, as Harrell (2001, § 4.4–4.5) notes, these results assume that the covariates collectively display ample variation and

[4] Peduzzi et al. (1996) used a dataset of 657 patients and 252 events to conduct a simulation study which showed that at least 10 observations per covariate were necessary to avoid bias in the coefficient estimates and ensure proper false positive rates of asymptotically justified tests (see also Whitehead (1993)).

thus are not overly multicollinear.

Another common research design in political science is the multilevel or clustered design. Yet, such designs often involve small numbers of units at the higher levels. For example, a cross-country study of political behavior may involve many thousands of individuals per country, but only 20 countries. In a vivid example of this problem in economics, Bertrand, Duflo, and Mullainathan (2004) collected data on 92 papers published in economic journals which used difference-in-difference estimation and created a simulation study based on real data (the wages of 900,000 observations of women in the Current Population Survey from 1979 to 1999 in 50 states and 21 years). Their simulation study involved randomly generating placebo laws for half of the states during random years 1985–1995 and testing the hypothesis of no effects of the fake laws for each run of the simulation. The simple *t*-test in OLS rejected the true null 67.5% of the time! When their simulations included 50 states, Bertrand, Duflo, and Mullainathan (2004) were able to diminish their false positive rate by using the block bootstrap (or "robust clustered standard errors") but when the number of states was 20, the false rejection rates again shot up, even in the presence of such a correction (see also Erikson, Pinto, and Rader (2010) for the same kind of results focusing on state-level policy interventions and their effects on individual-level behavior). This fact, that the asymptotic properties of multilevel models or "robust cluster standard errors" for clustering and multilevel designs depend on the numbers of clusters, is not new in political science, but fixes for it have been rare (Stoker and Bowers 2002*b*,*a*; Green and Vavreck 2007). For example, Bowers and Drake (2005) advocate eschewing statistical inference at all in such circumstances. Randomization inference, however, allows analysts to avoid this problem (Erikson, Pinto, and Rader 2010; Hansen and Bowers 2009, 2008; Small, Ten Have, and Rosenbaum 2008; Raab and Butcher 2005).

A final source of low information arises from weak instruments. We have long known that two-stage least squares (2SLS) estimation using instrumental variables is inconsistent in the face of weak instruments (Bound, Jaeger, and Baker 1995; Staiger and Stock 1997). For example, using a sample of 500,000 Americans, Imbens and Rosenbaum (2005*b*) replicated the famous Angrist and Krueger (1991) analysis of schooling and earnings but replaced the instrument (quarter of birth) with random noise. The confidence intervals for the effect of schooling on earnings should have been wide (indicating that the now noise-only instrument carried no information), but 2SLS reported narrow intervals; a manifestation of the kinds of overly liberal false positive rates that we have discussed in the context of clustered data or binary outcomes.

Because low-information designs are not uncommon in political science, a method which guarantees a given false positive rate regardless of information-content ought to be attractive to political scientists. Further, even if instruments are strong and samples are large and independent, the fact that Fisher's hypothesis testing framework does not require justification of choices about likelihood functions or functional forms may make it useful in situations where too much ink is being spilled over the statistical methods, detracting from discussion of the social science.

2.3 The Unmet Potential of Fisher's General Method

Even if current findings about binary outcomes, clustered designs, and weak instruments were not available in 1935, the promise of an approach like Fisher's was still clear at the time. However, a series of obstacles to widespread adoption have still remained. Testing the strict null of no effects only addresses hypotheses about whether any effect is manifest, but social scientists since Fisher have often preferred confidence intervals as summaries of sets of hypotheses over single hypothesis tests. Second, in order to elucidate the assumptions required to call an empirical comparison "causal," contemporary experimentalists have found it useful to conceptualize "a treatment effect" in formal, counterfactual, causal terms. If we do not know how hypotheses about "no effects" relate to potential outcomes, then one might worry about the meaning and validity of the causal inferences even if the validity of the statistical inference is clear.[5] Third, while the instructions to "repeat the experiment" are easy to state, they implied unsurmountable computational burdens during the decades following Fisher's initial idea, and may still imply such burdens today. Fourth, as developed by Fisher, this approach does not allow us to use subject matter knowledge to improve the precision in the estimation of treatment effects: experimentalists often use regression models for such "covariance adjustment". Furthermore, the exchange between Freedman (2008*b*,*a*, 2007) and Green (2009) and Schochet (2009) would make some experimenters worried about the use of regression

[5] For useful overviews of this formal framework for conceptualizing causal inference aimed at a political science audience see Sekhon (2008) and Brady (2008). The idea of potential outcomes is often traced to Neyman (1923 [1990]) and most prominently elaborated and developed by Rubin (1974, 2005)

in their particular designs, but not clear about what alternatives exist. So, a methodology focusing on hypothesis testing appears not to address the deep concerns of practicing political scientists.

This paper engages with each of these obstacles to the use of Fisher's method. Developments since Fisher allow us to broaden the scope of his simple idea — allowing us to understand his hypothesis test in terms of potential outcomes, and enabling the the creation of confidence intervals and the use of linear models to enhance precision of statistical inferences about substantively interesting causal comparisons. Some political scientists have begun to make use of Fisher's approach, although some of the more prominent examples of these applications have occurred in statistics journals. For example, Hansen and Bowers (2009) use Fisher's method to produce valid confidence intervals for treatment effects in a clustered multilevel randomized field trial with binary outcomes and an instrumental variable — in that study, Fisher's method did the work of what we might more commonly consider a two-stage, instrumental variables, multilevel logit model. Ho and Imai (2006) applied randomization inference to the question of ballot order effects. And other political science work shows how these basic ideas can be extended: to anova in experimental studies (Keele, McConnaughy, and White 2008); to tests of policy effects in difference-in-difference designs (Erikson, Pinto, and Rader 2010), and to instrumental variables (Imbens and Rosenbaum 2005*a*), to clustered assignment within pairs (Imai, King, and Nall 2008).

This paper contributes by providing a foundational exposition of Fisher's method to enable political scientists to extend and apply these ideas to their own studies and to understand and make use of the advanced applications that are beginning to appear.

3 Hypotheses about Effects: Confidence Intervals and Potential Outcomes

Imagine that past studies suggested that most elections in two-party systems were won or lost by small margins of victory (say, 2 percentage points). Would a 1.5 point mean difference be surprising from the perspective of the past literature's hypothesis of a 2 point effect? As shall be shown, testing hypotheses about effects allows the creation of confidence intervals. However, it will be easier to explain how Fisher's approach may be generalized beyond the testing of sharp nulls of no effects if we first understand how to describe such hypotheses using the notation of potential outcomes and formalize the testing of the "no effects" null.

3.1 Fisher's null of no effects reconceptualized and formalized.

Although he did not know it, Fisher's null hypothesis of no effects can be written as a statement about a causal quantity formally defined in the potential outcomes framework. Here, we write $Z_i = 0$ to indicate assignment to the control condition and potential response to control for city, i, as $r_{Z=0,i}$, or, more simply, r_{0i}.[6] Treatment has no causal effect on turnout in city, i, if $r_{0i} = r_{1i}$ (where r_{1i} is potential turnout after treatment with advertisements). Of course, the "fundamental problem of causal inference" Holland (1986) is that only one of the potential responses is realized and the other is missing. Fisher's hypothesis about no effects may be written as a comparison of potential outcomes and thus as a statement about a causal effect:

$$H_0 : r_{1i} = r_{0i}. \tag{1}$$

If the hypothesis were true, switching city i from treatment to control should not change the turnout we observe. Observed outcomes, R_i, are linked to potential outcomes and treatment assignment by the following identity:

$$R_i = Z_i r_{1i} + (1 - Z_i) r_{0i}. \tag{2}$$

Observed outcomes are potential outcomes realized. If, for the sake of argument, we granted the the null of no effects, then observed outcomes are potential outcomes absent treatment: $R_i = Z_i r_{1i} + (1 - Z_i) r_{0i} = Z_i r_{0i} + r_{0i} - Z_i r_{0i} = r_{0i}$. That is, *the null hypothesis and the identity linking observed to potential outcomes allows us to represent a hypothesis about a causal effect in terms of what we observe*. This is a big deal because it is one way to deal with the fundamental problem of causal inference. Entertaining the sharp null of no effects means that $R_i = r_{0i}$.

What is more, our null hypothesis also implies something about a function, $t(Z_i, R_i)$, producing a test statistic summarizing an observed relationship between Z_i and R_i.[7] Recall that in § 2.1 we calculated a mean difference for different replications of the experimental procedure to represent

[6] Upper-case letters represent random quantities (like treatment assignment which could have been otherwise) and lower-case letters represent fixed quantities (like potential outcomes or covariates).

[7] The design we use for an example here is paired, and so, for accuracy we should write Z_{ib} or r_{0ib} for city i in pair b. We suppress the b subscripts for now in the interest of clarity.

the distribution of the mean difference under the null. This hypothesis test was an assessment of a causal quantity because, under the null we were producing the distribution of $t(Z_i, r_{0i})$. Granting the null, we calculated $p(t(Z_i, r_{0i}) \geq 1.5) = 0.38$.[8]

The simple hypothesis about no effects has allowed us to link Fisher's developments with a modern perspective on causality. For example we might write Fisher's hypothesis about a causal effect, τ_i, where $\tau_i = r_{1i} - r_{0i}$, as $H_0 : \tau_{0i} = 0$. Fisher's hypothesis test is an assessment of causal effects. Nothing that we have said depends on means and their large-sample properties but allows the use of nearly any substantively meaningful or statistically powerful test statistic. Now, let us move beyond questions about "no effects" to questions about "some effects".

3.2 Confidence Intervals: Assessing Hypotheses About Effects

Recall that a $100(1 - \alpha)\%$ confidence interval is defined as the set of hypotheses not-rejected at some level α, where α quantifies the risk of falsely rejecting a true hypothesis.[9] For example, we have so far assessed $H_0 : \tau_{0i} = 0$, which follows from the strict null $r_{1i} = r_{0i} + 0$, and have p-values of 0.38 and 0.44 for the mean and rank-based test statistics respectively. So, we cannot exclude $\tau_i = 0$ from a 95% confidence interval defined as excluding those values of τ rejected at $p < .05$.

We were able to test a hypothesis of no effects because that hypothesis was specific enough to tell us how each unit would have acted if it were true. In fact, we can test any hypothesis which is this specific. What if elections tended to be won by two percentage points of turnout? One way to represent this informal hypothesis would be attribute 2 points of turnout to treatment for each unit. If the null of no effects implies $H_0 : r_{1i} = r_{0i}$, a hypothesis about 2 points of turnout might state $H_0 : r_{1i} = r_{0i} + 2$. More generally, one might imagine a general function that produces potential responses to treatment from potential responses to control. For example, a simple such function

[8] Nothing about this procedure requires a comparison of means. We could produce the same result with a standardized mean difference like the t-statistic. Or, if we worried about outliers (Lehmann 1998; Keele, McConnaughy, and White 2008), we could use a sum of ranks among the treated. For example, if $\mathbf{q} = \text{rank}(\mathbf{R})$, where bold-faced characters are vectors or matrices, 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, like means, are also functions of potential outcomes. A one-sided p-value for the test based on a paired signed rank-sum test statistic is p=0.44. If we had binary outcomes we might prefer simple sums or totals (Rosenbaum 2002*a*). Different test statistics might have different statistical power or might otherwise summarize results in more or less substantively meaningful ways.

[9] That is, given a choice of acceptable false positive error rate, one can create a confidence interval out of hypothesis tests (see, for example Rice 2007, §9.3).

states that $h(r_{0i}) = r_{0i} + \tau = r_{1i}$ which implies that $r_{0i} = r_{1i} - \tau$ or $\tau = r_{1i} - r_{0i}$. This "hypothesis generating function", or "model of effects" (Rosenbaum 2002*c*), suggests that the potential outcomes under treatment are merely the potential outcomes under control plus some constant τ which is the same for all cities. If $H_0 : \tau = \tau_0, \; \tau = r_{1i} - r_{0i}$, where τ_0 is a hypothesized value for τ, then the identity linking observed to potential outcomes implies something specific about our observed outcomes for cities i:

$$\begin{aligned} R_i &= Z_i r_{1i} + (1 - Z_i) r_{0i} && \text{the identity} \\ &= Z_i r_{0i} + Z_i \tau_0 + r_{0i} - Z_i r_{0i} && \text{substitute } H_0 \text{ implication for } r_{1i} \\ &= r_{0i} + Z_i \tau_0 \\ r_{0i} &= R_i - Z_i \tau_0. && (3) \end{aligned}$$

If our hypothesis were true, we could recover potential outcomes to control by removing the hypothesized effect from the observed outcomes of treated units: if $h(r_{0i}) = r_{0i} + \tau$ then $H_0 : \tau = \tau_0 \;\Rightarrow\; R_i = r_{0i} + Z_i \tau_0$. When we want to entertain the hypothesis that the treatment had no effects, $\tau_0 = 0$ and so $H_0 : \tau = 0 \;\Rightarrow\; R_i = r_{0i}$ — just as we saw earlier. For a given hypothesized value, τ_0, we can repeat the assignment process of the experiment by generating new $n \times 1$ vectors $\mathbf{z}$ that are consistent with the design of the study and calculate $t(\mathbf{z}, \mathbf{R} - \mathbf{z}\boldsymbol{\tau}_0)$ to represent the test statistic implied by the null hypothesis.[10]

Most of the writing on randomization inference talks about repeating the assignment process using a different language which helps formalize the process. So, here, briefly,we link our interpretation to what readers might find by reading works in statistics. That work follows Fisher (1935) and especially Rosenbaum (2010, 2002*c*) and uses a thought-experiment of "all of the possible ways for the assignment process to occur" rather than "many repetitions of the assignment process". Imagine a set Ω that contains all of the possible vectors describing treatment $\mathbf{z}$ (in this case, all of the $\mathbf{z}$ look like $\{1, 0, 1, 0, 1, 0, 1, 0\}$ — eight entries in pairs with one unit in each pair indicating treated (1)

[10] We use boldface font to represent vectors and matrices. Here, $\mathbf{R}$ is $n \times 1$ vector of observed responses, R_i, and $\boldsymbol{\tau}_0$ is also $n \times 1$ and simply contains τ_0 for all entries.

and the other indicating not treated (0)). For the simple strict null of no effect, 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:

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

where Ω is the set of all possible treatment assignments, and K is the total number of possible assignments in Ω. To generate the randomization distribution representing a hypothesis about constant, additive effects, substitute $(\mathbf{R} - \mathbf{Z}\boldsymbol{\tau}_0)$ for $\mathbf{R}$ in equation 4. If the p-value is greater than or equal to some α value, τ_0 is inside the 100α confidence interval, otherwise it is excluded from the confidence interval.

At the beginning of § 3 we raised the possibility that the treatment effect was 2 percentage points of turnout. The logic of equation 3 implies that, if our null hypothesis were true, potential outcomes to control among the treated would be potential outcomes to treatment minus 2: $r_{0i} = R_i - Z_i 2$. We can evaluate the evidence against this hypothesis using the same procedure as above: for each repetition of the experiment calculate the test statistic, now using the adjusted outcomes, $t(\mathbf{z}, \mathbf{R} - \mathbf{z}\boldsymbol{\tau}_0)$.

Repeating the assignment process entertaining this hypothesis tells us that our observed data would not be very surprising from the perspective of $H_0 : \tau_0 = 2$: we find p =0.5 using the difference of means test statistic against alternatives in which turnout after treatment is greater than control.

If we repeat this procedure for a range of τ_0, we produce a confidence interval. If we set $\alpha = .12$, for example, we would reject any $\tau_0 > 5$, leading to a one-sided confidence interval of $\tau \leq 5$. That is, we can reject hypotheses where the treatment effect is more than 5 percentage points of turnout if we used one-tailed tests as we have so far ($p = 0.188$ for $\tau_0 = 5$ but $p = 0.125$ for $\tau_0 = 5.1$).[11]

An 88% two-sided confidence interval contains those τ_0 not rejected at $\alpha = .12$ using a two-tailed test. In this case the two-sided interval is $\tau \in$ [-7.00, 6.00]. Within this interval the two-sided p-values were all greater than or equal to .25 and outside the interval the p-values were smaller than

[11] The mean difference and paired signed rank test statistics produced identical intervals in this case although the particular p-values inside the intervals differed.

or equal to .125.[12]

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. In this case, our 88% CI actually could encompass an 80% CI or even a 75 % CI 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* $1 - \alpha$. In this way, confidence intervals created using randomization inference are guaranteed to have correct coverage, and will be conservative if their significance level ($\alpha = .12$) is not exactly the same as their size.[13] Rosenbaum (2002*c*, Chapter 2) also proves that these intervals are unbiased and consistent (such that more information leads to smaller intervals) but that the correct coverage of the intervals does not depend on the sample size or correctness of some model of outcomes.

We could produce a point estimate by shrinking this confidence interval. For example, a 66% interval narrows the range of not-rejected hypotheses considerably: [-2.00, 5.00]. The difference of means test statistic provides a point estimate of 1.5. In this simple case, the point estimates are the hypotheses least surprising from the perspective of the observed data: the hypotheses with the largest p-values.[14]

3.3 Beyond the Constant Effects Hypothesis

Stating a hypothesis is not the same as assuming a model. The question to be asked of a hypothesis is not whether it is correct but whether it is substantively useful or interesting. Are the hypothetical outcomes under treatment generated by the constant, additive effects function, $h(r_{0i}) = r_{0i} + \tau$, scientifically useful or interesting counterfactuals? In this case, one might imagine so — the same

[12] We cannot use $\alpha = .05$ (for a 95% CI) here because the atom of the probability distribution in this data is 1/16=.0625: in the case of independent assignment across pairs or blocks, b, $K = \prod_b \binom{n_b}{\sum_i^{n_b} Z_i}$. In our case $K = \prod_{s=1}^{4} 2 = (2)^4 = 16$. There are only 16 ways to assign a binary treatment within 4 pairs and 8 units.

[13] The size of a test is the probability that it will reject a true null hypothesis. The level of a test is the *maximum* probability that it will reject a true null hypothesis (Lehmann and Romano 2006, § 3). (Rosenbaum 2010, page 365) summarizes this distinction nicely: "In typical use, the level of the test is a promise about the test's performance and the size is a fact about its performance, where the achieved fact may be better than the promised performance." Here, the promise is that $\alpha = .12$ but the achieved fact is $\alpha = .25$. The size and level of tests tend to only differ when information is low (ex. sample sizes are small).

[14] Actually, in this case they are the medians of such sets of hypotheses. This is an informal way to talk about what are known as Hodges-Lehmann point estimates. More formally, these point-estimates are defined as the value of the hypothesis that makes a test statistic equal to its expectation when the null of no effects is maintained. See the theory in (Hodges and Lehmann 1963) and elaborated in (Rosenbaum 1993).

kind of media campaign within U.S. cities chosen to have similar institutional and electoral features in the same year might reasonably have the same effect in each city. Yet, these hypotheses ought not exhaust our imagination. Any scientifically interesting hypothesis may be formulated in this way and assessed using the machinery described above. Rejection of a null hypothesis means that the data are surprising from the perspective of the hypothesis. The data could be surprising either because our hypothesis generating function is not well supported by the data (say, if $r_{1i} = \tau_i + r_{0i}$ rather than $r_{1i} = \tau + r_{0i}$) or because our hypothesis generator is plausible but the particular value of τ is not plausible. In either case, the confidence interval produced by rejecting hypotheses will have the correct coverage even if it might be wider than desired.[15]

To demonstrate that Fisher's approach is not limited to assessing hypotheses about constant effects, we here build a hypothesis generator that is a bit more tailored to our study and more complex. We observe baseline turnout for each city, so a difference-in-differences type of design. For example, we can calculate a treated vs. control difference of 22-16=6 in the pair where advertisements ran in Sioux City but not in Saginaw. Yet, in the mayoral election prior to treatment, the difference in turnout between Sioux City and Saginaw was 21-17=5. One might imagine, based on this comparison of baseline outcomes, some of the 6 point difference observed post-treatment might be attributable to the 5 point baseline difference. This reasoning suggests suggest a generator like: $h(r_{0i}) = r_{1i} = r_{0i} + \tau - (x_i - x_j)$, where x is baseline turnout and i and j are the paired units. When we solve for τ, $\tau = (r_{1i} - r_{0i}) - (x_i - x_j)$, we see that the post-treatment difference is adjusted by the pre-treatment difference. However, we also wonder whether the actual effect of the treatment depends on the level of baseline turnout. At the extreme, we imagine different ceiling and floor effects: changing turnout in a place where no one turns out (or everyone turns out) might require much more effort than changing turnout in places where some but not everyone turns out. For now, and for simplicity, we presume a linear relationship — as the level of baseline turnout increases, we

[15] For example, Gadbury (2001) and Robins (2002, § 2.1) both show that the coverage remains correct when the true hypothesis generator is $r_{1i} = \tau_i + r_{0i}$ but $r_{1i} = \tau + r_{0i}$ is tested. Rosenbaum (2002*d*, § 3–6) explains the equivalences between estimating an average treatment effect and testing a sequence of hypotheses about individual causal effects. Other examples of hypothesis other than the constant, additive effects hypothesis include Hansen and Bowers (2009), who assess additive hypotheses about binary outcomes which vary by individual ($r_{0i} = r_{1i} - \tau_i, \tau \in 0, 1, r_{1i} \geq r_{0i}$), and Rosenbaum (2002*e*) who inspects hypotheses about $r_{1i} > \theta > r_{0i}$ where θ is some value of the order statistics of r_{0i}. See also § 2.4.5 in Rosenbaum (2010) for an assessment of "Tobit" style effects in which $r_{0i} = \max(r_{Ti} - \tau_0, 0)$.

imagine that the treatment effect itself will increase. These two considerations lead to the following generator:

$$h(r_{0i}) = r_{1i} = r_{0i} + \beta\tau x_i - (x_i - x_j) \tag{5}$$

Equation 5 is a difference-in-differences-style of hypothesis but adding dependence on the level of x_i — on pre-treatment turnout for the unit about which we are hypothesizing. The $x_i - x_j$ term is the difference between the paired units in baseline turnout. The causal effect of interest, τ, is still a constant, but any given city receives a proportion of the treatment effect $\beta\tau$ depending on the baseline turnout of the city, x_i, and this effect is adjusted by the baseline or placebo difference between the two cities in the pair.

Just as (3) showed how the hypothesis generator $h(r_{0i}) = r_{1i} = r_{0i} + \tau$ implies that $r_{0i} = R_i - Z_i\tau$, equation 5 and (2) imply that when we entertain this hypothesis we can recover potential responses to control by adjusting our observed outcomes such that $r_{0ib} = R_{ib} - Z_i\left(x_{ib} - \beta\tau(x_i - x_j)\right)$.

For simplicity, we set $\beta = .2$ and tested hypotheses about τ: the largest treatment effect not rejected at $\alpha = .12$ was 1.2: our one-sided 88% confidence interval is $(\infty, 1.2]$ We could have also tested hypotheses about *both* β and τ here — showing either a 2-dimensional confidence region or some version of an interval about τ conditioning on tests of β following the example of Nolen and Hudgens (2011). We do not do such a test here because the point of this more complex hypothesis was not to claim that this hypothesis is the most scientifically interesting way to think about how treatment turns into outcomes in this study, but to dispel myths about the requirement of constant effects which have hindered the adoption of these methods and to show another way that social scientific theory or research design and substantive knowledge may assist and enhance statistical inference in Fisher's approach.

4 Using what we know: Model-assisted, Randomization-justified Inference

Political scientists know a lot about turnout. And we have shown how we can assess hypotheses in which we directly represent some of our knowledge about turnout and potential outcomes and our design. Yet, a more common approach among political scientists is to use linear models including well-known predictors of outcomes to increase the precision with which treatment effects are estimated. Since we are most comfortable with linear models, and linear models are very useful and flexible tools, can Fisher's approach be compatible with such models? In addition, we have long known that one can shrink the sizes of confidence intervals using such information to "adjust" our test statistics for covariates (often known as "covariance adjustment").[16] Yet, Fisher's original approach did not make it clear how one might use the additional information in randomization-based hypothesis testing. And, the extensive literature on adjustment of randomized experiments is contentious because of fears that exploration of adjustment strategies will led to misleading characterizations of treatment effects: estimating many different linear regression models in the process of adjustment may yield a treatment effect that appears particularly large, but which, in fact, is merely an artifact of the kinds of linear interpolation and extrapolation of a given specification. More over, statistical inference for treatment effects after such data snooping is also suspect: one hundred tests with pre-specified rejection level $\alpha = .05$ will falsely reject the true null of no effect five times. These problems of data snooping are well known. Current best practice in the world of clinical trials involves a public declaration in advance of the experiment of an adjustment strategy using one of the online trial registries and also to report unadjusted results.[17] The Experiments in Governance and Politics group is currently working to translate such procedures to the social sciences.[18]

At first glance it appears that any attempt to adjust experimental data without advance public registration will run afoul of such criticisms. Yet, we will briefly show here that covariance adjusted randomization inference offers a particular benefit in this regard, allowing a search over adjustment specification without risking the problems of data snooping. First, however, let us briefly explain how one may "adjust" treatment effects in the spirit of Fisher's randomization inference.

Recall that the procedure for randomization inference depends on focusing attention on a specific substantively interesting set of hypotheses: claims about counterfactual comparisons. For example, $H_0 : \tau = \tau_0$ and $\boldsymbol{\tau} = \mathbf{r}_1 - \mathbf{r}_0 = \tau\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, implies a particular $\mathbf{r}_0$ that we can observe by adjusting observed responses such that $\mathbf{r}_0 = \mathbf{R} - \tau_0\mathbf{Z}$. The design of the experiment

[16] See Cox and McCullagh (1982); Bowers (2011); Keele, McConnaughy, and White (2010) for only a few of the overviews of such adjustment.

[17] http://www.consort-statement.org/

[18] http://e-gap.org/resources/standards-project-registration/

allows us to instruct our computers to repeat it while calculating $t(\mathbf{z}, \mathbf{r}_0; \mathbf{z} \in \Omega)$ for each hypothetical repetition. The collection of such test statistics is the null randomization distribution against which we compare our observed $t(\mathbf{Z}, \mathbf{R})$ in an effort to discredit H_0.

The width of 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 distinguish control from treated observations. Imagine that we could regress $\mathbf{r}_0$ on some set of covariates in the $n \times p$ 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 $\mathbf{x}$. Such a regression would not involve looking at effects of treatment, and thus, protects our inferences from concerns about data mining. But such a regression is impossible since we do not observe $\mathbf{r}_0$ for cities where $Z_i = 1$. Fisher, however, shows that the process of hypothesis testing allows us to replace $\mathbf{r}_0$ with some function of what we do observe. Rosenbaum (2002*b*, § 2.4) 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 hypothesis generating function, one can test hypotheses $H_0 : \tau = \tau_0$ using $t(\mathbf{Z}, \mathbf{e})$. To summarize, a regression model can aid the production of randomization justified confidence intervals via the following steps:

Define a function to produce residuals (outcomes purged of covariate-based noise). $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x}) = \mathbf{e}$. The residuals, $\mathbf{e}$ will be calculated from fixed quantities $\mathbf{r}_0$ and $\mathbf{x}$ and so will be fixed just as $\mathbf{r}_0$ itself is fixed under the null.

Compute adjusted outcomes based on some hypothesis. For example, hypotheses of the form $H_0 : \tau = \tau_0, \mathbf{r}_1 = h(\mathbf{r}_0) = \mathbf{r}_0 + \tau$ imply that 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 $\mathbf{R}$ and τ is a scalar.

Compute $t(\mathbf{Z}, \mathbf{e}_0)$ and compare to $t(\mathbf{z}, \mathbf{e}_0; \mathbf{z} \in \Omega)$ for a p-value. Recall that Ω can be thought of as the collection of possible treatment assignment configurations (accessed by repeating the assignment process using our computer).

Any attempt to use covariates must engage with a number of questions: Which covariates ought to be included? Which function of covariates ought to be fit? Which fitting procedure (least squares? least absolute deviations? outlier resistent least squares?)? Luckily, the procedure outlined above allows analysts to use the data to answer such questions as long as treatment effects are not calculated until after the exploration is complete: covariance adjustment in the context of randomization inference separates adjustment from assessment of treatment effects.

Here, very briefly, we describe how we applied these general ideas to eight city study. Our sample is small and so concerns about overfitting loomed somewhat large for us — a model fit to 8 observations taking pairing into account would lead us to an $R^2 \approx 1$ merely by including 3 covariates. To heavily penalize overfit solutions we used two strategies. First, we departed from Rosenbaum's article, and restricted our model fitting and selection (of variable and of tuning parameters) to the control group only. And second, we did model selection and penalization using median regression (which is influential point resistent) (Koenker 2005) with a lasso penalty (Tibshirani 1996) for each parameter (which is very overfitting resistent). We chose tuning parameters following the adaptive lasso procedure of Zou (2006) but as applied to quantile regression (fitting only the median) (Koenker 2005, § 4.9.2) and extrapolated the final fit to the treatment group in order to produce residuals.[19] We made this choice in large part because the process of inverting the hypothesis tests in such a small sample was overly sensitive to influential points.

We also chose this procedure in order to show how easily the basic ideas from Fisher can relate to the most modern approaches to model-fitting in statistics and machine learning. Moreover, not only is the adaptive lasso new (and its link to median regression even newer), but statistical inference for these estimators is not well worked out. For example, Chatterjee and Lahiri (2011) propose a residual bootstrap approach for inference about the coefficients in lasso models reacting to disappointing theoretical findings about the large sample properties of lasso estimators.[20] Thus, we can highlight another benefit of Fisher's basic ideas: we can fit models, using the most modern model selection and variable selection technology (so new that statistical inference is on somewhat shaky ground

[19] More details on how we chose a covariance adjustment specification are available in the source code of this document available online at `http://`.

[20] The software we used, `quantreg` (Koenker 2011), refuses to produce standard errors let alone p-values for coefficients estimated by it.

even if the model selection part is firm) but our statistical inference occurs after such fitting is over and does not depend on any of the same assumptions.

Using the adaptive lasso-penalized median regression procedure outlined above, the same simple hypothesis generating function of $h(r_{0i}) = r_{0i} + \tau = r_{1i}$ $H_0 : \tau = \tau_0$ we find that we can reject hypotheses greater than $\tau_0 = 2.1$ at $\alpha = .12$ using a one-tailed test — about half the length of the unadjusted interval. A two-sided confidence interval is bounded by -3.2 and 5.0 points of turnout difference after this covariance adjustment.

This change in the confidence interval does not change its operating characteristics: Fisher's method rejects a true null at $\alpha = (1/8)$ exactly 1/8 of the time with or without adjustment. And, we could have assessed the more complex hypotheses generated by equation 5 on these less noisy outcomes as well.

4.1 Regression without Regrets?

Models relating covariates to outcomes are meant only to reduce noise in the outcomes in this method. We emphasized $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x})$ above — a noise-reduction or residual-producing function — rather than $\hat{\mathbf{R}} = \mathbf{x}\hat{\boldsymbol{\beta}}$ because we never need to examine the coefficients relating covariates to outcomes 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 $\mathbf{Z}$. This is not to say that one must not be thoughtful in choosing such a model. Incorrect specifications will add noise and will thus make the confidence interval wider but will not change the coverage of the confidence interval. Finally, notice that this is a method for hypothesis testing, not estimation. We are not using the residuals to produce point estimates.

Recall that regression in this approach does not estimate a causal effect, rather it removes noise from outcomes to enable more precise tests of hypotheses about causal effects. In this way, attention turns to using substantive knowledge to specify hypotheses of interest among the many possible to test, and also to produce useful regression specifications (useful, in that they soak up non-treatment related noise in outcomes based on knowledge about the substance).

5 Summary: Fisher's promise is now feasible

Statistical inference, as inference, enables guessing about unobserved quantities using the principles of probability. We have shown that the unknowns in Fisher's approach can be understood as the potential responses to treatment among the control group (or to control among the treatment group). And we have shown how this approach can fruitfully answer questions of interest to political scientists using analytic objects like confidence intervals or linear models which do not immediately suggest themselves as useful when one considers the Fisher's 1935 test of the sharp null of no effects.

We have shown that statistical inference does not require a model of outcomes or specific hypotheses about non-zero 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 (i.e. believes reports about how randomization occurred — not about how treatment was actually administered, but about how the random numbers themselves were generated). Inference about effects (rather than about the plausibility of no effects) requires adding substantive knowledge about the process relating treatment to potential outcomes. Statistical 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 an outcome-data-generating-process model (like a common likelihood function) or on asymptotics or on correctness of some function linking covariates and treatment to each other and to outcomes, although we did add more and more structure to the inference. What is nice about this, we think, is that statistical inference in experiments is reliable in a special way. Even if one prefers a t-test or common regression model for statistical inference in experiments, randomization inference offers a check on such approximations. Those who know the history of statistics will not find these facts 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, who did know about Fisher's work, will be, we hope, pleasantly surprised.

6 Discussion: Fisher and Neyman.

In a sense, randomization inference allows scientific attention to focus on the political phenomena of interest: what causes what? what counterfactuals ought we to entertain? what units received the treatment, in what way? That more information is better within this framework is a good thing. Randomized experiments make models of assignment more credible (in general) than observational studies. And randomized experiments are a natural place to apply these methods.[21] 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 principle although one may pragmatically approximate randomization-based results using other methods (while knowing that such approximations are assessable using the randomization-based methods).

There are two other goals interrelated goals in the analysis and design of experiments that are not directly addressed in this paper but which are focuses of other modes of statistical inference: testing weak null hypotheses and extrapolation or forecasting. The fact that one mode of inference cannot do it all is not a reason to ignore it, yet it is reasonable to expect some comparison with extant methods in an exposition such as this.

Before Fisher (1935), Neyman (1923 [1990]) invented a version of statistical inference that depended on a physical randomization process for validity and justification. Neyman's randomization inference used knowledge of the sampling process and the size and nature of the finite population to characterize a sampling distribution for a difference of means (or totals). The Normal shape of this sampling distribution arose from large-sample arguments and the use of the mean (or sums) to summarize causal effects. This mode of large-sample random-sampling randomization inference naturally attracted early survey statisticians, and it has since been very well developed and plays central roles in most major survey sampling projects (see for only three exemplar textbooks Kish 1965; Lohr 1999; Särndal 2003). Neyman's mode is particularly useful when information is high (i.e. samples are large enough to justify a central limit theorem), the sampling process from the popula-

[21] Keele, McConnaughy, and White (2008) provide a detailed argument in favor of applying these kinds of techniques to laboratory experiments.

tion is known, and attention is focused on means or sums (for which central limit theorem results are easily available).

Neyman (1923 [1990]) not only proposed a new method of randomization inference, but he also proposed the idea of potential outcomes as a way to formalize thinking about causal effects. He was especially worried about the fundamental problem of causal inference. His solution was to average over the unobserved potential outcomes and to use the fact of random sampling and the large-sample properties of expected values to show that, in expectation, the expected value of r_{1i} in the treatment group is the same as the expected value of r_{1i} in the control group. That is, Neyman proposed to aggregate potential outcomes to solve the fundamental problem of causal inference whereas Fisher proposed handling the missing data problem with hypotheses. Neyman's solution is very clever not only because the large-sample properties of means are mathematically tractable but also because it linked the problem of causal inference with a method for statistical inference.

Neyman's work thus provided a way to state hypotheses about causal effects. His solution says nothing about the effects on any given unit, but rather posits effects at the level of aggregates: hypothesizing about no average treatment effect rather than no treatment effect. Freedman, Pisani, and Purves (2007, A-32–A-33) call this model a "weak null" compared to the "strong null" discussed in this paper. A weak null hypothesis of no effect of advertising on turnout would posit that, in the unobserved population of cities, the average turnout of treated cities is no difference from the average turnout of control cities. The weak null of no average treatment effects would be satisfied, for example, if the treatment caused some cities to have large negative effects from advertising as long as these negative effects are balanced by other cities having large positive effects. The "strong null", as described here, says that, in the observed set of cities, each city would display the same turnout regardless of advertising. Weak null hypotheses may be tested using randomization inference, as was originally shown by Neyman (1923 [1990]) where the sample in hand arises as a random sample from the population. Knowing the random sampling process allows one to repeat it and thus describe a sampling distribution for the average treatment effect under the weak null.

Weak nulls tend to be most useful when the point of the statistical analysis itself is extrapolation. Strong null hypotheses tend to be of scientific interest when the question is about what happened in the data in hand, and then, as Cox (2006, page 191) notes, "Any question of extrapolation is then one of general scientific principle and method and not a specifically statistical issue." In this paper, we restrict attention to sharp null hypotheses in part because the relationship between these cities and some population of cities is not clearly defined. And also in part because statistical, rather than scientific, extrapolation to other cities, other elections, or other times, would have required a more complex model; a model that could be built on Fisher's foundation, but a model that might detract from the discussion of Fisher's particular insights. This feature of the study, in which we have a set of observations that are not formally a sample from a well-defined population, and in which scientific interest lies in statistical inference about a counter-factual defined within a given study for a given set of units is not uncommon in political science.

Although Fisher and Neyman were famously and acrimoniously at odds with each other, we see that their ideas as complementary. For example, throughout this paper we used confidence intervals (Neyman and E. Pearson's invention) to summarize Fisher's sharp hypotheses. Other examples combining ideas from both Fisher and Neyman include: Hansen and Bowers (2009) use of Neyman's sampling-based inference as an approximation to the complex sharp null hypothesis about heterogeneous treatment effects which allows every unit may have its own treatment effect, and Imai (2008) showing how to assess Neyman's weak null yet comparing the variance estimation for the average treatment effect in-sample (i.e. without extrapolation) with the with variance of the average treatment effect for the population.

7 Conclusion

When might Fisher's method prove particularly useful to political scientists? Recall that the promise of Fisher's method arises in part from what is not required: the validity of the tests (and thus confidence intervals and point-estimates) arises from the design of the study, not the size of the sample or amount of information in the study or information brought to the study from outside the study. When the topic at hand is highly controversial, such that assumption-laden analyses are likely to be attacked on many different grounds, randomization-inference of even large datasets with strong instruments and non-skewed outcomes may complement and check more complex models in which many more moving parts may be criticized. Finally, sharp null hypotheses allows the use of social scientific theory very directly in the process of statistical inference: complex hypotheses may be stated, and with modern computers, assessed to create confidence sets.

Social scientists are comfortable with linear models, and we often know which variables are important for our outcome, but we may not know the exact functional form relating them to each other and to the outcome, let alone to treatment. Randomization inference allows us to use this information in a principled manner. Thus, as elaborated and synthesized here, we can see that randomization inference allows analysts to use scientific knowledge without having to exaggerate confidence in exactly what is known.

In a paper summarizing many of David Freedman's concerns about quantitative social science, Mason (1991) asks for help. Among many pleas, two stand out:

> 3. I'd like somebody to tell me how to make meaningful statistical inferences in the social sciences. When do I really have a population? Or what is my superpopulation, and should I care? …
>
> 10. Analyses based on "all" the data are paradoxical. I once spent a lot of time trying to do an analysis of tuberculosis mortality (Mason and Smith 1985). My analysis was based on population counts. I used maximum likelihood to estimate logistic regressions. There's a problem here. If I've got all the data, why do I need a statistical procedure? If I've got a sample, what do I have a sample of, and how do I figure out what the standard errors are? For that matter, how do I figure out what the right estimation procedure is? My answer at the time was that it was convenient to do what I would have done had I been working with a sample in the usual sense. I am not satisfied with this. Neither are Freedman et al. (1978), who warn their readers to watch out for circumstances like these. Statisticians need to give us more instructive and concrete advice for cases of this kind. (page 349)

Fisher answered both questions in 1935: Statistical inferences in randomized experiments are meaningful as summaries of design-based uncertainty for hypotheses about counterfactuals within the experimental pool. A correct and valid estimation procedure tests the sharp null of no effects based on the randomization done in the study: the right estimation procedure, from this perspective, is one that reflects the design of the study and allows us to repeat it under the null hypothesis. Of course, although Mason probably knew of Fisher's answer, the answer as of 1935 (or even 1991) may have appeared impractical. This paper has shown that advances in statistics and computing enable Fisher's methods to do the kinds of things that political scientists want their statistical tools to do.

Appendix A Simple Assessment of Type-I Error Rates

Here we show the R code that one can use to assess the false positive rate (Type I error rate) using the *t*-test on the newspapers data. Since this is a very small dataset the randomization distribution is defined by the $K = 16$ ways that one may assign treatment within these pairs.

We use sampling from the set of possible treatment assignments here to enable other researchers to use this code most easily when K is very large. In our particular case there are only 16 ways to assign treatment so sampling 1000 times not necessary (but it does not change our substantive results).

First, we define functions and setup the matrix of all of the possible ways to assign treatment:

R Code

```r
set.seed(20110401)
## Produce a matrix (Om) containing all possible randomizations compatible with the design
totrands ← prod(unlist(tapply(news.df$z,news.df$s,
                             function(z){choose(length(z),sum(z))})))
Om ← as.data.frame(matrix(0,nrow=length(news.df$z),ncol=totrands,
                         dimnames=list(row.names(news.df),NULL)))
all.z.1 ← produceRandomizations(news.df$z,news.df$s,totrands)
for(i in 1:16){Om[all.z.1[,i],i] ← 1}
paired.t.stat ← function(r,z,s){
  ## Produce a paired t-statistic
  paired.R ← mapply(FUN=function(r,z){r[z==1]-r[z==0]},r=split(r,s),z=split(z,s))
  mean.R ← mean(paired.R)
  var.R ← var(paired.R)
  mean.R/sqrt(var.R/length(paired.R))
}
p.null.t.test ← function(r=news.df$r,shuf.z,s=news.df$s){
  ## Compare the observed t-statistic to a t-distribution
  obs.t ← paired.t.stat(r,shuf.z,s)
  pt(obs.t,df=length(r)-1,lower.tail=FALSE)
}
p.null.rand.test ← function(r=news.df$r,shuf.z,s=news.df$s,enumerate,nsamps=500,Om=Om){
  ## Compare the observed t-statistic either to the enumerated/exact
  ## randomization distribution implied by the null OR by sampling from that distribution
  obs.t ← paired.t.stat(r,shuf.z,s)
  if(enumerate){
      dist.t ← sapply(Om,function(om.z){ paired.t.stat(r,om.z,s) })
    }else{
      dist.t ← replicate(nsamps,paired.t.stat(r,with(news.df,shuffle.z(z,s)),s) )
    }
  mean(dist.t ≥ obs.t)
}
shuffle.z ← function(z,s){
  ##Shuffle treatment assignment z within blocks defined by s.
  unsplit(lapply(split(z,s),sample),s)
}
```

Next, we repeat the assignment process many time (either as many as possible, i.e. 16, or just very many, i.e. 1000), representing the null of no effects by simply shuffling the values of the assignment variable:

R Code

```
nsims ← 1000 ## for the sampling method
null.t.test.sim ← replicate(nsims,p.null.t.test(shuf.z=with(news.df,shuffle.z(z,s))))
null.rand.test.sim ← replicate(nsims,
                                p.null.rand.test(enumerate=FALSE,
                                                       shuf.z=with(news.df,shuffle.z(z,s))))
## These next lines are for the enumerated method:
##null.t.test.sim ← sapply(Om,function(newz){p.null.t.test(shuf.z=newz)})
##null.rand.test.sim ← sapply(Om,function(newz){
##  p.null.rand.test(enumerate=TRUE,shuf.z=newz)})
```

Finally, we print out a few nominal false positive rates (α) and the corresponding realized false positive rates:

R Code

```
alphas ← seq(.0625,.25,.025)
sapply(alphas,function(a){round(c(nominal.alpha=a,
                            realized.alpha.rand=mean(null.rand.test.sim ≤ a),
                            realized.alpha.t=mean(null.t.test.sim ≤ a)),2)
                          })
```

Print Nominal and Realized False Positive Rates

```
                    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
nominal.alpha       0.06 0.09 0.11 0.14 0.16 0.19 0.21 0.24
realized.alpha.rand 0.02 0.06 0.07 0.11 0.13 0.16 0.20 0.21
realized.alpha.t    0.06 0.13 0.13 0.20 0.20 0.20 0.20 0.20
```

Notice that the randomization based test has a controlled false positive rate — never exceeding the nominal α — while the t-distribution based test does not have a controlled false positive rate and is often overly liberal. This is true for all α but shown for a few values here for illustration.

References

Angrist, J.D., and A.B. Krueger. 1991. "Does Compulsory Schooling Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106(4): 979–1014.

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119(1): 249–275.

Bound, J., D.A. Jaeger, and R.M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak." *Journal of the American Statistical Association* 90(430).

Bowers, Jake. 2011. "Making Effects Manifest in Randomized Experiments." In *Cambridge Handbook of Experimental Political Science*, ed. James N. Druckman, Donald P. Green, James H. Kuklinski, and Arthur Lupia. New York, NY: Cambridge University Press chapter 32.

Bowers, Jake, and Katherine Drake. 2005. "EDA for HLM: Visualization when Probabilstic Inference Fails." *Political Analysis* 13(4): 301–326.

Brady, Henry E. 2008. "Causation and explanation in social science." *Oxford handbook of political methodology* pp. 217–270.

Brazzale, Alessandra R., Anthony C. Davison, and Nancy Reid. 2006. *Applied Asymptotics*. Cambridge University Press.

Chatterjee, A., and SN Lahiri. 2011. "Bootstrapping Lasso estimators." *JASA* .

Cox, D. R. 2006. *Principles of statistical inference*. Cambridge: Cambridge University Press.

Cox, DR, and P. McCullagh. 1982. "Some aspects of analysis of covariance (with discussion)." *Biometrics* 38: 541–561.

Davison, A.C. 2003. *Statistical Models*. Cambridge University Press.

Druckman, J.N., D.P. Green, J.H. Kuklinski, and A. Lupia. 2006. "The growth and development of experimental research in political science." *American Political Science Review* 100(04): 627–635.

Erikson, R.S., P.M. Pinto, and K.T. Rader. 2010. "Randomization Tests and Multi-Level Data in US State Politics." *State Politics & Policy Quarterly* 10(2): 180.

Fisher, R.A. 1935. *The design of experiments. 1935*. Edinburgh: Oliver and Boyd.

Freedman, David A. 2007. "On regression adjustments in experiments with several treatments." *Annals of Applied Statistics (To Appear)* .

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

Freedman, David A. 2008*b*. "Randomization does not justify logistic regression." *Statistical Science* 23(2): 237–249.

Freedman, David, Robert Pisani, and Roger Purves. 2007. *Statistics*. 4th ed. New York: W.W. Norton.

Gadbury, G.L. 2001. "Randomization inference and bias of standard errors." *The American Statistician* 55(4): 310–313.

Green, Donald P. 2009. "Regression Adjustments to Experimental Data: Do David Freedman's Concerns Apply to Political Science?".

Green, D.P., and L. Vavreck. 2007. "Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches." *Political Analysis* .

Hansen, B.B., and J. Bowers. 2008. "Covariate Balance in Simple, Stratified and Clustered Comparative Studies." *Statistical Science* 23: 219.

Hansen, Ben B., and Jake Bowers. 2009. "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association* 104(Sep): 873—885.

Harrell, Frank E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer.

Ho, D.E., and K. Imai. 2006. "Randomization inference with natural experiments: An analysis of ballot effects in the 2003 california recall election." *JOURNAL-AMERICAN STATISTICAL ASSOCIATION* 101(475): 888.

Hodges, J.L., and E.L. Lehmann. 1963. "Estimates of location based on rank tests." *Ann. Math. Statist* 34: 598–611.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(December): 945–960.

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

Imai, K., G. King, and C. Nall. 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, Guido W., and Paul R. Rosenbaum. 2005*a*. "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+.

Imbens, G.W., and P.R. Rosenbaum. 2005*b*. "Robust, accurate confidence intervals with a weak instrument: quarter of birth and education." *Journal of the Royal Statistical Society Series A* 168(1): 109–126.

Keele, Luke, Corrine McConnaughy, and Ismail White. 2008. "Statistical Inference For Experiments.".

Keele, Luke, Corrine McConnaughy, and Ismail White. 2010. "Adjusting Experimental Data: Models versus Design.".

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

Koenker, Roger. 2005. *Quantile Regression (Econometric Society Monographs).* Cambridge University Press.

Koenker, Roger. 2011. *quantreg: Quantile Regression.*

Lehmann, EL. 1998. *Nonparametrics.* Revised first ed. Springer.

Lehmann, E.L., and Joseph P. Romano. 2006. *Testing Statistical Hypotheses (Springer Texts in Statistics).* Springer.

Leisch, Friedrich. 2002. Dynamic generation of statistical reports using literate data analysis. In *Compstat 2002 - Proceedings in Computational Statistics*, ed. W. Haerdle, and B. Roenz. Heidelberg, Germany: Physika Verlag pp. 575–580.

Leisch, Friedrich. 2005. *Sweave User Manual.*

Lohr, S. 1999. *Sampling: Design and Analysis.* Brooks/Cole.

Mason, W.M. 1991. "Freedman Is Right as Far as He Goes, but There Is More, and It's Worse. Statisticians Could Help." *Sociological Methodology* 21: 337–351.

Neyman, J. 1923 [1990]. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (1923)." *Statistical Science* 5: 463–480.

Nolen, T.L., and M. Hudgens. 2011. "Randomization-Based Inference within Principal Strata." *JASA* forthcoming.

Panagopoulos, Costas. 2006. "The Impact of Newspaper Advertising on Voter Turnout: Evidence from a Field Experiment.".

Peduzzi, P., J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. 1996. "A simulation study of the number of events per variable in logistic regression analysis." *Journal of Clinical Epidemiology* 49(12): 1373–1379.

Raab, Gillian M., and Isabella Butcher. 2005. "Randomization inference for balanced cluster-randomized trials." *Clinical Trials* 2(2): 130–140.

Rice, J.A. 2007. *Mathematical Statistics and Data Analysis.* 3rd ed. Belmont, CA: Duxbury Press.

Robins, James M. 2002. "[Covariance Adjustment in Randomized Experiments and Observational Studies]: Comment." *Statistical Science* 17(3): 309–321.

Rosenbaum, Paul. 2009. "Design of Observational Studies.".

Rosenbaum, Paul R. 1993. "Hodges-Lehmann Point Estimates of Treatment Effect in Observational Studies." *Journal of the American Statistical Association* 88(424): 1250–1253.

Rosenbaum, Paul R. 2002*a*. "Attributing Effects to Treatment in Matched Observational Studies." *Journal of the American Statistical Association* 97: 183–192.

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

Rosenbaum, Paul R. 2002*c*. *Observational Studies*. Second ed. Springer-Verlag.

Rosenbaum, Paul R. 2002*d*. "Rejoinder." *Statistical Science* 17(August): 321–327.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. Springer.

Rosenbaum, PR. 2002*e*. "Attributing effects to treatment in matched observational studies." *Journal of the American Statistical Association* 97(MAR): 183–192.

Rubin, D.B. 1974. "Estimating the Causal Effects of Treatments in Randomized and Nonrandomized Studies." *J. Educ. Psych.* 66: 688–701.

Rubin, Donald B. 2005. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions." *Journal of the American Statistical Association* 100(March): 322–331.

Särndal, Carl-Erik. 2003. *Model Assisted Survey Sampling (Springer Series in Statistics)*. Springer.

Schochet, Peter. 2009. "Is regression adjustment supported by the Neyman model for causal inference." *Journal of Statistical Planning and Inference* .

Sekhon, Jasjeet S. 2008. "The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods." *Oxford handbook of political methodology* pp. 271–.

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

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3): 557–586.

Stoker, L., and J. Bowers. 2002*a*. "Erratum to "Designing multi-level studies: sampling voters and electoral contexts" [Electoral Studies 21 (2002) 235-267]." *Electoral Studies* 21(September): 535–536.

Stoker, Laura, and Jake Bowers. 2002*b*. "Designing Multi-Level Studies: Sampling Voters and Electoral Contexts." *Electoral Studies* 21(June): 235–267.

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Whitehead, John. 1993. "Sample size calculations for ordered categorical data." *Statistics in Medicine* 12(December): 2257–2271.

Zou, H. 2006. "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association* 101(476): 1418–1429.