The randomization mode of statistical inference.

Jake Bowers * Costas Panagopoulos†

August 28, 2010

Abstract

How should one estimate and test comparative effects from a field experiment of only 8 units? What does statistical inference mean in this context?

In a randomized experiment the most basic and important inference is between the treatments: after all, the point of randomizing is to allow us to say how the treatment group would have behaved had treatment been withheld.

In this paper we show how one can make such inferences based on *models of the design of the study*, and specifically, on the random process by which the values of the explanatory, or treatment, variable were produced. These *models of assignment* form a basis for valid hypothesis tests, confidence intervals with correct coverage, and point estimates. And this mode of statistical inference is often called "randomization-based" inference or "design-based" 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 eight observations. In addition, we present some innovations in the methodology of randomization inference in the use of linear models as a way to allow outcome- and parameter-models to assist the randomization-based inference without requiring commitments to the usual assumptions that would be required for direct causal inferences using those methods and which would protect the analyst from charges of data snooping.

keywords: randomization inference, randomized experiments, covariance adjustment, linear models

*Assistant Professor, Dept of Political Science, University of Illinois @ Urbana-Champaign (jwbowers@illinois.edu). *Acknowledgements:* Thanks to Chris Achen, Dan Carpenter, Tommy Engstrom, Mark Fredrickson, Don Green, Ben Hansen, and Joe Bowers. Part of this work funded by NSF Grants SES-0753168 and SES-0753164. 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

This paper uses a randomized field-experiment of eight cities to introduce the ideas behind randomization-based inference to political scientists. It also proposes some new methodology to enable the use of regression models for the analysis of randomized-experiments in principled ways.

1 Did newspaper advertisements increase turnout: 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.[1] 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	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. Treatment with the newspaper ads is coded as 1 and lack of treatment is coded as 0 in the 'Treatment' column.

In three of the four pairs, the treated city showed higher turnout than the control city. Did the treatment have an effect? Here, the pairwise-difference of means is 1.5 percentage points of turnout. Is 1.5 percentage points of turnout compatible with a hypothesis of no effects? What do we mean by "no effect"? Fisher (1935, Chap 2) suggests that a treatment has no effect when units would display the same outcome regardless of treatment condition. If that null hypothesis were known to be true, Saginaw would show 16% turnout if given advertising and Sioux City would show 22% turnout if no advertising were done. What do we mean "compatible"? Imagine we re-ran this experiment many times, each time reassigning treatment within pairs

[1] 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. For Costas to check numbers and description.]

but leaving turnout unchanged to represent the hypothesis of no effects, and each time recording the pairwise-difference of means. The set of numbers we receive from this replication represents the values that our mean difference could take on under the thought experiment that the treatment had no effect. Is the number 1.5 surprising in comparison to this "no effect"-distribution? A p-value would encode how surprising 1.5 is from the perspective of the null hypothesis.

Engaging in the computational thought experiment with the eight city data, the "no effect"-distribution (more commonly called the "null-" or "randomization-" distribution) for the mean-difference ranges from -5 to 5. About 1/3 of these differences are greater than or equal to 1.5. Did the treatment have an effect? Our observed value does not cast doubt on the null of no effects. Of course, it is possible that the treatment had an effect that we cannot detect here, or that it would have an effect if it were given to a different set of cities. All we can say here, without adding assumptions, is that we can't easily rule out the idea that treatment had no effect.[2]

To summarize: Our simple scientific investigation centered around stating a substantively meaningful hypothesis and then doing the computational thought-experiment in which the null hypothesis held and in which the assignment mechanism was repeated many times. How often did those repetitions give rise to a test statistic that is as surprising from the perspective of the null hypothesis as the statistic that was obtained in the real-life experiment? Only about 1/3 of the time. We do not-reject the null of no effects.

What are some of the strengths of this approach for this design and these data? We were able to execute a substantively meaningful hypothesis test without having to know either how these cities relate to any larger population of cities nor having to know anything about how turnout is distributed in such a population. Usually, not knowing that turnout is Normally distributed in the population, one might use asymptotic arguments to get p-values. But our sample size is eight. Luckily, we did not have to make asymptotic arguments here.

What are some of the weaknesses of this approach as developed by Fisher? First, testing the strict null of no effects does not tell us much about the likely size of the effect: it addresses questions about whether any effect is manifest (in Fisher's terms), but not questions about the size of the effect. Second, we did have to repeat the experiment and did not, in this case, rely on close-form expressions that allow nearly instantaneous computation as would be the case if we had used a t-test.[3] Third, as developed by Fisher, this approach does not allow us to use subject matter knowledge to improve the precision as regression models tend to allow. Fourth, it might appear that relaxing distributional and asymptotic assumptions is relatively minor compared to the assumptions commonly required of causal inference, such as SUTVA (cite), which do not, on their face, appear any less binding of randomization inference than

[2] All computations reported in this paper are available for reproduction and exploration by interested readers from `http://jakebowers.org`. This paper is written in the mixture of R and LaTeX known as Sweave (Leisch, 2002, 2005). I encourage those those interested in learning more to download the source code of the paper and apply themselves to adapting it for their own purposes.

[3] Fisher did provide a closed-form expression in his Lady Tasting Tea example in 1935. But, each design could be seen to require its own closed-form expression. This would be another weakness but it is rendered irrelevant by our computational thought-experiment.

they would of other forms of inference. Finally, the statistical inferences made here say nothing about cities or treatments not observed, thus, this method only speaks to what Cook and Campbell have called "internal validity" rather than "external validity", or the ability to directly apply, in a statistical sense, what one learns from a sample to a population.

This paper contributes to political science methodology by addressing each one of these weaknesses in turn either by explaining developments extant in the field of statistics or by proposing new methodology. Developments since Fisher allow us to much broaden the scope of his simple idea. By overcoming the weakness listed above, I hope to show that randomization inference now has the potential to become a useful tool for those using randomization to study political phenomena.

2 Testing Hypotheses About Effects

We have seen how to assess a hypothesis about no effects. What if we had a hypothesis about some effects? For example, if we knew that most elections in two-party systems were won or lost by small margins of victory (say, 2 percentage points), we might ask whether our observed result of 1.5 percentage points would be surprising from the perspective of a 2 percentage point increase hypothesis. In order to assess such a hypothesis, and to enhance clarity in the rest of the paper, I first link Fisher's ideas about scientific hypotheses and statistical inference to a formal conceptual and notational framework for causal effects.

2.1 Causal effects as comparisons of potential outcomes

What is the causal effect of newspaper advertisements on aggregate vote turnout in the Panagopoulos Newspapers dataset? By "causal 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.[4] 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}.[5] We could represent these potential outcomes in the Newspapers design as follows in Table 2.

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 how the

[4] We can write r_{0i} because we are training attention on hypotheses in which treatment given to one city does not matter for another city. If we wondered about hypotheses in which 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.

[5] 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.

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.

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?

2.2 Hypothesizing about causal effects.

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 is very specific in that it specifies the potential outcomes for each unit in the experimental pool. For example, a strict null hypothesis of "no effect" implies that, for all i, $r_{1i} = r_{0i}$. If this hypothesis were known to be true, the missing potential outcomes in Table 2 would be known.

2.3 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: after all it is within pairs that treatment was assigned.

Write R_{bi} for the observed percent of registered voters who voted in 2005 in each city $i \in \{1, 2\}$ in pair $b \in \{1, 2, 3, 4\}$. 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, writing $\mathbf{Z}$ as a vector of ones and zeros and $\mathbf{R}$ as a vector of observed outcomes, as:

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

Where $m_b = \sum_i Z_{bi}$ is the number of treated units i, $i = 1, \ldots, n_b$, in block b, $b = 1, \ldots, B$. The total number of units in the study is $n = \sum_{b=1}^{B} n_b$. Here we assume only two treatments (thus the number of control units in a block is $(n_b - m_b)$.[6]

And we can calculate the overall difference as the average of the within pair averages from equation 1 weighting simply by the size of block (for the case when blocks are of equal size, for example, in the paired case):

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

The within-block formula written in equation 2 is not generalizable to situations where blocks contain more than one treated unit and/or more than one control unit but is simple and useful for this data example.[7]

Applying equation 2 to the Newspapers dataset gives an average difference of treated and control units within pairs in turnout after treatment of 1.5 percentage points.

2.4 Simple Inference for a Null of No Effect

Now we can link the informal discussion of hypothesis testing § 1 that to the potential outcomes framework: Fisher's null hypothesis understood in this frame work states: $H_0 : \tau_0 = 0$ which implies $r_{1i} = r_{0i}$. Notice that we can link this hypothesis about the only partially observed counterfactual of substantive interest to what we observe by the following identity: $R_i = Z_i r_{1i} + (1 - Z_i) r_{0i}$ (dropping the block subscript for the moment in the interests of clarity). Why should shuffling treatment assignment and recalculating the test statistic, $t(Z, R)$, represent the null hypothesis of no effects? If, for the sake of argument, we granted the null, then we would be saying that $r_{1i} = r_{0i}$, and this in turn would imply $R_i = Z_i r_{1i} + (1 - Z_i) r_{0i} = Z_i r_{0i} + r_{0i} - Z_i r_{0i} = r_{0i}$. That is, the null hypothesis and the identity linking observed to potential outcomes allows

[6] We can think of $m_b = \sum_i Z_{bi}$, $n_b - m_b$ as fixed quantities either by design or via conditioning argument of the type articulated in Hansen and Bowers (2008, §3.3) or more generally in Cox (2006). Cox (2006, 1989–190) highlights the fact that outcomes are considered fixed here as a feature of the design and that the conceptual framework of randomization-based inference.

[7] 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 such that:

$$
t(\mathbf{Z}, \mathbf{R}) = \left(\sum_{b=1}^{B} \frac{n_b}{m_b (n_b - m_b)} \right) \cdot \sum_{b=1}^{B} \frac{n_b}{m_b (n_b - m_b)} t(\mathbf{Z}_b, \mathbf{R}_b). \tag{3}
$$

See Hansen and Bowers (2008) for formal arguments justifying the weight used here. See Imai (2008) for another randomization-based argument in favor of weighting unequal sized blocks equally.

a hypothesis about a causal effect to have an implication on what we observe. Entertaining the sharp null of no effects means that $R_i = r_{0i}$ (all observed responses are those that we would observe under the control condition). Thus, our null hypothesis also implies something about a test statistic summarizing a relationship between Z and R: that is, in calculating $t(Z, R)$ for each re-shuffled assignment, under the null, we are actually calculating $t(Z, r_{0i})$. Thus, the distribution of the test statistic as generated by repetition of the experiment can be directly linked to a causal quantity (where by "causal" here in the context of a randomized experiment we mean, narrowly, a quantity that we can write down as a counter-factual).

Above, we used mean differences (equation 2) as our test statistic to summarize the evidence in the data that might speak to this null. And we followed the classical frequentist approach to quantifying the relationship between a piece of observed data and a hypothesis: we used a one-sided p-value. In our case, we would expect to observe a $t(Z, r_{0i}) \geq 1.5$ under the null about $1/3$ of the time: where "of the time" refers to repetitions of the null-thought-experiment.[8]

The same causal hypothesis can be assessed with difference 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, like means, 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.[9] Lehmann (1998) and Keele et al. (2008) among others suggests rank based tests for their insensitivity to outliers. A one-sided p-value for the test based on a paired signed rank-sum test statistic is p=0.4375.

Now we are prepared to address the weaknesses mentioned above: First, political scientists have at least a rough sense of what kinds of covariates relate to turnout. That is, although precise models of outcomes may be a burden, we would like to use what we know to increase the precision with which we evaluate hypotheses. 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.

[8] I don't report a two-sided p-value here, although it would be easy to do so. A common definition of a two-sided p-value is twice the one-sided p-value for test statistics with only positive domains (like ranks) or twice the smaller of the two one-sided p-values for asymetric randomization distributions which is equivalent the sum of the absolute value of the mass at or greater than the observed value in symmetric distributions. See Rosenbaum (2010, page 33) and Cox et al. (1977) for arguments justifying twice the minimum of the two one-sided p-values.

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

2.5 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 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 0.375 and 0.4375 for the mean and rank-based test statistics respectively. Regardless of test statistic, we so far cannot exclude $\tau_0 = 0$ from 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?

2.5.1 Hypothesis generation functions: Models of Effects

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, and these hypotheses are not limited to the $r_{1i} = r_{0i}$. A "model of effects" is a function that produces such specific statements about potential outcomes. For, example, a common and useful model states that

$$\tau = r_{1i} - r_{0i} \tag{4}$$

which implies that $r_{1i} = r_{0i} + \tau$. This model 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. Are the hypotheses generated by the constant, additive model of effects scientifically useful or interesting? In this case, one might imagine so — the same kind of media campaign within US cities in the same year might reasonably have the same effect in each city. If, however, one were interested in some other hypothesis, say, that cities which have high potential outcomes in response to control (measure, perhaps, as cities with high turnout at baseline) ought to have stronger responses to the treatment than cities which have low potential outcomes in response to control, we might say: $r_{1ib} > \theta > r_{0ib}$ where θ is some value of the order statistics of r_{Cib}. This is a model of "displacement effects" (Rosenbaum, 2002e, Chapter 5). Any scientifically interesting hypothesis may be formulated in this way and assessed using the machinery described above. Rejection of a null hypothesis means that our data are surprising from the perspective of the hypothesis. Our data could be surprising either because our model generating hypotheses is not well supported by the data (say, if $r_{1i} = \tau * r_{0i}$ rather than $r_{1i} = \tau + r_{0i}$) or because our model of effects 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 they might be too wide if, say, $r_{1i} = \tau_i + r_{0i}$ as shown by (Gadbury, 2001) and Robins (2002, § 2.1).[10]

Consider $H_0 : \tau = \tau_0$ to generalize from the simple hypothesis of no effect. If $\tau = \tau_0$

[10] Rosenbaum (2002f, § 3–6) explains the equivalences between estimating an average treatment effect and testing a sequence of hypotheses about individual causal effects.

and entertaining as substantively interesting a model of constant, additive effects from equation 4 where $r_{Tsi} = r_{Csi} + \tau_0$, then:

$$\begin{aligned} R_{bi} &= Z_{bi} r_{1bi} + (1 - Z_{bi}) r_{0bi} \\ &= Z_{bi} r_{0bi} + Z_{bi}\tau + r_{0bi} - Z_{bi} r_{0bi} \\ &= r_{0bi} + Z_{bi}\tau_0 \qquad (5) \\ r_{0bi} &= R_{bi} - Z_{bi}\tau_0 \qquad (6) \end{aligned}$$

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 $H_0 : \tau = \tau_0 \;\Rightarrow\; R_{ib} = r_{0ib} + Z_{bi}\tau_0$.

The logic of § 2.4 can apply directly here. Now the test statistic is $t(\mathbf{Z}, \mathbf{R} - \mathbf{Z}\boldsymbol{\tau}_0)$ rather than $t(\mathbf{Z}, \mathbf{R})$. Thus, for a given hypothesized value of τ, τ_0 where the "0" stands for null hypothesis, we can repeat the assignment process of the experiment by generating new 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.

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, I link my interpretation to what readers might find by reading works in statistics. That work 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) 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 7 summarizes the doubt cast by our observed test statistic against the null hypothesis:

$$\Pr(t(\mathbf{z}, \mathbf{R}) \geq t(\mathbf{Z}, \mathbf{R}) | \tau = \tau_0) = \frac{\sum_{\mathbf{z} \in \Omega} \mathbf{1}\{t(\mathbf{z}, \mathbf{R}) \geq t(\mathbf{Z}, \mathbf{R})\}}{K} \qquad (7)$$

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 $K = \prod_{s=1}^{4} 2 = (2)^4 = 16$. To test a hypothesis about constant, additive effects, substitute $\mathbf{R} - \mathbf{Z}\boldsymbol{\tau}_0$ for $\mathbf{R}$ in that equation. If the p-value is greater than or equal to our α value, τ_0 is inside the confidence interval, otherwise it is excluded from the confidence interval.

For example, say we want to test $H_0 : \tau = 1$. Our model of effects and the logic of equation 6 says that, if our null hypothesis were true, potential outcomes to control among the treated would be potential outcomes to treatment minus $\tau_0 = 1$: $r_{0bi} = R_{bi} - Z_{bi}\tau_0$. 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 repetition of the experiment calculate the

test statistic, now using the adjusted outcomes, $t(\mathbf{z}, \mathbf{R} - \mathbf{z}\tau_0)$, and refer to equation 7 for a p-value.

Using the paired rank sum statistic 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 cannot 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 $100(1-(1/16)) \approx 94\%$ one-sided CI but in practice such a CI would be incredibly wide. The mean difference test statistic did not reject constant, additive effects hypothesis between [-3.4, 6.4] at the 88$ confidence level.[11]

We could produce a point estimate by shrinking this confidence interval.[12] Yet, for this application, the confidence intervals themselves summarize the evidence adequately: we can reject the hypotheses, at the 88% level, that the advertisements increased or decreased turnout, at the outside, by more than about 6 percentage points.

Notice, that 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. 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, 2002e, 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 (2002e, Chapter 2) also proves that these intervals are unbiased and consistent (such that more information is better — produces smaller intervals) but that the correct coverage of the intervals does not depend on the sample size or correctness of some model of outcomes.

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

We must recall that political scientists are not ignorant about turnout. 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. In fact, 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"). Bowers (2011) and Keele et al. (2010) provide randomization-inference

[11] We used two-sided tests here to create confidence intervals with well-defined ends. We could have also easily, used one-sided tests to produce a one-sided interval assessing null hypotheses about only positive effects of the putatively turnout enhancing intervention.

[12] This is a very informal way of talking about what is called a Hodges-Lehmann point estimate (Hodges and Lehmann, 1963; Rosenbaum, 1993).

oriented overviews to the use of covariates for adjusting experiments. The problem with adjustment, has long concerned experimentalists and has created extensive debate in the methodology of randomized clinical trials. In brief, here is the problem using this study as an example:

We imagine that current vote turnout ought to relate to past turnout. In fact, the list of variables thought to relate to aggregate turnout is probably longer than the number of cases in this dataset. "Adjustment" for covariates in experiments like this one would usually means regressing outcomes on a treatment indicator plus some linear function of covariates. Some of these covariates might be entered into the equation as interactions with other covariates, or as polynomials, or cut into discrete categories, or simply assuming a linear and constant relationship with outcomes and treatment and other covariates. Whereas treatment effects assessed without adjustment are not subject to criticisms that the results arise from what Leamer (cite) might call "whimsical" modeling decisions, some scholars (Deaton and/or Senn cite) have claimed that the act of adjustment removes all of the benefits of randomization. Choosing one particular adjustment strategy among many requires justification of that choice: after all, in one among many regressions the treatment effect may appear large and "statistically significant" merely through the operation of multiple testing or because the regression equation itself has so many covariates that the actual treatment-vs-control comparisons are based entirely on extrapolation (i.e. there are no highly competitive, small number of candidates, large population size, low past turnout cities with both treated and control observations). And, critics may be concerned that such adjustment might induce non-independence with as yet unobserved covariates — removing the benefits of randomization. Further, the coverage of confidence intervals based on regression models (including all common maximum likelihood justified models) are based on getting the model correct. So, it would seem as if the appealing simplicity of randomization disappears and the burdens of specifying and justifying and checking statistical models reappear upon the decision to adjust.

As I highlight above, this need not be so. First, Rosenbaum (2002d) provided a deep yet simple insight that allows for covariance adjusted randomization inference for treatment effects that has the same statistical properties of any other randomization-based inference. Second, here I advance a proposal for building on his insight as a way to address the otherwise ever present concerns about data snooping that attend to regression modeling. Third, I explain a large-sample but randomization-based method of covariance adjustment that allows not only the use of information from within the sample, but the use of information on units not sampled (units which provide us information about the relationship between turnout and other covariates but which are neither formal controls or treated units) developed by Hansen and Bowers (2009) (discovered previously by Peters (1941) and Belson (1956)). Both methods of using regression models allow for randomization-based inferences that use substantive knowledge about the outcome under study to shrink our confidence intervals.

3.0.2 Rosenbaum(2002) style covariance adjustment

Recall that the procedure for randomization inference depends on focusing attention on a specific substantively interesting set of hypotheses about counterfactual comparisons. 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 $\mathbf{r}_0 = \mathbf{R} - \tau_0\mathbf{Z}$. The design of the experiment is allows us to instruct our computers to repeat it while calculating $t(\mathbf{Z}, \mathbf{r}_0)$ 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 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 (2002d, §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 regression 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 just as $\mathbf{r}_0$ itself is fixed under the null when we do randomization inference without adjustment.

Compute adjusted outcomes based on a $H_0 : \tau = \tau_0$ So far we have been interested in investigating hypotheses of the form: $\mathbf{r}_0 = \mathbf{r}_1 - \tau$. This hypothesis-generator or model of effects implies 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 $\mathbf{R}$. Of course, we can specify any substantively meaning function to generate hypotheses and adjust outcomes in this way.

Compute $t(\mathbf{Z}, \mathbf{e}_0)$ **and compare to** $t(\mathbf{z}, \mathbf{e}_0; \mathbf{z} \in \Omega)$**for a p-value** where Ω can be thought of as the result of repeating the assignment process of the experiment many times.

Models relating covariates to outcomes are meant only to reduce noise in the outcomes in this method. A correct specification is not required. Incorrect specifications will add noise and will thus make the confidence interval wider but will not change the coverage of the confidence interval. We emphasized $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x})$ above — a noise-reduction or residual-producing function — rather than $\hat{R} = \mathbf{x}\hat{\boldsymbol{\beta}}$ 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. This is not to say that one must not be thoughtful in choosing such a model — a bad choice could inflate the confidence interval.

In his article introducing these ideas, Rosenbaum (2002d), 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. [13] 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. Such 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.[14]

In addition, while relatively expensive information about treatment-vs-control comparisons may only be available within an experimental pool (such as the eight cities discussed here), information about covariate-outcome comparisons may be more plentiful (and cheaper). Hansen and Bowers (2009) re-discovered a method for covariance adjustment published by Peters (1941) and Belson (1956) but elaborated the links between it and the assessment of counterfactual comparisons using randomization inference. Hansen and Bowers (2009) relied on Central Limit Theorems to speed computation of confidence intervals (thus limiting the applicability of their method to large-samples) but also provided a stress-testing framework to check whether a given sample was large enough for such limit theorems to operate.

3.1 The Peters-Belson Method

TODO: show the Peters-Belson method using extra cities.

3.1.1 Using non-sample control units to further decrease non-treatment-related noise in outcomes

Now, before, I demonstrate (1) how covariates which predict the outcome (often called "prognostic covariates") can decrease the width of confidence intervals and (2) covariates which do not predict the outcome may only increase the width of the intervals but do not change the coverage rates of the intervals, let me engage with the elephant in the room: choice of adjustment strategy.

3.2 Principled Regression Model Specification for Randomized Experiments

Rosenbaum reminds us that:

[13] 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 (2002d, § 2.4) in using Huber's M-estimator as implemented in the `rlm` function packaged with `R`.

[14] 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).

> 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, 2002d, page 290)

How should one choose covariance adjustment specifications among the many? Running many many different linear models searching for the one which provides the lowest p-value on the treatment effect is not the way to go. [more on multiple testing problems and relatedly discretion]

One rather bullet-proof option is to declare, in advance of the assigning treatment (or at least, in advance of inspecting outcomes) (1) the covariates for which one plans to adjust and (2) the regression modeling specification one plans to use. [cite to experiment registers in clinical trials on this]. Such public declarations protect analysts from claims of data snooping. Another method, advanced here, will be useful if, for some reason, new data become available for use in adjustment that were not available before the experiment was run, or if, for other reasons, advance registration of covariance adjustment strategies is impractical (say, during secondary analysis of already collected experimental data).

Recall that from the perspective used so far we assess treatment effects using linear models merely to transform our response (a residual is just a transformed response). It is well known that different functions of the response may have more or less power to detect treatment effects under different data configurations. For example, when $\mathbf{R}$ is approximately Normal, tests using means tend to offer tighter confidence intervals than tests using ranks, conversely, when $\mathbf{R}$ has long tails, is skewed, or otherwise has large outliers, then tests using ranks out-perform tests using means [cites to Lehmann among others on this]. If one is not estimating treatment effects, one may compare the performance of a means-based with that of a ranks-based test without worry that one is actively inspecting the treatment effect [cite to Rosenbaum and/or Imbens and Rubin]. That is, Fisher-style randomization inference allows us, in principle, to compare the performance of test statistics. The $t(\mathbf{Z}, \mathbf{e})$ calculated from the different regression specifications in § 3 are each different test statistics, and are thus comparable in terms of their power.

Imagine that the true causal effect were $\tau = 0$. Say, we tested the null hypothesis that $\tau_0 = 5$ when the true effect was 0. We would want to reject $\tau_0 = 5$ if $\tau = 0$, but we also know that we might not reject $\tau_0 = 5$ in a small sample — a small sample might not give us enough information to distinguish between 0 and 5 percentage points of turnout effects. As sample size grows we would imagine that our ability to distinguish between 0 and 5 percentage points of turnout would also grow. We would be more likely to reject $\tau_0 = 5$ if $\tau = 0$ when the sample size increases. That is, our tests would be more powerful as sample size grows.

Tests may vary in their power even when sample size remains the same. And, as I explained, different regression specifications amount to different test statistics. To choose a regression specification that maximizes power without data snooping (without looking at comparisons of treated to control outcomes) one can engage in a thought experiment like the one just described. Here is our procedure:

1. Posit that $R_i = r_{0i}$ to start. Represent a "true" (for the sake of the simulation) effect using, for now, and for convenience, the idea of treatment effects occurring in a constant and additive fashion such that $r_{1i} = r_{0i} + \tau$. That is, we generate r_{1i} for each unit by adding τ to its observed outcome (which we take to represent r_{0i} for the sake of the power analysis).[15]
2. Now, test null hypotheses, say, $\tau = \{0, ..., 10\}$. More powerful test statistics (aka regression specifications+test statistics) will be more likely to reject a given null than will less powerful test statistics.
3. Repeat this procedure, choosing different values for the "truth" (say, also $\tau_0 = \{0, ..., 10\}$).
4. Assess the strict null of no effects against the one-sided alternative of $\tau > 0$ using both rank-based and mean-based test statistics on the residuals from a variety of covariance adjustment specifications.
5. The proportion of p-values less than some rejection threshold α (say, here, $p = .125$ to denote inclusion in an 88% confidence interval) is an indication of the power of the tests.

The proportion of p-values lower than some fixed number α is a reasonable comparative measure of power here: For each entertained treatment effect τ, we test the null that $H_0 : \tau = 0$. Recall that a powerful test is one which rejects the null when the alternative is true. Here, we set the alternative to a series of values (that is, the alternative is always true, and it is just more or less easy to detect as τ increases — larger τ ought to be easier to detect). Thus, larger τ should be associated with smaller p-values. Consider two tests of $H_0 : \tau = 0$ against $H_A : \tau \geq 0$. With a sample size of 8, an entertained $\tau = 5$ could count as a different amount of evidence against the null: a powerful test might consider it a lot of evidence, a less powerful test might consider it little evidence.

Figure 1 shows one way to inspect the results of such a power analysis. The test statistic with the least power against the alternatives given a true null of no effects is the unadjusted version (labeled with "r_0" alone). Two regression specifications appear to similar power against alternatives close to and far from zero: one using baseline turnout and number of candidates in the election (r_0 $r_{0,t-1}$ + candidates) and the other using the median age of the city and the number of candidates (r_0 age + candidates).

I call this method of choosing a regression specification "principled" because at no point do we inspect the relationship between treatment and outcomes. Thus, we do not worry that we are choosing a specification in an effort to argue in favor of some idiosyncratic choice of treatment effect.

[15] Previous versions of this paper generated control outcomes from a Normal distribution based on the information in the control group in the sample and then added the true null to these Normal variates. That method is also a reasonable approach with Normal-ish outcomes. The method used here, now, makes no claims about the distribution of the outcomes even for the sake of the power analysis

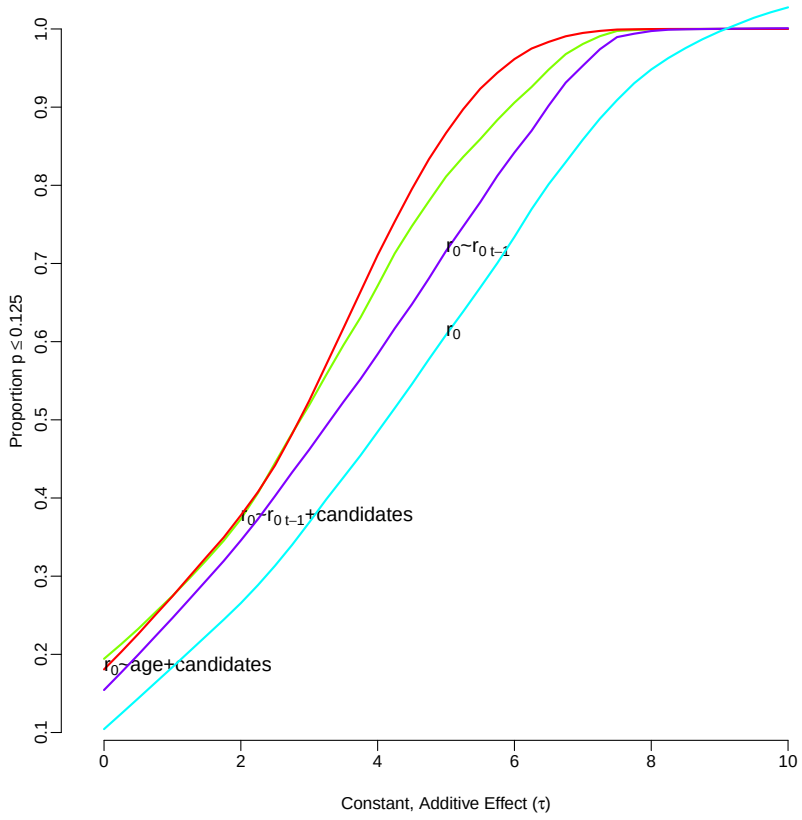

Figure 1: Randomization-based power assessments for OLS regression models (r_0 stands for no model or equivalently a constant model) using the paired signed rank sum test statistic. Specifications which reject the true null of $\tau_0 = 0$ at a one-sided test using $\alpha = .12$ more frequently have more power. The power curves have been smoothed using loess with a span of $1/2$ to decrease distraction caused by discreteness.

3.3 Covariance-Adjusted Randomization-Inferences for the Newspapers Study

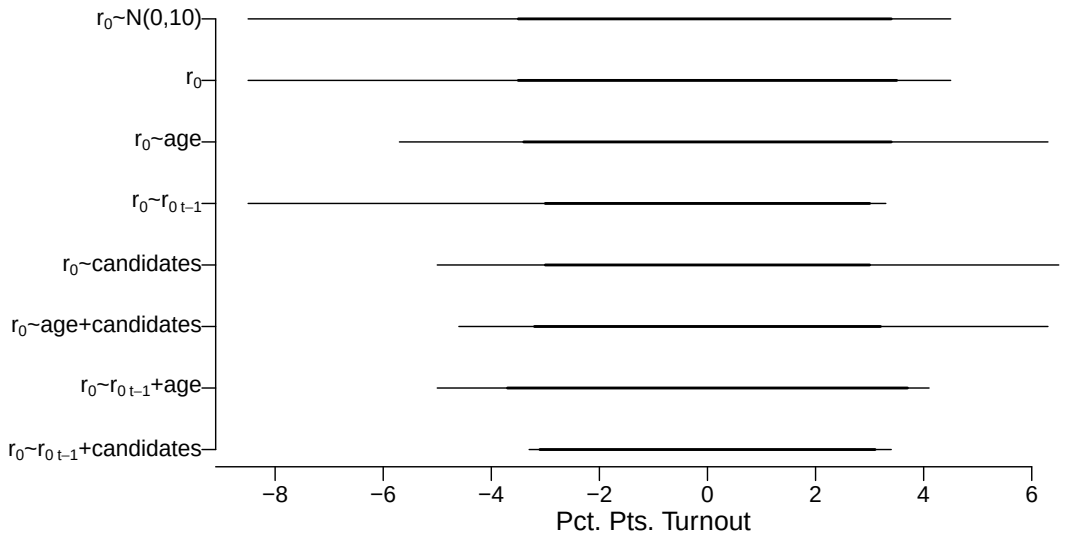

Figure 2: 88% (thin lines) and 2/3 (thick lines) Confidence Intervals for effects of Newspapers advertisements on Turnount (Percentage Points of Turnout). All models run with pair-aligned data [i.e. treated-control outcomes and covariates] (equiv. fixed effects for pair). Here $\tilde{\varepsilon}(\mathbf{r}_0, \mathbf{x})$ uses OLS and $t(\mathbf{Z}, \mathbf{R})$ is the paired signed rank sum statistic. Turnout in the previous election is labeled $r_{0\,t-1}$. "Age" is median 2000 census figures for the city. "Candidates" is number of candidates in the election. Noise only model uses a draw from a Normal distribution mean=0, sd=10 (labeled N(0,10)).

Figure 2 shows the 88% and 2/3 confidence intervals that arise from different covariance adjustment specifications. Each confidence interval is labeled with the linear model formula from the noise-reduction function, and the intervals are plotted in order of width (widest on top, narrowest on bottom). The confidence interval without covariance adjustment [-6–7] is labeled r_0 — it is the 2nd line from the top of the plot. The widest confidence interval arises from the model regressing outcomes on N(0,10) uncorrelated noise. Adding noise expands the intervals. Also, as expected, covariates which predict the outcome reduce the interval: the bottom line is an 88% CI running from -3.3 to 3.4 percentage points of turnout after removing the linear and additive relationships between baseline turnout and the number of candidates and post-treatment turnout. Recall that this was one of the regression specifications with the highest power arising from our power analysis as shown by Figure 1. Notice also that this simple covariance adjustment decreased the width of the 88% interval (the thin line) by half.

In this particular case, the sample size limits the possibilities for covariance adjustment — only so many covariates may be used in a regression model with 8 observations and pair indicators (or, equivalently, with 4 paired observations).

3.4 Regression without Regrets?

[TODO: Summary of the Rosenbaum and the Peters-Belson styles of randomization-based yet model-assisted statistical inference.]

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). The statistical inferences do not depend on the correctness of the regression model nor assumptions about any particular error process.

4 Beyond the constant, additive effects set of hypotheses: Assessing hypotheses about interference

Fisher's sharp null hypothesis of no effects is an excellent representation of what we mean when we say "no effects." We have focused on one set of potentially interesting substantive hypotheses, those in which $r_{1i} = r_{0i} + \tau$. In the context of binary outcomes, other work has entertained hypotheses positing that each unit has its own unique response to treatment $r_{1i} = r_{0i} + \tau_i$ (Hansen and Bowers, 2009; Rosenbaum, 2002c, 2001a) and (Rosenbaum, 2002b, Chap 5). What about hypotheses in which treatment given to one unit changes the potential outcome to another unit? Such hypotheses are required to be assumed away in most other approaches to causal inference from randomized experiments as part of the invocation of the SUTVA assumptions, but, in this framework, our substantive concerns in this paper have so far lead us to focus on hypotheses implying the SUTVA assumption, but this is not required. Here, I demonstrate that one may entertain and test hypotheses implying interference among units.

TODO!

5 What does "goodness of fit" mean for hypotheses?

Recall that we have been careful to talk about not hypotheses as assumptions, but hypotheses as hypotheses, as substantive questions posed against which we consider the data as evidence. How might we know when our particular hypothesis-generating function (or model of effects) "fits" the data? What does "fit" mean in this context? We know that a model of effects fits the data well when, if we remove the hypothesized effect from each treated observation, any remaining differences between the treatment and the control group ought to be due only to chance.

Say we want to consider the model of $r_{1i} = r_{0i} + \tau$, and, given this model, the the median of the confidence intervals produced using the most powerful covariance adjustment formula (the one using age and % black) is 1.5.[16] Figure 3 shows how one might go about this kind of assessment (inspired by Rosenbaum (2010, Chapter 2)). The left panel shows the data as collected — the vertical axis is proportion turning out the vote in the city minus the mean proportion turning out to vote in the pair, thus removing the "pair effect" from the data. The dark lines in the middle of the boxplots

[16] See Hodges-Lehmann (and Rosenbaum, or Cox 2006) for why the median hypothesis not-rejected might be a good operationalization for "best guess".

are the control and treatment group means. Visual inspection suggests some kind of positive treatment effect, and the general shifting of the distribution in the treatment group relative to the distribution in the control group supports the model of constant additive effects (although the different within-pair differences might argue against such a model). The right panel repeats the same plot for the control group but applies the estimated τ to the treatment group in accord with our model of effects. Now the mean turnout in the treatment group is the same as the mean in the control group. And the middle portions of both distributions (indicated by the box in the boxplot) also are the same. However, the medians still differ, with the median in the treatment group higher than the median in the control group. This model does bring 3 out of 4 of the individual cities closer to their matched control, but it increases the distance within the triangle pair, and more importantly, still leaves some of the turnout in the treatment group unaccounted for.

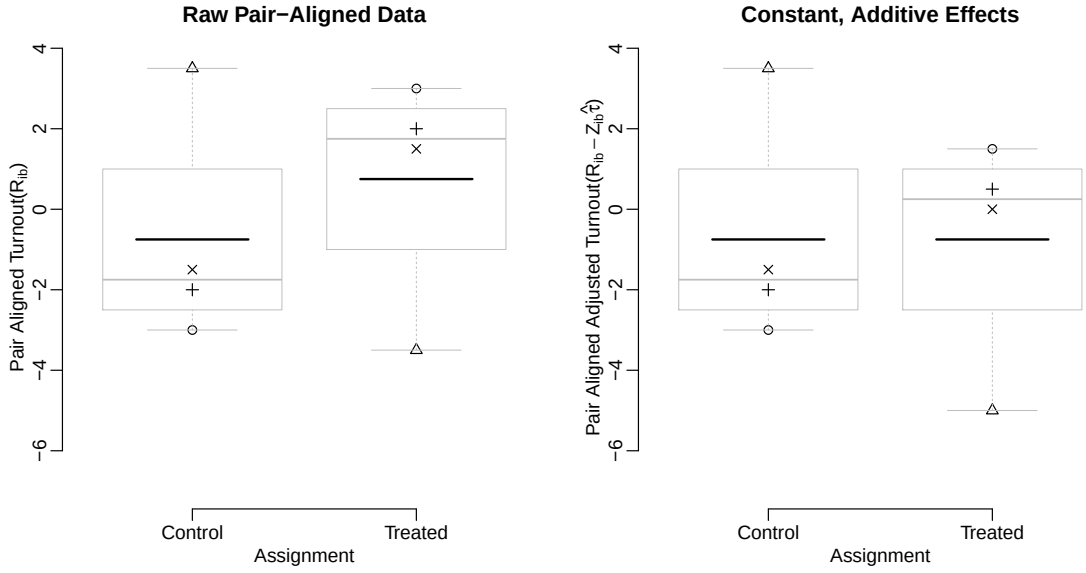

Figure 3: Raw responses of control and treated cities (left panel) and raw control responses compared to treated responses adjusted for a constant, additive model of effects (right panel). A good fitting model of effects would make the two groups look equivalent but for random noise. Pairs are marked with symbols. Since the baseline turnout differed between pairs, turnout here has been "aligned" or "pair-mean centered" to enable easier visual comparison of responses.

What about a weaker hypothesis? Perhaps something more like $r_{1i} = r_{0i} + \tau_i$ but where we are predominantly interested in the average difference in potential outcomes? There is a sense in which our assessment of the constant, additive effects model is an assessment of our model of an average treatment effect. Of course, there are many possible collections of $\boldsymbol{\tau}$ which could produce the same average of 1.5. (for example, we could have some enormously positive effects within some pairs and enormously negative effects within other pairs — or no effect in all but 1 pair, but a large and counterbalancing effect in the other pairs.) Thus, although this diagnostic for models of effects was designed to compare what happens given sharp and focused models of effects, it could also be useful to assess the usefulness of the more fuzzy average treatment effect presumption.

6 Discussion

6.1 Randomization Can Justify Statistical Inference

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 (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 no effect) requires adding substantive knowledge about the process relating treatment to potential outcomes.[17] Rejecting null hypotheses about effects may either indicate that $\tau \neq \tau_0$ or that the model of effects is not supported by the data. 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) nor on asymptotics nor 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 one can base 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, whose only training in statistics has occurred within departments of political science, sociology, or economics, will be, we hope, pleasantly surprised.

In a sense, randomization inference allows scientific attention to return back to 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.[18] 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).

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 validity

[17] Note that this is one way in which Fisher's framework differs from Neymans: Neyman poses a "weak" null (about average differences) and focuses attention on estimating a difference of averages — that is, focuses on a particular model of effects at the aggregate level. Fisher begins with no model of effects but then requires them once hypotheses about effects are entertained.

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

in the eventual judgement of the scholarly community about the value of a study or theory or finding.[19] 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, researchers have a much more solid sense of the specific from which to address the general. [Cites to Rosenbaum 2010 and Cook and Campbell: good generalization requires confidence in the internal validity of a study. This is what randomization and randomization-based inference provide.]

There is an alternative way to state hypotheses about causal effects that is worth mentioning. A common model says nothing about the effects on any given unit, but rather posits effects at the level of aggregates: the average treatment effect is one common such model. Freedman et al. (2007, A-32–A-33) call this model a "weak null" compared to the "strong null" discussed here. A weak null hypothesis of no effect of advertising on turnout would say that, in the population, treatment with advertising produces no higher turnout than no advertising, on average. The "strong null", in contrast, says that, in this set of cities, each city would display the same turnout regardless of advertising. Weak null hypotheses may be tested using randomization, as was originally shown by Neyman (1990), and require a model whereby the sample in hand arises from the population — they 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, I restrict attention to sharp null hypotheses in part because the relationship between these cities and some population of cities not clearly defined. 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 what happened in a given study for a given set of units, is not uncommon in political science.

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 citing Fisher on elaborate theories]. 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 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

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

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.

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).

6.2 Extensions and Elaborations

Past work shows how these basic ideas can be extended: to instrumental variables (Imbens and Rosenbaum, 2005b), 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.

6.3 Benefits of Randomization Inference Worth Re-emphasizing

When datasets are small or information is weak otherwise (for example, Imbens and Rosenbaum (2005a)'s analysis of the 500,000 person study with a very weak instrument) the Fisher-style randomization inference is particularly useful because it need not require large sample approximations.

When outcomes are very skewed or make central limit theorem claims hard to justify, again, the fact that Fisher-style randomization inference makes no claims about the distributions of outcomes, and need not require averages to summarize effects, recommends these techniques.

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 be preferred over, say, multi-stage, multi-level Bayesian models in which many more moving parts may be criticized.

Binary outcomes are a situation in which information can be surprisingly low [cites to Harrell etc.. on this], and the links from the potential outcomes model of causal effects and extant binary outcome models can often appear strained when we ask about what they imply for individual units. Thus, the attributable effects framework is particular useful Hansen and Bowers (2009); Rosenbaum (2002a, 2001b, 2002b).

Randomization-based inference focuses our attention on articulating very specific hypotheses about what actually happened in a given study. Confidence intervals and point estimates are summaries of assessments of families of such hypotheses. And such hypotheses ought to be chosen based on scientific knowledge to represent scientific interest. Common assumptions often invoked to justify causal inferences (such as constant treatment effects or SUTVA) can be considered exclusions of classes of hypotheses — hypotheses involving interference or non-constant effects may be of less scientific interest (inviting confusion, for example, in the case of general hypotheses involving interference) — and less as restrictions to the scientific enterprise.

Since we often know which variables are important for our outcome, but we may not know the 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.

Randomization inference allows you to use scientific knowledge without having to exaggerate what you know.

Randomization inference is not limited to simple analyses of two-by-two tables anymore.

7 Conclusion

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)

Experimenters are lucky. In a sense they do have "all" of the data, or at least the fact of randomization allows them to make inferences to their entire experimental pool without worrying about the questions Mason raises. Although this paper has been about randomization inference for randomized studies, we also think that there are many circumstances in the social sciences in which observational studies might be better analyzed using fake-randomization-based inference (i.e. fake because no randomization occurred) than analyzed using fake-population-based or fake-model-based inference (i.e. fake because no one knows how the data on hand were sampled from any presumed super-population or because the grounds on which the data generating process model stand are shaky). Of course, such a speculation ought to require another paper on that very topic: we'll be writing one soon.

References

Belson, W. (1956), "A technique for studying the effects of a television broadcast," *Applied Statistics*, 195–202.

Bowers, J. (2011), "Making Effects Manifest in Randomized Experiments," *Cambridge Handbook of Experimental Political Science*.

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

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

Cox, D., van Zwet, W., Bithell, J., Barndorff-Nielsen, O., and Keuls, M. (1977), "The Role of Significance Tests [with Discussion and Reply]," *Scandinavian Journal of Statistics*, 4, 49–70.

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

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

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

Freedman, D., Pisani, R., and Purves, R. (2007), *Statistics*, New York: W.W. Norton, 4th ed.

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

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, 1360–1383.

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

— (2009), "Attributing Effects to A Cluster Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association*, 104, 873—885.

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

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

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

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.

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

Imbens, G. W. and Rosenbaum, P. R. (2005b), "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, 109+.

Keele, L., McConnaughy, C., and White, I. (2008), "Statistical Inference For Experiments," Unpublished manuscript.

— (2010), "Adjusting Experimental Data: Models versus Design," Unpublished manuscript.

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

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

— (2005), *Sweave User Manual*.

Mason, W. (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. (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.

Peters, C. (1941), "A method of matching groups for experiment with no loss of population," *The Journal of Educational Research*, 606–612.

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

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

Rosenbaum, P. (1999), "Choice as an Alternative to Control in Observational Studies (with discussion)," *Statistical Science*, 14, 259–304.

— (2002a), "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association*, 97, 183–192.

— (2002b), *Observational Studies*, Springer-Verlag, 2nd ed.

— (2007), "Interference Between Units in Randomized Experiments," *Journal of the American Statistical Association*, 102, 191–200.

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

— (2001a), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies with a Discrete Pivot," *Biometrika*, 88, 219–231.

— (2001b), "Effects Attributable to Treatment: Inference in experiments and observational studies with a discrete pivot," *Biometrika*, 88, 219–231.

— (2002c), "Attributing effects to treatment in matched observational studies," *Journal of the American Statistical Association*, 97, 183–192.

— (2002d), "Covariance adjustment in randomized experiments and observational studies," *Statistical Science*, 17, 286–327.

— (2002e), *Observational Studies*, Springer.

— (2002f), "Rejoinder," *Statistical Science*, 17, 321–327.

— (2010), "Design of Observational Studies," Springer, forthcoming.

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

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

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, 271–279.

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