

Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity*

Luke Keele[†]

Corrine McConnaughy[‡]

Ismail White[§]

September 20, 2010

Abstract

Experiments have become an increasingly common tool for political science researchers over the last decade, particularly laboratory experiments performed on small convenience samples. We argue that the standard normal theory statistical paradigm used in political science is ill-suited to the needs of these experimenters and outline an alternative approach to statistical inference based on randomization of the treatment. The randomization inference approach not only provides direct estimation of the experimenter's quantity of interest—the certainty of the causal inference about the observed units—but also helps to deal with other challenges of small samples. We offer an introduction to randomization inference, outline some existing randomization tests, and develop a randomization test for two-way factorial designs as an alternative to the commonly used two-way ANOVA model. Finally, we reanalyze data from two political science experiments using randomization tests to illustrate the inferential errors that can be made when classical tests are used with data from the lab.

* Authors are in alphabetical order. For helpful comments and discussion, we thank Jake Bowers, Sanford Gordon, Kosuke Imai, Cindy Kam, Walter Mebane, David Nickerson, Jasjeet Sekhon, Nicholas Valentino and Lynn Vavrek. We also thank James Fowler and Cindy Kam for generously sharing their data. A previous version of this paper was presented at the 2008 Annual Meeting of the Midwest Political Science Association, and the 2008 Annual Meeting of the American Political Science Association, Boston, MA.

[†]Associate Professor, Department of Political Science, 2137 Derby Hall, Ohio State University, Columbus, OH 43210
Phone: 614-247-4256, Email: keele.4@polisci.osu.edu

[‡]Assistant Professor, Department of Political Science, 2018 Derby Hall, Ohio State University, Columbus, OH 43210
Phone: 614-292-9658, Email: mconnaughy.3@polisci.osu.edu

[§]Assistant Professor, Department of Political Science, 2008 Derby Hall, Ohio State University, Columbus, OH 43210
Phone: 614-292-4478, Email: white.697@osu.edu

Experimentation has been a growth industry in political science over the last decade or so. This growing presence of experimentation in the methodological toolkit of the political scientist reflects an interest in making valid causal inferences about political phenomena: properly designed randomized experiments are understood to rule out confounders that threaten the internal validity of such inferences.¹ Randomly assigning subjects to either receive or not receive a treatment that represents a causal factor of interest enables researchers to employ the assumption that they have two groups that are equivalent, with the exception of the groups' reception of the treatment. Thus, any observed differences across the groups in the outcome of interest are validly ascribed to the treatment - but not without some uncertainty.

Uncertainty enters the experimenter's analysis, first and foremost, not because of the nature of the sample used to estimate a treatment effect, but rather from the realization that randomization is a tool that delivers equivalence in expectation. There remains a chance in any one experiment that the treatment and control groups were different on the observed outcome variable before the treatment was applied. It is this sort of uncertainty that the experimenter would like to estimate and report upon: how certain is it that the observed difference on the outcome variable is due to the treatment, and not to chance. This type of uncertainty is, of course, about the internal validity of the causal inference made from the experiment. While the very choice to employ the experimental approach, to randomly assign a potential cause, signals a heightened concern for confidence about causality, the experimenter is not uninterested in questions of external validity. In the interest of sufficient control for strong causal inference, however, the experimenter often forgoes random sampling as a tool for generating estimates of external validity. To make generalizations of the findings to other populations and settings the experimenter is reliant upon his or her understanding of and confidence about the causal process revealed in the experiment, rather than turning to a sampling mechanism for an estimate of uncertainty about the findings' generalization to a particular population (Brewer 2000; Imbens 2009; Morton and Williams n.d.).

Unfortunately, while experimentation has grown in political science, that growth has not corresponded with attention to the need for statistical inference tools matched to the political science experimenter's purposes. We argue that the standard tools of statistical inference employed in political science are often ill-equipped to estimate the political science experimenter's quantity of interest, especially in the realm of the laboratory experiment. For many experimenters, particularly those working in lab settings, we argue for the utility of an alternative set of inferential tools—randomization tests. Randomization tests use the random assignment built into the design of the experiment as the

¹Proper design, of course, includes elements other than randomization, such as isolation of the quantity of interest in the manipulated treatment and control over confounders not remedied by randomization, such as maturation or reactivity, perhaps through the use of a proper placebo in a control group, and proper measurement.

statistical basis for inference, offering direct and exact estimates of uncertainty about internal validity, about whether an observed difference in outcomes can be explained away by the random assignment to treatment and control groups. The randomization inference approach thus enables the experimenter to proceed without assumptions such as random sampling or normal distributions. These tests, we show, fulfill the needs of the experimenter not only because they provide direct estimation of the quantity of interest, but also because they can address a number of common analytic challenges. Difficulties of small samples and skewed distributions are reduced through flexibility in choice of statistic, including the availability of rank-based statistics, and increased efficiency, enabling more effective and confident comparison of treatment and control groups, particularly in lab applications or field experiments with smaller samples. Finally, understanding of the randomization inference framework does not necessarily entail the use of new statistical techniques. Indeed, a number of randomization tests can be approximated by standard (normal theory) tests when sample sizes are large enough. Yet, a firm understanding of the randomization inference approach clearly delineates which standard tests are appropriate and when.

To be clear about current practice and inferential needs in political science, we first review the recent experimental literature. To convey the intuition of the randomization inference framework, including the use of ranks, we then walk the reader through a simple example before moving on to the technical details of the approach. We cover not only the technicalities of simple hypothesis testing, but also the estimation of treatment effects and confidence intervals. Given that the range of randomization inference tests is far too large to cover in one paper, we select and present three tests that we identify as particularly relevant to current experimentation in political science. We then illustrate the use and implications of randomization inference tests in data from two experiments conducted by political scientists, including the differences between the inferences that would be made from these data using the randomization inference tests rather than standard normal theory tests. Finally, we conclude with some thoughts about what the experimenter ought to consider as s/he approaches the analysis of her data and where s/he can head for extensions of the randomization approach we have outlined.

1 Inference in Experimental Political Science

As experimentation has grown in political science, experiments conducted in laboratory settings, rather than in the field or “real world” settings, have been the most common type of experimentation in political science (Green and Gerber 2002). Lab environments have held particular appeal for tests of positive formal theories and psychological theories of political behavior, with scholars finding the lab most amenable to implementing representations of the abstract concepts from those theories that may be unmeasurable in other contexts (Brewer (2000), p.6, Kinder and Palfrey (1993), p.17). The lab also

offers unparalleled control to rule out confounding factors and ensure compliance with the treatment of interest (Falk and Heckman 2009). As causal theory testing vehicles, lab experiments reflect a concentration on certainty about a treatment causing a difference in outcomes across treated and control groups among the observed units—sometimes termed the local average treatment effect (LATE) (Imbens and Angrist 1994). Although this quantity is the same as the sample average treatment effect (SATE), the LATE terminology is used to emphasize that the set of units under study is not, in fact, being treated as a “sample” in the sense that the purpose is not to use these units to offer an estimate of a population average treatment effect (PATE). Again, this focus on confidence about the LATE, in turn, implies that lab experiments are rarely if ever used in conjunction with large, representative samples of national or regional populations. At the most basic level, even if such samples were desired, the costs of bringing them into the lab are typically prohibitive. Moreover, simply given the abstraction and isolation of the lab, the LATE may not be a quantity straightforwardly linked to the substantively meaningful quantity of interest about some larger population. The effect on approval of a fictional candidate due to a carefully crafted racial prime in a manipulated piece of political communication to which undistracted subjects are exposed, for example, is not meant to be an estimate of how an electorate would respond to a racial prime in a campaign. Rather, it is meant as one test—likely of many—of a theory that explains how racial primes in campaigns can and cannot affect electoral choices. Hence, the predominant samples for lab experiments in political science are both small and non-random, typically termed convenience samples.²

With a concentration on the LATE, often with a small convenience sample on hand, turning to the standard tools of statistical inference used in political science can be an uncomfortable move for the experimentalist. The classical statistical inference tests typically applied to observational data, of course, rely upon an assumption of random sampling. In political science applications where the central inferential task is connecting a sample to a specified population, this makes sense. In the experimental context, it very often does not. To be sure, there are several ways that one can recast convenience samples to provide a basis for classical statistical inference. The most coherent story is to treat the convenience sample as a population and assume that random assignment of treatment forms a sampling mechanism for that population (Hansen and Bowers 2008). One can also assume that the convenience sample is a random sample from some unknown or hypothetical population (Lehmann and Romano 2005). Yet, while classical inference can be applied to experiments, even with convenience samples, there are reasons to seek an alternative. First and most simply, the conceptual fit is awkward as the sampling metaphor must be maintained. Reporting a p -value as an estimate of the uncertainty about whether we would observe a difference value in the sample that large or larger if there were no

²For discussions of the use of convenience samples in social and behavioral science experiments, see Sears (1986), Brewer (2000), and Kam et al. (2008).

treatment difference in the population was not the original aim of the experimenter. Second, using classical inference relies on a willingness to make parametric assumptions. That is, one must assume that the test statistic follows some parametric distribution such as the t or Normal. That this sort of assumption is often questionable in small samples gives the experimentalist another reason to desire an analytic alternative. Finally, the randomization inference approach provides a number of tests that deal well with extreme values or skewed distributions, common challenges especially to experimenters working with small samples.

Despite the number of ways in which standard inferential tools seem inappropriate to the political science experimenter's analytic goals and constraints, we find that their use is nonetheless commonplace, suggesting that alternatives are not widely known or accessible. To assess the inferential tools being used by experimenters in political science, we used JSTOR to search for all articles that contained the word "experiment" published between 1995 and 2007 in the *American Political Science Review*, the *American Journal of Political Science*, and the *Journal of Politics*. This search, which allowed us to cast a wide net in finding experimental studies and examining the sampling strategies and statistical analyses employed therein, returned 258 articles. We found that normal-theory based analyses are standard procedure, even among experimental studies conducted upon convenience samples. For simple comparisons of treatment and control groups, we found experimental political scientists regularly use the difference in means t -test and single factor (or one-way) analysis of variance (ANOVA) (Sigelman et al. 1995; Cobb and Kuklinski 1997; Valentino, Hutchings, and White 2002; Gibson 2002; Druckman and Nelson 2003; Hibbing and Alford 2004). Of the experimental research conducted on convenience samples turned up in our search, about one-third of the studies used either t -tests, ANOVA, or both. Even more common than the t -test and one-way ANOVA, however, are multivariate techniques such as multi-factorial (two-way) ANOVA, analysis of covariance (ANCOVA), and multiple regression analyses. Nearly 70 percent of the studies performed upon convenience samples that we examined used some form of these techniques. These multivariate approaches are commonly employed to test hypotheses about 2 and 3-way interactions, often driven by theories that specify factors that should moderate treatment effects. Specified interactions include interactions of multiple treatment variables when experimenters employ multi-factorial designs (see for example Golebiowska (1996); Mitchell et al. (2003)) and interactions of experimental conditions or treatments with some other variable of interest, usually presumed to remain fixed across experimental conditions (see for example Nelson, Clawson, and Oxley (1997); Gilliam and Iyengar (2000); Miller and Krosnick (2000); Druckman (2001); Valentino, Hutchings, and White (2002); Brader (2005)). Multivariate regression, multi-factorial ANOVA and ANCOVA have also been used as ways of adjusting for imbalance of treatment and control groups (despite randomization). In models assessing differences across conditions,

experimentalists often attempt to account for the imbalance of relevant covariates by entering these variables as controls (for example see Nelson, Clawson, and Oxley (1997); Valentino, Hutchings, and White (2002); Brader (2005)).

Our search also helped to confirm just how commonplace the sort of samples are that present the greatest questions about reliance on normal theory tests. Twenty percent of the 258 articles turned up in our search proved to be experimental studies that used convenience samples of either students or non-student adults. Since not all of the studies captured by our search were true experiments, however, this percentage understates the ratio of convenience to random samples in experimental research. Kam, Wilking, and Zachmeister (2008) found 25% of the experiments in these journals use student samples. This number again understates the total number of convenience samples since they group any form of adult convenience sample with random samples of adults. A survey of the experimental studies published in the *American Political Science Review* conducted by Druckman et al. (2006) found a total of 57 studies, 63% of which were conducted in a laboratory or classroom, where convenience samples are almost always used. Similarly, a search of the journal *Political Psychology* during the 1995-2007 time frame turned up 89 articles employing experimental research, 48 percent of which relied upon convenience samples. Kam, Wilking, and Zachmeister (2008) found that 70% of the experiments in *Political Psychology* and *Political Behavior* used student samples. In sum, the conduct of experiments with convenience samples is widespread in political science.

Given the state of experimentation in the discipline, we advocate for an alternative set of inferential tools called randomization tests—design-based tests that use the experiment’s random assignment of the treatment as the basis for statistical inference. The range of test statistics available in the randomization inference framework is large and varied, easily meeting the analytic needs of the experimenter in terms of the comparisons s/he might wish to make. The resultant p -values from these tests are exact estimates of the probability that the treatment effect found in the sample at hand could be ascribed to the stochastic process that delineated the treatment groups. Thus for direct estimation of the experimenter’s quantity of interest, all experimenters should find the randomization inference framework conceptually appealing. Those working with small and/or non-random samples face constraints that make these tests even more desirable. First, because the tests we outline are nonparametric, they do not require the analyst to make the perhaps questionable assumption that the test statistic follows a parametric distribution. While the particular assumptions of each specific randomization test vary, the key—and obviously justified—assumption of random assignment of the treatment is typically joined only by an assumption that the underlying distribution has some basic property, such as being continuous (Hollander and Wolfe 1999, pg. 106).³ Although some of these randomization tests can be

³There is also a parametric version of design based inferences. We discuss this briefly below.

asymptotically approximated by standard (normal theory) tests, small samples are particularly vulnerable to inferential errors due to the asymptotic approximations. Moreover, many normal theory tests are not approximations of any randomization inference tests. Hence, an understanding of the randomization basis for statistical tests can help experimenters to avoid choosing tests that would invoke assumptions unfounded in their circumstances. Finally, the randomization inference framework allows for the use of rank-based statistics in testing differences across treatment and control groups. The appeal of these tests is their utility in testing differences when the outcome of interest is prone to extreme values—a particularly vexing problem when the number of observations is small.

These tools are certainly not new. In fact, they date back to the basic theory of randomization in experiments as formulated by Fisher (1935) in the first half of the twentieth century. They have been largely absent, however, from experimental methodology in political science (Hansen and Bowers 2008, 2009; Sekhon and Titiunik 2008; Fowler and Kam 2007). The most significant barrier to use in the past has been insufficient computing power. That constraint now rarely applies, and the tests are finding their way into many standard software programs. That our broad search of the main disciplinary journals of political science, however, found no studies that employed an explicit randomization test suggests that these tests have not yet come to the attention of the field. We expect that practitioners offered an informed choice between standard, normal theory approaches and the randomization inference approach will find the latter useful, particularly in the small convenience sample environment often encountered in the lab.⁴

Finally, we should note that randomization inference does not in any way solve the statistical problem of how to make inferences about populations with nonrandom samples. While randomization tests may provide better estimates of statistical uncertainty about internal validity, they do not provide any statistical estimates of external validity. Of course, the standard tests currently used don't provide statistical leverage over external validity either. They simply approximate the randomization tests that we outline below. Questions about external validity can only be addressed statistically with an appropriate sampling mechanism.

2 An Introduction to Exact Randomization Inference

Before providing a formal description of randomization tests, we start with an example to illustrate the logic and implementation of the randomization inference approach. The example introduces both the

⁴This is not to say that randomization tests are totally unknown to political scientists. Both Hansen and Bowers (2008) and Hansen and Bowers (2009) investigate and advocate the utility of randomization tests using political science applications. Ho and Imai (2006) provide another example with a political application. These all appear in statistics journals, which only seems to underscore the rarity of these tests in political science. We know of one article published in the main disciplinary journals that uses randomization tests to analyze an experiment: Fowler and Kam (2007), an article that fell just outside the time bounds of our initial JSTOR search.

idea of randomization as the source of uncertainty for statistical inference and the use of ranks as the basis for comparison of treated and control groups. While the use of ranks is not necessary to employ randomization inference, it is a type of test possible within the randomization inference framework that should hold particular appeal for experimental applications where the outcome of interest is skewed and/or prone to outliers.

Consider an experiment where we select seven students to play a dictator game with a computer program which asks them to allocate 100 dollars between themselves and a charitable donation.⁵ Three of the students are randomly chosen to receive the treatment, a prime expected to make them more altruistic. If the treatment is effective, we would expect that the students who receive it would give away more of their money. If there is no effect of the treatment, we would expect no such difference across the two sets of students. In other words, we are interested in testing the following null hypothesis:

H_0 : The prime has no effect on amount given.

against our alternative:

H_a : The prime has a positive effect on amount given.

Relying on the standard toolkit of the political scientist, we might translate these expectations into a t -test comparing the mean difference in dollars given we observed to a null of no mean difference. We would then calculate a p -value using the critical value from a t -distribution with five degrees of freedom. For this inference, we must assume the test statistic follows a t -distribution.

As an alternative, we derive the test from the randomization of treatment assignment and avoid the parametric assumption. Our expectation, again, is that treated students should give away more than non-treated students. We know that the best evidence that the prime increases altruistic behavior would be if the three students who received the treatment rank 1, 2, and 3 in terms of the amount given away. So we might ask: what the probability is that the three students in the treatment group would happen to rank 1, 2, and 3 in terms of the amount given away if the null hypothesis were true, and, in fact, there was no effect of the prime? To develop a probability statement about this hypothesis that hinges on the composition of the treatment and control groups - the feature that was assigned by a random process - we turn to basic combinatorics. Knowing that the number of ways of selecting r objects from a set of n is $n!/[r!(n-r)!]$ tells us that there are 35 ways to select three students from a set of seven. Table 1 contains all the possible combinations of ranks that we could observe for our set of treated subjects. If the treatment had no effect, and simple chance were the only factor governing which students were in the control group and which were in the treatment group, then each of these 35 combinations would be equally likely. Seeing in the table that our best evidence outcome is 1 of

⁵This example is adapted from one in Sprent and Smeeton (2007).

35 possible treatment group compositions tells us that there is a $1/35 \approx 0.0286$ chance that we would observe the best evidence case if the null were true. That is, if the treatment has no effect, the chance that random assignment will produce this exact outcome is $1/35$. This p -value indicates that the best evidence outcome indeed enables us to be fairly confident that the prime has an effect on student behavior in our divide the dollar game.⁶

Table 1: Possible Combinations of Ranks in Treatment Group

1,2,3	1,2,4	1,2,5	1,2,6	1,2,7
1,3,4	1,3,5	1,3,6	1,3,7	1,4,5
1,4,6,	1,4,7	1,5,6	1,5,7	1,6,7
2,3,4	2,3,5	2,3,6	2,3,7	2,4,5
2,4,6	2,4,7	2,5,6	2,5,7	2,6,7
3,4,5	3,4,6	3,4,7	3,5,6	3,5,7
3,6,7	4,5,6	4,5,7	4,6,7	5,6,7

We can make this approach to hypothesis testing more general by introducing a summary statistic that enables us to translate ranks into a single measurement of the outcome among the treatment subjects. One possible statistic for this purpose is the sum of the ranks of the treated subjects. This statistic will be lower if the treated subjects are generally higher in their giving than the control subjects, and higher if they are not.⁷ Using this statistic we can answer the question of what the chance is of observing an outcome of a specific degree (or smaller/larger) among the treatment subjects if the treatment actually has no effect. For example, suppose the outcome we observed among the treated subjects was the ranks 1, 2, and 7. Our summary statistic would be 10 ($1 + 2 + 7 = 10$). This seems close to the “best evidence” outcome we just considered, where the sum of the ranks would be $1 + 2 + 3 = 6$, but is it close enough to be convincing evidence of a treatment effect? To answer this question, we work out the probability of observing a rank sum of the same amount or less than the one we did under the null hypothesis that the treatment had no effect on giving by returning to our enumeration of all 35 possible combinations of the three ranks and calculating the rank sum for each combination. Four of the 35 possible rank combinations produce a sum of 10, 3 more produce a sum of 9, 2 result in a sum of 8, 1 set sums to 7, and another to 6. Thus, if the treatment had no effect, the chance of observing an outcome like the one we did or smaller would be $p = 11/35 \approx 0.314$. Put another way, if the prime has no effect, we could expect to see a value for the summed ranks as low as or lower than the one we observed 31 out of every one hundred times we randomly assigned the

⁶If we had a two-sided alternative hypothesis, we would double the resultant p -value. However, there are a few other methods for calculating two-sided p -values. See Lehmann (1975) for details.

⁷Note the inverse relationship is due to the direction of our ranking - that those who gave away more were given lower values for their ranks (our highest giver being ranked 1, etc.). The subjects could just as easily be ranked in the opposite manner, and then higher rank sums would be associated with higher amounts given. Having directionality - and attending to it - is what is important here.

treatment to these particular subjects. Using the traditional hypothesis testing threshold of .05, the p -value we calculated would not allow us to reject the null hypothesis; the observed outcome did not provide sufficient evidence that the prime had an effect on our subjects' behavior.

Note that the interpretation of the p -values in this example reference exactly the information we believe the political science experimenter is after: the probability that the result observed among his/her specific set of experimental subjects can be explained away by the chance constitution of the treatment groups. We can calculate such an exact p -value for any combination of values that we might observe in the experiment without ever using a parametric distribution. Despite this fact, we are still able to make statistical inferences because we can use the randomization of the treatment to create a meaningful probability distribution for the null hypothesis.

2.1 Randomization Inference for No Effect

Here, we provide a more formal outline of using randomization inference to test that the experiment is completely without effect. Of course, calculating the p -value for rejecting the null of no treatment effect is only one statistical quantity of interest; later we will turn to confidence intervals and point estimates. However, the test for no effect deserves special attention since it is a very rare thing: a statistical test with no assumptions. That is, if the randomization has been successfully implemented and the researcher has no concerns about noncompliance with the treatment, the test of no effect can be done without any assumptions—beyond the one that randomization was, indeed, used (Rosenbaum 2002b).

A statistical test built on the randomization inference framework has the same elements as any statistical test: data, a null hypothesis, a test statistic, and a distribution of the test statistic under the null hypothesis. It is the derivation of the last element, the null distribution, that is unique to the randomization test. To provide a formal derivation of the null distribution for randomization tests we first define \mathbf{T} as a random vector that assigns subjects to either the treatment or control group. This treatment vector provides the basis for the assumption of exchangeability: each unit has equal probability of being assigned to the treatment and control groups. This will be true if the values for \mathbf{T} are chosen according to a randomization mechanism like a coin flip. To illustrate, if the first, third, and fourth subjects out of seven subjects are selected to receive the treatment, \mathbf{T} would have the following form: (1,0,1,1,0,0,0).

Next, we denote the quantity \mathbf{y} as a vector of the outcomes for the subjects, and we define the test statistic as

$$S = f(\mathbf{y}, \mathbf{T}) \tag{1}$$

That is, the test statistic, S , is the result of a function, f , that operates on both the outcomes and the

treatment assignment. In the example in the previous section, S took the form of the summed ranks for the treatment group. The function f and the test statistic S can take several possible forms; we discuss some of those possibilities in greater detail in the following sections.

We further denote all possible responses under the treatment as Ω with elements s_i . This Ω contains all outcomes under all possible realizations of \mathbf{T} . In our example, summing the elements in Table 1 forms Ω for the experiment. We use Ω to calculate the probability that the observed value of S is as large or larger (or as small or smaller, depending on the direction of our expectations) than what we would observe if the null hypothesis were true. This p -value is the sum of the randomization probabilities that lead to the referenced values of S , relative to all possible values of S :

$$Pr(S \geq s_i) = \frac{\sum I(S \geq s_i)}{|\Omega|} \quad (2)$$

where $I(\cdot)$ is an indicator function, and $|\cdot|$ denotes the cardinality of a set.

The implications of this inferential mechanism bear repeating. Probability enters our calculations only through randomization of the treatment, and does not rely on any parametric probability distribution or sampling mechanism. That is, the inference is entirely design-based—assuming only random selection of units or random allocation of units to experimental conditions—and invokes no modeling assumptions external to the study design. Here, the null distribution is completely known. Moreover, the p -values directly translate into the experimenters’ original quantity of interest: the probability that what they observe as evidence of a treatment effect can be explained away by the chance process that assigned their observed subjects into treatment and control groups. The p -value is exact, which implies that our test will provide exact coverage probabilities for confidence intervals without relying on assumptions that the data are from any parametric distribution (Hollander and Wolfe 1999).

The randomization inference paradigm we have outlined in this section was developed by Fisher (1935). It is designed to allow analysts to make inferences about sample treatment effects through tests of a *sharp* null hypothesis. Under the sharp null hypothesis, we test whether the treatment effect is zero for all units. This approach is different than the randomization inference approach developed by Neyman (1923), which is designed to test hypotheses about the average treatment effect (ATE), and the familiar null hypothesis is that the ATE is zero. While the Neyman approach may be of interest in some applications, there are several compelling reasons to adopt the Fisher approach, particularly in the typical political science experiment performed upon a convenience sample. First, the Fisher approach enables the analyst to proceed without distributional assumptions, while the Neyman-style tests of SATE require assuming that the sample is large enough for normal based test statistics.⁸ Second, the sharp null hypothesis is simply a strict test of the presence of a treatment

⁸In brief, the Neyman approach involves first using the mean difference serves as an unbiased estimator of the causal

effect; it eliminates the possibility of finding in favor of a treatment effect when evidence of that effect is actually mixed. Yet, it does not preclude the researcher from specifying moderators of the effect and testing for effects differentially according to values of those moderators. Moreover, recent work on Fisher style randomization inference allows an analyst to diagnosis the extent of nonconstant treatment effects for some randomization tests (Rosenbaum 2003). These techniques allow analysts to understand just how plausible the sharp null is. In what follows, then, when we refer to randomization inference or tests, we mean Fisher’s version of these tests. As we will discuss later, however, in large samples the distinction between the two types of hypotheses vanishes.

Promising as the randomization inference method seems, there is one caveat, albeit a practical one. For large samples, the number of possible outcomes can be quite large and, even with modern computing, the time required to compute an exact p -value can be lengthy. For such situations, we can simulate the distribution of null outcomes and derive approximate exact p -values. Simulation demonstrates that these simulated tests very closely approximate the exact tests. In fact, in examples where the entire null distribution can be elaborated, approximate tests based on simulation produce accurate inferences when the number of simulations is considerably smaller than the total number of permutations.⁹

2.2 Asymptotic Approximations to Randomization Tests

Some standard parametric tests are asymptotic approximations to particular randomization tests. Fisher (1935, chap. 21) hypothesized that the t -test could be derived from the permutations of randomized treatments and that a t -test based on permutations and the t -distribution should be similar. Hoeffding (1952) later proved that asymptotically the two tests are equivalent, justifying Fisher’s belief that the usual t -test can be viewed as an approximation to the distribution-free exact version of the test. The results in Hoeffding (1952) also imply that parametric tests based on the F-distribution in one- and two-way ANOVA models provide an approximation to randomization based tests for multiple treatments. The convergence of the exact distribution and parametric distributions in large samples also implies that the distinction between tests for average treatment effects and tests of sharp null hypotheses becomes less important as the sample size grows.

The accuracy of the approximation for any given application is, of course, questionable, and the asymptotic approximation provided by the parametric test will be less reliable than that for the

estimand. The analyst next finds an unbiased or upwardly biased estimator for the variance of the average causal effect. An appeal to the central limit theorem allows the analyst to form a confidence interval for the average causal effect based on these estimated quantities. Tests of average treatment effects come in two versions: tests of hypotheses about the population average treatment effect (PATE) and tests about the sample average treatment effect (SATE). The tests differ in the form of the estimator for the variance.

⁹There are several exceptions. Fisher’s exact test for a binary outcome and a binary treatment has a closed form solution, as the permutation distribution follows a hypergeometric distribution. Other tests with closed form solutions are the Mantel and Haenszel test and the d statistic of Hansen and Bowers (2009).

randomization test, itself. As we have noted, political science experiments often involve small numbers of subjects, and in these contexts asymptotic approximations provide little comfort. Note, in particular, that the asymptotic approximation is based not on the overall sample size, but the group size for each treatment (i.e., the cell size) (Lehmann 1975). General guidelines on when the exact and the asymptotic approximation converge are difficult since it depends on the test. Thus, while use of standard tests as approximations of their randomization inference counterparts can be appropriate, we illustrate with the empirical examples that follow how different the results from a randomization test and the asymptotic approximation can be.

Reliance on asymptotic approximations poses another danger. While there is work justifying asymptotic approximations for randomization tests, simple alterations of the analysis can render these approximations incorrect. For example, Freedman (2008*b*) demonstrates that a bivariate regression model provides inferences about treatment effects identical to those from the t distribution, which is justified as an approximation of a randomization inference test. One might suspect that moving to a multiple regression model produces a similar approximation. Freedman proves, however, that this is not the case. The multiple regression model provides biased estimates of the treatment effect in small samples and fails to correctly estimate the variance in any sample size (Freedman 2008*b,a*). So what might seem like a simple step of adding regressors results in biased estimates and misleading inferences. For binary data, a similar problem occurs. For binary data, an exact test known as Fisher’s exact test is available. Alternatively, one can test for a difference in proportions and rely on a Normal approximation to the exact test. One might assume by extension that a logistic regression model is another method one could use to test for treatment effects. Freedman (2008*c*) demonstrates that treatment effect estimates from logistic regression models are inconsistent. Hansen and Bowers (2008) provide an example of this inconsistency when logistic regression is applied to a voter turnout field experiment. Thus, we strongly emphasize that care must be taken with asymptotic approximations.

2.3 Interval Estimation

Thus far, we have discussed only the calculation of exact p -values for tests of sharp null hypotheses. Of course, many experimenters will be interested in more than simply evaluating the null hypothesis based on a p -value. Notably, they may wish to say something about the range of plausible treatment effect sizes given the evidence in their data. That is, experimenters are likely to also be interested in a point estimate for the treatment effect and a confidence interval for that point estimate. Both are possible within the randomization inference framework, though they will require assumptions beyond random assignment. Still, the assumptions required are transparent, and the experimenter may feel quite comfortable in making them.

The underlying logic of estimating confidence intervals under the randomization inference paradigm

is no different than for the standard parametric method for interval estimation. Confidence intervals are still simply ranges of values that would not be rejected as tested hypotheses under a specified α level. Since confidence intervals are, in fact, a series of hypothesis tests, one way to obtain a $100(1-\alpha)\%$ confidence interval is to “invert” a series of $100(1-\alpha)\%$ level tests (Walsh 1996). This is true regardless of whether one relies on a parametric distribution or not when forming confidence intervals. Under the randomization inference paradigm, we will invert the test of a sharp null hypothesis to form a confidence interval. In short, we test a series of sharp null hypotheses, and we invert each test and use the calculated p -values to find the values at the endpoints of the $100(1-\alpha)\%$ confidence interval.

Estimating confidence intervals, however, does require some additional notation and assumptions that were not required for calculation of the exact p -value for the test of no effect. First, for each unit i observed in the experiment, the indicator $T_i = t$, $t \in \{0, 1\}$, records the randomly assigned treatment status. Each unit also has potential outcomes or responses—values that could have been observed for that unit—which we denote as y_i . Each unit has a potential outcome under treatment and a potential outcome when not treated, though we only observe one or the other. We thus write the responses as y_{it} such that $y_{it} = y_{i1}$ when $T = 1$ and $y_{it} = y_{i0}$ when $T = 0$. Combining potential responses with the treatment indicator, we can define the observed outcome for each unit i :

$$Y_i = T_i y_{i1} + (1 - T_i) y_{i0}. \quad (3)$$

Two additional assumptions are needed for interval estimation. First, we must assume that the outcome of each unit is unaffected by the treatment status of other units (Cox 1958). Analysts often refer to this as the “stable unit treatment value assumption” (SUTVA)(Rubin 1996). Typically, the credibility of this assumption must be evaluated in light of the design and implementation of the experiment. If an experiment involves subjects privately interacting with a computer within a lab setting, for example, we are likely to be quite confident that one subject’s receipt of the treatment does not affect the observed outcome of any other subject. Next, the analyst must make an assumption about the nature of responses to the treatment. Rosenbaum (2002*b*) refers to this assumption as a model for the effects. Rosenbaum (2002*b*, ch. 5) outlines a number of different effects model, but the most widely used model of effects is that of a constant-additive effect, which is the sort of effect implied by a linear model. That is, we assume that the treatment raises the response of each unit by a constant amount: $y_{i1} = y_{i0} + \tau$, where τ will serve as a point estimate for the treatment effect.¹⁰

Under this model of effects, we can rewrite the outcomes as $Y_i = y_{i0} + \tau T_i$. Notice that the adjusted

¹⁰The model of the effects may or may not correspond to the test statistic. For example, the constant-additive model of effects may be used with both the average difference of means test statistic as well as the the rank sum test statistic. The constant-additive model of effects has a natural correspondence with the first test statistic but not necessarily with the second. However, this model of effects is valid for either test statistic.

responses: $Y_i - \tau T_i = y_{i0}$ are fixed and do not vary with treatment assignment (Rosenbaum 2002b). We exploit this fact to invert the test for interval estimates. A $1 - \alpha$ confidence interval may be obtained by testing nonzero values of τ , using the same method we outlined in Section ???. The values of τ not rejected at level α form a $1 - \alpha \times 100\%$ confidence interval.

This method of interval estimation is best understood with a simple example. Assume we conduct an experiment testing whether the tone of a media campaign affects intention to vote. Here, intention to vote is measured with a seven-point scale, with higher values indicating a higher propensity to vote. Among twelve subjects, six are randomly exposed to a negative campaign advertisement, and we observe the following outcomes:

$$\begin{aligned} \text{Treatment Group} &= (1, 4, 5, 1, 5, 5) \\ \text{Control Group} &= (7, 7, 5, 4, 6, 5) \end{aligned}$$

Seeing no outliers in the data, we decide to use the absolute difference in means across the treated and control units as our test statistic. We first test the usual sharp null hypothesis that the treatment effect is zero. To calculate an exact p -value, again, we compare the number of times the test statistic occurs relative to the universe of test statistics computed with all possible permutations of the data. In this case, there are 924 possible ways to form a treatment group of seven subjects from a pool of fourteen. Comparing the observed test statistic to the 924 permutations, we find that the exact p -value is 0.002, and thus we would reject the sharp null hypothesis that the treatment effect is zero. To construct our confidence interval, we first assume a model of constant additive effects. We then test a series of sharp null hypotheses, testing the null hypothesis that $\tau = 1$ and then $\tau = 2$ and so on. In this case, we tested a series of null hypotheses, beginning at 0, in increments of 1. We compute adjusted responses according to our model of effects, using the equation $Y_i - \tau T_i = Y_{i0}$. In sum, we subtract the value of τ for each null hypothesis from the treatment observations and then calculate the test statistic. We then calculate the exact p -value associated with this test statistic. We form the confidence interval from the hypotheses tests for the values of τ where we do not reject at a chosen level of α . The results from this iterative process are in Table 2.

To form a 95% confidence interval for our treatment effect, we find the values of sharp null where we reject the null at $\alpha = .05$. Given the discrete nature of exact p -values, it is often the case that we can't form a precise $100(1 - \alpha)\%$ confidence interval. This is true in our example. We don't observe values right at the α value of 0.05. Rather, we observe that a null hypothesis of 1 has a p -value of 0.048 and the null of 6 has a p -value of 0.052. Thus, the smallest effect we would reject at the approximate 0.05 α level is a mean difference of 1, and the largest value we would accept at the approximate 0.05 level is a mean difference of 6, making our confidence interval [1,6].¹¹

¹¹Rosenbaum (2001) calls the series of sharp nulls that are tested attributable effects. He demonstrates how this framework

Table 2: Inverting the Null Hypothesis To Form Confidence Intervals

95 % Confidence Interval	Sharp Null Hypothesis	Exact p -value
	0	0.002
Lower Bound	1	0.048
	2	0.234
	3	0.701
	4	0.727
	5	0.251
Upper Bound	6	0.052
	7	0.004

2.4 Point Estimation

Point estimation for treatment effects, like interval estimation, depends directly on the model of effects chosen by the analyst. Here, we only discuss point estimation for models of constant additive effects. There are a number of methods for point estimation of τ . The most widely used point estimate for τ is simply the observed difference between the average response in the treated and control groups. This, of course, is exactly the amount calculated in the test statistic in Equation 4. It is well known that in a randomized experiment this is an unbiased estimate of the treatment effect. For this point estimator, we need not assume that the treatment effect is constant. In fact, the only assumption needed for unbiased estimation of τ is that SUTVA holds. The reader may have noticed that in this case, there is a direct correspondence between the test statistic and the point estimate. In other instances, however, there may be such a direct link between the test statistic and the point estimate. For rank based tests, typically, the test statistic does not have any correspondence with point estimate. The point estimator for τ used in conjunction with rank based test statistics is one developed by Hodges and Lehmann (1963). The Hodges-Lehmann point estimator for τ is the value of $\hat{\tau}$ such that the adjusted responses, $Y_i - \tau T_i$, are exactly without treatment effect. It is beyond the scope here to provide more details about the Hodges-Lehmann point estimator other than to say that it is closely related to estimators that are based on differences in medians. Moreover, point estimation of treatment effects is not the goal for many more complex experimental designs. For example, in twoway factorial designs there is not a single parameter that summarizes the treatment effect.

can be extended to any experimental design. He also proves that these attributable effects are a random variable that map the different outcomes that might have been observed in the treatment group under the sharp null hypothesis.

2.5 Confidence Intervals and Parametric Assumptions

The nonparametric confidence intervals that we outlined above have a singular property not found with the parametric confidence intervals that analysts are used to working with. To form parametric confidence intervals, the analyst must assume that the data follow a particular distribution. When the sample size is small, however, the analysts essentially adds information to the data with the parametric assumption which will result in confidence intervals that are overly narrow and fail to maintain correct coverage (Imbens and Rosenbaum 2005). This is analogous to using an informative Bayesian prior with the data. Informative priors are most likely to influence our answer when sample sizes are small (Gelman et al. 2003). One advantage of the confidence intervals from Fisher-style randomization tests is they maintain correct coverage regardless of how much data is used. The exact $100(1 - \alpha)\%$ confidence set for the treatment effect estimate will always maintain its stated coverage of $100(1 - \alpha)\%$, but when the data do not contain enough information, the interval may achieve this coverage by becoming infinite in length (Imbens and Rosenbaum 2005). That is, we may find that there are no values of the sharp null where we are able to reject at a chose confidence level. This is an attractive feature of these nonparametric confidence intervals: they reveal whether additional data are required to increase the power of the test. The randomization inference confidence interval is wide or narrow depending on the evidence available in the data, not based on assumptions about the distribution of the data. For experiments, this property provides important post-hoc information about the power of the test.

Consider an example from an experiment we conducted. In the experiment, subjects in the treatment group viewed a story about a mugging in the local newspaper. The control group was exposed to a story about changes to the iPhone from the same local newspaper. Subjects were then asked to rate whites and African Americans on racial stereotypes. The difference across the measures of racial stereotypes measures whether the treatment caused subjects to rate African Americans lower relative to whites when primed on the topic of crime. In this experiment, there were 19 subjects in the control condition and 22 subjects in the treatment condition. Table 4 contains the point estimate and confidence intervals for both a standard t -test and the rank sum test. If we proceed with a standard parametric analysis based on the t -test, the difference in mean ratings is -1.3. That is, subjects in the treatment condition rated African-Americans 1.3 points lower than whites on the racial stereotypes scales. The normal theory confidence interval for this estimate is [-2.47, -0.19]. As we will see, this confidence interval is overly narrow.

The point estimate for the treatment effect from the rank sum test is -1.5 with an exact p -value of 0.004. The confidence interval for the nonparametric estimate, however, is $[-\infty, 0]$. The confidence intervals formed from inverting the sharp null hypotheses maintain that correct coverage when there is little information in the data by rejecting few if any hypotheses. Therefore, the randomization test

Table 3: Parametric and Nonparametric Confidence Interval Comparison

	<i>t</i> -test	Exact test
Estimated Effect	-1.3	-1.5
95% Confidence Interval	[-2.47, -0.19]	$[-\infty, 0]$
N_C		19
N_T		22

reveals that there is not enough information in the data to say anything more about the treatment effect other than it is negative. To be more specific about the treatment effect would require us to invoke a parametric assumption. Thus, the precision in the parametric test is overstated unless the analyst is willing to defend the distributional assumption. In our example, we see that by assuming the data are distributed normally adds information to the data resulting in confidence intervals that are overly narrow unless the parametric assumption is correct. With the nonparametric test, we observe that the treatment effect is clearly negative, and we can reject the null that the sharp null is zero, but to learn more about the treatment effect requires a larger sample size. As such, the confidence intervals from nonparametric tests are far more revealing about the post hoc power of an experiment to detect effects. Based on these results, we conclude that in future iterations of the experiment on a larger number of subjects would be wise.

3 Statistical Tests with Randomization Inference for Political Science

As mentioned in the last section, the analyst must choose a function f that operates on treatment assignment and outcomes to produce S , a test statistic. A range of valid test statistics is available to the experimenter, often including several possible choices for evaluating the treatment effect for a particular design. Some of those choices may not, however, be very powerful. If the distribution of outcomes has outliers or heavy tails—as is often the case with, for example, reaction time data—test statistics based on averages will have little power. The decision about which statistic to employ, therefore, should take into consideration not just the nature of the design, but also the distribution of the outcome of interest.

Consider the choice of a statistic to test a difference across two experimental conditions. A familiar choice of statistic for the experimentalist is the difference in means for the treatment and control groups, a function that calculates a measure of location shift across the treatment and control distributions:

$$\Delta = \frac{1}{n_C} \sum Y_i^C - \frac{1}{n_T} \sum Y_i^T = \bar{y}^C - \bar{y}^T \tag{4}$$

where n_C is the number of units in the control group and n_T is the number of units in the treatment group. While this statistic is easily interpretable as an additive treatment effect, it will have low power in the presence of outliers, skewed distributions, or heavy-tailed distributions (Diaconis and Lehmann 2008).

A transformation before comparing average levels across treatment and control can produce other useful statistics. For example, taking natural logarithms of the outcomes and estimating as follows:

$$\Delta = \frac{1}{n_C} \sum \ln(Y_i^C) - \frac{1}{n_T} \sum \ln(Y_i^T) \quad (5)$$

produces a statistic that we can interpret as a constant multiplicative treatment effect (Imbens and Rubin 2008). The log transformation will increase the power of the test if the raw data have a skewed distribution. Ranking is another common transformation of the data. As we illustrated with our example in Section 2, the outcomes are transformed to ranks before estimating differences between the treatment and control groups. The advantage of the rank transformation is that it improves the power of the test in the presence of skewed distributions, outliers and symmetric, heavy-tailed distributions (Hollander and Wolfe 1999). As we show later, the loss of power with ranks when the data are normal is slight. In most statistical software, randomization tests are implemented as rank tests. Many other statistics are also possible. One might use the absolute difference in medians, for example, or the test statistic from the t -test to perform a permuted t -test. The availability of a particular statistic in software is the only real constraint on the choice of statistic.

In this section, we outline three randomization inference tests that we expect to be of use to political science experimenters. The list presented here is by no means exhaustive. We focus on two that most directly relate to the basic tests commonly used in the political science literature: the Wilcoxon rank sum test as an alternative to a difference of means t -test or one-way ANOVA on two groups, the Kruskal-Wallis test as an alternative to a one-way ANOVA with multiple groups. We then present permuted F-tests as an alternatives to standard two-way ANOVA analysis of multi-factorial designs.¹² Again, while we focus here on rank-based tests given their robust properties and the current infrequent use in political science, ranks are not necessary for randomization inference.

¹²Note the important distinction between nonparametric rank-based statistics and randomization inference. Statistical tests such as the Wilcoxon rank test are commonly referred to as distribution free or nonparametric. While asymptotic normal approximations are available for this test, the test does not rely on parametric distributional assumptions. Such nonparametric tests can be justified on two grounds. The first assumes that the data are a random sample from some unknown distribution; depending on the test, this unknown distribution may be assumed to be symmetric or, more weakly, continuous (Hollander and Wolfe 1999, pg. 2). Thus derived, rank tests are designed to replace parametric tests in settings where specific distributional assumptions are questionable, but the random sampling assumption holds. Nonparametric rank-based statistics, however, can also be derived quite naturally from the assumption of treatment randomization (Lehmann 1975). For the tests that we outline below, we derive the form of the statistical test assuming not random sampling but randomization of the treatment.

3.1 Comparing Treatment and Control: The Wilcoxon Rank Sum Test

For experimental designs with a single treatment group and a control group (or two alternative treatment groups and no strict control), basic assessment of the hypothesis that the treatment causes an outcome change depends on a comparison of the realizations of that outcome in the treatment and control groups. Such an assessment is typically conducted with a difference of means t -test or a one-way ANOVA. The rank-based randomization alternative is the Wilcoxon rank sum test, which is what we applied in the illustrative example in the previous section.

For the rank sum test, we let Y_1, \dots, Y_n be n treated observations and X_1, \dots, X_m be m control observations for a total of $N = n + m$ observations. Under the null hypothesis, the distributions of X and Y are identical. The alternative hypothesis asserts that the location of Y is larger or smaller than that of X . Assuming random assignment of the treatment, we wish to estimate the following:

$$\Delta = G(X) - F(Y)$$

The parameter Δ is the location shift or treatment effect, and F and G are distribution functions. The null for this location shift model is

$$H_0 : \Delta = 0$$

the hypothesis that the treatment has no effect. To test this hypothesis, we order the combined sample of X - and Y -values from the least to greatest. Let R_n be the ranks of the n Y -values in the joint ranking. Let W be the sum of the ranks of the Y -values. More precisely,

$$W = \sum_{j=1}^n R_n.$$

If W is sufficiently large enough to exceed a test statistic, c , one may reject the null hypothesis and conclude that the treatment is effective. More formally, we reject the null hypothesis if the following is true:

$$W \geq c.$$

We select the value for c such that the probability of W being greater than or equal to c is equal to α the selected level of statistical significance. For a standard test at the 0.05 level, we would select c such that:

$$P(W \geq c) = 0.05$$

where this probability is computed under the null hypothesis of no effect. Values of W greater than or equal to c are unlikely when the null is true and the probability of observing such values by chance

are equal to α . What is different from standard parametric tests is how we find values for c . For this, we must find the probability that W has a specified value under the null hypothesis.

For a parametric test, we would assume that the data are from a known parametric probability distribution and select c accordingly to the value of α . To avoid use of a parametric distributional assumption, we can rely on possible permutations under the randomized treatment to select a value for c . Let $\#(w; n, m)$ denote all the possible divisions of the N ranks into the n and m control ranks for which the treatment ranks are summed. The universe of possible divisions of the ranks is given by the n subsets that can be formed from N , known as a binomial coefficient in combinatorics. Using the binomial coefficient, each possible division of the ranks for the treatment group occurs with probability:

$$\frac{1}{\binom{N}{n}}.$$

Therefore, it follows that:

$$P(W = c) = \frac{\#(w; n, m)}{\binom{N}{n}}.$$

This provides an exact distribution free test of the null hypothesis based on the randomization of the treatment (Lehmann 1975). This exact p -value only requires that the treatment is randomized. Confidence intervals for a constant-additive treatment effect are formed through inverting a series of null hypotheses. This confidence interval, as discussed previously, is especially informative about what can be learned from small sample sizes. The rank sum test also has an associated Hodges-Lehmann point estimator, which is the point estimate median of the $m(N - m)$ pairwise differences formed by taking the each of the m treated responses and subtracting each of the control responses (Hodges and Lehmann 1963). The rank sum test we have described assumes that there are no ties in the ranks. If there are ties, the tied observations are given the average of the ranks for the tied observations. For example if the scores for five observations were: 0.7, 1.2, 1.7, 1.7, and 2.8, the ranks for the tied data would be $(3+4)/2 = 3.5$. Once midranks are assigned, exact p -values may be calculated in a similar fashion as before using permutations of the data (Hollander and Wolfe 1999).

The reader might wonder about the power of the rank sum test in comparison to more traditional tests, such as the t -test. The power of a test depends upon a number of factors: the sample size, α , the largest p -value that indicates statistical significance, the size of the departure from the null hypothesis, and some evaluation of whether the test assumptions hold. The standard method of comparing the

power of tests relies on the asymptotic relative efficiency (ARE), which was developed by Pittman (1948) in a series of unpublished lecture notes, and thus is often also referred to as the Pittman efficiency. For two tests T_1 and T_2 , the ARE provides a useful comparison of test efficiency depending on the underlying distribution of the data.¹³ If the $ARE = 2$, this implies that T_2 requires twice as many observations as T_1 to achieve a given level of power. ARE calculations for the rank-sum test and the t -test have been known for some time (Hodges and Lehmann 1956). Table 4 contains the ARE for the rank-sum compared to the t -test for several different population distributions. We see that when the data are drawn from a normal distribution the rank sum test requires approximately 4% more data to achieve the same power as the t -test. For a heavy-tailed distribution, such as the exponential, a t -test would require 3 times as many observations to be as powerful as the rank sum test. The ARE comparison further underscores the utility of the rank test. The loss of power is minimal when the assumptions of the t -test are met but are considerable when those assumptions are not met.

Table 4: Asymptotic Relative Efficiency of Rank Sum Test and t -test

	Normal	Uniform	Logistic	Double Exponential	Exponential
t/W	0.96	1.00	1.10	1.50	3.00
Source: Hollander and Wolfe (1999, pg. 140)					

3.2 Comparing Multiple Treatments: The Kruskal-Wallis Test

Many experimental research designs are not limited to a single treatment, but instead may involve the comparison of three or more treatment conditions. Here, the analyst may administer k treatments and compare the locations for each of the treatment groups, or one of the treatments may be a control group, and one may wish to detect whether any of the $k - 1$ treatments differ from this control. More formally, the null hypothesis of no difference among treatments is:

$$H_0 : \tau_1 = \dots = \tau_k$$

where τ is the treatment effect. Under this null hypothesis, the distributions of each treatment group is the same. Standard practice in political science for evaluating such a hypothesis is to perform a one-way ANOVA based on the F distribution. The rank-based randomization test alternative we outline is the Kruskal-Wallis test. This test does not have a specific estimated quantity of interest associated with it. Tables of medians would be one appropriate summary statistic of the experimental outcomes.

¹³For two tests T_1 and T_2 , we set α to a given level and use a sequence of k values under the alternative hypothesis denoted as θ_k . For sample sizes of n_1 and n_2 , one calculates the following Type II error rates: $T_1(\theta_k) = \beta$ and $T_2(\theta_k) = \beta$. The asymptotic relative efficiency of T_1 with respect to T_2 is: $ARE = \lim_{k \rightarrow \infty} \frac{n_2}{n_1}$. Importantly, the ARE is the same for any choice of α and β (Sprenst and Smeeton 2007).

The Kruskal-Wallis statistic, H , is based on a joint ranking of the data. To compute H , all N observations from the k treatment groups are combined and ranked. Let r_{ij} denote the rank of each observation in the joint ranking. From this joint ranking of the data, we calculate two values. The first is:

$$R_j = \sum_{i=1}^{n_j} r_{ij}$$

which is the sum of the joint ranks for the j th treatment group. The second value is:

$$R_{.j} = \frac{r_j}{n_j}$$

which is the average rank for the j th treatment group. If the treatments are effective, we would expect large differences in the values of each R_j . If the treatments are ineffective, the values of R_j should be similar and hence close to the overall average of the ranks. The Kruskal-Wallis statistic compares a weighted average of the overall ranks to the average rank within each treatment group. It is:

$$H = \frac{12}{N(N+1)} \sum_{j=1}^k n_j \left(R_{.j} - \frac{N+1}{2} \right)^2$$

H will be large when there are substantial differences among the $R_{.j}$ terms and is zero when they are all equal. To test the null hypothesis at the traditional 0.05 level of significance, we would select c to make the following statement true

$$P(H \geq c) = 0.05$$

Selection of c at any level of significance requires that we calculate the null distribution for H . We calculate the null distribution of H using a straightforward application of combinatorics. From N subjects, we randomly allocate n subjects to each treatment group so that:

$$n_1 + \cdots + n_k = N$$

The number of possible assignments of the subjects is given by the multinomial coefficient. Under the assumption of random assignment, all possible combinations are equally likely, therefore each possible outcome occurs with equal probability. By counting the number of times an observed treatment value for H occurs under the null distribution, we can straightforwardly calculate an exact distribution free p -value. For large values of N , calculation of exact p -values can be cumbersome. Early practice in such situations was to rely on an asymptotic approximation between H and a χ^2 distribution

(Lehmann 1975). Modern computing allows analysts to use Monte Carlo simulation for a more accurate approximation of the null distribution for H (Sprent and Smeeton 2007). Also note that the Kruskal-Wallis statistic assumes that there are no ties in the ranks of the data. If ties are present, midranks are used, as in the case of the Wilcoxon sum rank test, and some adjustment is required to calculate the null distribution of H , but this presents no complications. Hollander and Wolfe (1999) provide details on this adjustment. The Kruskal-Wallis test can be generalized in several ways. For example, the test can be altered to accommodate randomized block designs or increasing (or decreasing) treatment effects. The Kruskal-Wallis test does not have an associated point estimate or confidence interval, though one could use the attributable effects framework for such quantities.

3.3 Randomization Tests for Two-way Factorial Designs

As we emphasized earlier, a randomization inference is built from the design of the experiment. For simple randomized and oneway factorial designs, building the statistical test does not require much special care by the analyst. More complex designs, however, need more attention. This is true for twoway factorial designs which require the experimenter to define the test statistic of interest depending on the the question of interest. In a twoway design, two randomized treatments are applied to all subjects. The experimenter is interested in testing whether the presence of one treatment modifies the effect of the other treatment. The data from such designs are almost always analyzed with a regression model. When using the regression model, the test for an interaction between the two treatments is easily misapplied (Clark, Gilligan, and Golder 2006). This may be in part because the regression model does not require definition of which treatment is the moderator when testing for the presence of an interaction. Moreover, if the interaction is statistically significant interpretation of the main effects is no longer transparent in the regression model.

The test statistic for the randomization inference requires the analyst to precisely define the nature of the interaction before conducting the test. That is, one must declare which of the two treatments is the modifying factor. The treatment effect of factor B must be tested with the levels of A. Moreover, the nature of the test statistics can be fairly complex for designs with many levels in each factor. We start with a 2×2 design and then generalize to $R \times C$ factorial designs. We assume that two binary treatments are randomized and applied to all subjects in the study. Let A be the first factor with levels $1, \dots, i$; $i = 2$, and B is the second factor with levels $1, \dots, j$; $j = 2$. Each unit response, we designate as Y_{ijk} where the k th replicate receives the i th level of factor A and the j th level of factor B with $1 \leq k \leq n_{ij}$. Each unit response can be written as the following linear function

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

where μ_{ij} is the effect of the i th level of factor A and the j th level of factor B . We assume e_{ijk} are random errors that are independently and identically distributed. If equal numbers of subjects are assigned to each treatment combination the design is said to be balanced, and if unequal numbers of subjects are assigned to each treatment combination, the design is said to be unbalanced, which must be taken into account for accurate assessment of the treatment effects. Political science designs are often unbalanced, though small departures from a fully balanced design are usually inconsequential. The rank test we outline below makes no assumptions about whether the design is balanced or not. Regression model analyses assume the data are balanced and require the use of Type III sum of squares for correct tests for interaction (Maxwell and Delaney 2003). The standard regression model is easily misapplied when the data are unbalanced, but the randomization inference protects against such misapplication.

The usual data analytic strategy uses the following linear regression model:

$$Y_{ijk} = \mu + \alpha_j A + \beta_k B + (\alpha\beta)_{jk} A \times B + \epsilon \quad (6)$$

where Y_{ijk} is the outcome for individual i in treatment combination jk . The term α_j represents the effect of treatment A and β_k is the effect of treatment B , while $(\alpha\beta)_{jk}$ represents any joint interactive effect of the two treatments. This model assumes that the variances within each treatment combination are equal (Maxwell and Delaney 2003). Using a regression model does not require the analyst to declare which treatment modifies the other. The usual F-test provides an omnibus test for the presence of an interaction.

We outline a rank based test focusing on testing for an interaction between the effect of A and B . If there is no interaction between these two treatments, then the design can be analyzed as if two separate experiments had been conducted. We designate A as the modifying factor which implies that the effect of B varies with the levels of A . Thus we define two within strata null hypotheses for the effect of B . First, we state the null for the effect of B within the first level of A :

$$H_0 : \mu_{11} - \mu_{12} = 0$$

Second, we state the null for the effect of B within the second level of A :

$$H_0 : \mu_{21} - \mu_{22} = 0$$

If we believe no interaction is present, the within strata effects of B should be constant which implies

the following.

$$\mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}$$

To test for an interaction, we use the following null hypothesis:

$$H_0 : (\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = 0$$

To test this null hypothesis, we need a rank based statistic for the within strata effects of B . We use the approach of Patel and Hoel's (1973) rank-based statistic. This rank based statistic relies on the following sign function:

$$\psi(X_{ijk}, X_{i'j'k'}) = \begin{cases} 0 & \text{if } X_{ijk} < X_{i'j'k'} \\ \frac{1}{2} & \text{if } = \\ 1 & \text{if } > \end{cases}$$

The rank statistic within each strata of A is:

$$V_i = \frac{\sum_{j=1}^{n_{i1}} \sum_{k=1}^{n_{i2}} \psi(X_{ijk}, X_{i'j'k'})}{n_{i1}n_{i2}}, \quad i = 1, 2$$

The estimator for the interaction test is simply:

$$T = V_1 - V_2$$

Asymptotically this statistic is normally distributed with a variance estimate as defined in Patel and Hoel (1973). Instead of relying on this asymptotic approximation, we can simply permute observations within each level of the modifying factor to obtain a randomization test p -value for the interaction. If the interaction is statistically significant, each V_i serves as the within strata test statistic that is nearly identical in form to the usual rank sum test.

For higher order designs, the definition of the interaction test statistic remains unchanged but one may devise a variety of within strata test statistics. For example, assume that B now was three instead of two levels. The within-strata test statistics can be defined in multiple ways. Patel And Hoel's (1973) outline a test statistic that makes graduated comparisons. More complex statistics are also possible. For example, the within strata test could include all possible comparisons or test for a dose response. However the within strata test statistics are defined, the differences across these statistics forms the interaction test statistic. For example, suppose that instead the design is such that A has three levels while B has two. Here, the rank statistic defined above generalizes but the test statistic for the interaction is now the the difference of three within strata test statistics. Analyst must also take care to permute observations across the levels of B within each strata of A .

4 Examples from Political Science Experiments

Having laid out our case for randomization tests, we now turn to applying these tests to two datasets from political science experiments. We use a dataset from Fowler and Kam (2007), whose published results represent a rare example of the use of randomization tests in political science, and part of a dataset produced by White (2003) from which results have not been previously published. Both experiments were performed on convenience samples – one relying entirely on student subjects and the other recruiting both students and non-student adults. We compare the results from standard statistical tests to those from randomization tests. In addition to offering a more direct estimate of the type of uncertainty that concerns the experimentalist, we find that randomization tests can produce p -values that would lead to different substantive conclusions.

4.1 Partisan Generosity

Fowler and Kam (2007) performed a series of experiments to test a set of hypotheses about individuals' propensity to give to others. The authors brought student subjects into a laboratory environment and asked them to play a dictator game, wherein each subject was given a set of 10 lottery tickets and asked to divide the tickets between themselves and an anonymous recipient. By manipulating the identity of the anonymous recipient, Fowler and Kam intended to test for differences in giving that could be attributed to the effects of social identities. Thus three experimental conditions were employed, the treatment being the identity of the recipient: no identifying information, registered Democrat, or registered Republican.

Among Fowler and Kam's expectations was the hypothesis that subjects would display an in-group preference and thus give more when the recipient was revealed to be in the same party as themselves. One way to use their data to assess this hypothesis is to compare the amounts given to the recipient when: 1) the subjects' partisan identities matched the recipient, 2) the subjects' identities diverged from the recipient, and 3) the subjects had no information about the identity of the recipient. The authors looked at these comparisons separately among those who strongly identified with their party label and those who weakly did so; we do the same. In the replication, we perform three different tests. First, we use the standard one-way ANOVA. We next use a randomization test where the test statistic is an F-test statistic from the same ANOVA table but the p -value is calculated from all permutations of the data. Finally, we calculated the Kruskal-Wallis test based on ranks. This allows us to compare different test statistics under the randomization testing paradigm. In Table 5, we display, side-by-side, the p -values from the three tests across the three conditions for the two partisan groups. Among the weak partisans, the differences are minimal. For the strong partisans, the permuted test provides a p -value similar to the standard test. The rank test, however, appears to be more powerful

due to the presence of several outliers in the data. Among strong partisans, in fact, the decrease is enough to change whether or not the null hypothesis would be rejected at the conventional .05 level.

Table 5: ANOVA and Kruskal-Wallis p -value Comparisons

	ANOVA	Permuted F-test	Kruskal-Wallis
Strong Partisans	0.055	0.045	0.006
Weak Partisans	0.515	0.530	0.499

Note: There are 127 subjects per cell for the first row tests, and 125 subjects per cell for the second row tests.

If the researchers had specific hypotheses about differences not across all three conditions, but rather across any two, we could employ either a permuted difference in means or a rank sum test as an alternative to the asymptotic approximation provided by the t -test on the difference in amounts given.¹⁴ For example, the researchers might have specifically expected a difference in subjects' giving in the same-party recipient condition as compared to the "control" condition where no information about the recipient is given. Alternatively, they might have been interested specifically in the difference in giving when the recipient is identified with the in-group as compared to when the recipient is identified with the out-group. Here, we compare a standard t -test to a permuted t -test and the rank sum test for Fowler and Kam's data that would be used to test these two pair-wise comparison hypotheses among strong partisans. The results from the three tests are in Table 6.

The asymptotic t -test does provide similar results to the permuted t -test. The test statistic based on ranks provides greater power once again.¹⁵ An inspection of the data reveals several noticeable outliers, and the rank sum test has greater power in the presence of such outliers. So we see that under the rank test, the chance that random assignment would produce such a level of partisan generosity among strong partisans is well below 1%, therefore this provides strong evidence against the null hypothesis of no treatment effect. Both analyses underscore the robust properties of rank based test statistics in the presence of outliers.

For one-way ANOVA models and the Kruskal-Wallis test, there is not a summary statistic for the effect just a p -value for the test, but under the randomization inference framework, we can develop a test statistic, confidence intervals and a point estimator. First, we need a test statistic for some relevant aspect of the experimental design. We use the average median difference across the three partisan categories as a test statistic for the one-way design. That is, we calculate the median number

¹⁴The t -test in this instance can be interpreted as either a test of whether the ATE is different from zero or an asymptotic approximation to the test of the sharp null under the randomization test.

¹⁵Note that if we proceeded through a series of pair-wise tests to test for differences across conditions we would really want tests that account for the multiple comparisons we were making. Appropriate randomization tests exist (Hollander and Wolfe 1999), or one could simply use a Bonferroni correction.

Table 6: t -test and Wilcoxon Rank Sum p -value Comparisons

	t -test	Permuted t -test	Rank Sum
In-Group vs. Control	0.050	0.053	0.015
In-group vs. Out-group	0.029	0.032	0.003

Note: There are 127 subjects per cell for the first row tests, and 125 subjects per cell for the second row tests.

of lottery tickets given away across the three treatments and take the average. This provides use with a measure of how behavior changed across the experimental conditions. We can calculate an approximate exact p -value for this test statistic using 10,000 simulations from the null distribution. We find that the approximate exact p -value for a one-sided test is 0.081, so it is statistically significant at the 0.10 level.

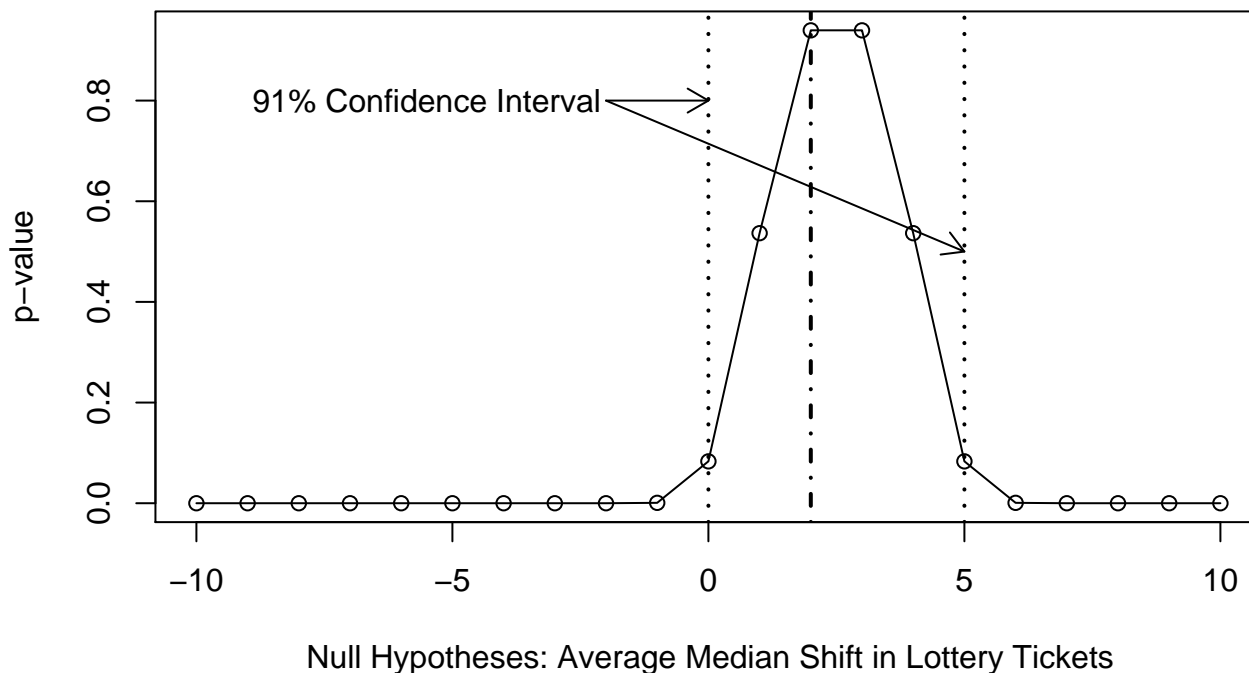


Figure 1: Attributable Effects Against Exact p -values for Dictator Game

The next step is to develop a model of effects. We adopt the usual constant-additive model of effects. Thus we assume that the treatment effect is constant and additive across each condition. In this case, we use the observed summary statistic for the data as the point estimate, which here is 2.

That is, the average median difference across treatment categories was 2 lottery tickets. For interval estimation, we specified integers from -10 to 10 as the values for A , the range of null hypothesis values. For each value in the range of A , we subtracted this value from the outcomes of the out-party condition, then calculated the approximate exact p -value based on the true null distribution. We next construct a confidence interval for this estimate. Given the discrete nature of exact p -values, we don't observe a value at the point necessary to construct a 95% confidence interval. Thus we draw the 91% confidence interval instead. In the figure, the exact p -value is plotted against the range of null hypotheses. Based on the randomization null distribution, this point estimate has a 91% confidence interval of [0, 5]. This example demonstrates how one can easily construct summary statistics built from the randomization in the experiment and then conduct hypothesis tests and estimate confidence intervals for these summary statistics.

4.2 Racial Cues

In our second example, we analyze data from White (2003). White designed an experiment to test the effects of two types of racial cues in political communication: a source cue and a racial frame. The treatment received by all subjects was a news magazine article laying out arguments for opposition to the war in Iraq; the article was manipulated across conditions to vary both the frame of the opposition argument and the source of the article. Two frames were employed in the experiment: an explicitly racial frame and an implicitly racial frame. Each of the frames was presented inside a news story appearing in either a black news magazine (*Black Enterprise*) or a mainstream news magazine (*Newsweek*). Thus the experiment is a 2 x 2 factorial design with a total of four conditions, and subjects were randomly assigned to the conditions. Subjects were then asked to report their level of belief, on a 1-7 scale, in three arguments for the war: whether the U.S. was too quick to use military force, and whether they approved of how George W. Bush handled the situation in Iraq. The experiment was run separately on both white and black subjects, as the theory implied that blacks and whites would respond differently to the treatments.

Among the expectations was the hypothesis that blacks would be persuaded to be less supportive of the war and less receptive to arguments used to justify the war when they were exposed to any of the racial cues. Additionally, it was hypothesized that the effect of the frame might depend on the source in which the article appeared, notably that the two types of racial cues (source and frame) might have mutually reinforcing effects. Thus the frame was postulated as the modifying factor. We concentrate on these hypotheses, and thus only analyze the data from the black subjects. Given that this is a two-way design, we test for two "main" effects and for an interaction. We assessed these hypotheses using a standard twoway ANOVA which relies on F-tests and the rank-base method outlined in Section 3.3. For each of the permutation tests, we used 10,000 permutations to form the null

distribution, though we found that using 1,000 permutations made little difference. That is we took 10,000 random permutations of the data and calculated the test-statistic to form a null distribution. Table 7 contains a comparison of the p -values that resulted from both the standard two-way ANOVA and the rank-based method. We report the results for three tests for each outcome. We test the effect of the racial cue within each level of the media source treatment and then we test whether those effects differ.

In this experiment, we find that the randomization tests produce generally lower p -values for the test of the interaction, often changing whether or not the null hypotheses would be rejected. In the first example, our inference is maintained but the difference is p -values is 0.13. For the other two outcomes, we narrowly conclude that an interaction is present based on the asymptotic test. The randomization based inference, however, provides stronger evidence that an interaction is present. While the randomization inference in this instance lowers the p -value in all three instances, such differences cannot be assumed to hold in other data sets.

Table 7: Comparison of Asymptotic and Permuted p -value for Two-way ANOVA

	Parametric p -value	Approx. Exact p -value
Security Council Approval for War		
Racial Cue within Media Source Level 1	0.353	0.235
Racial Cue within Media Source Level 2	0.546	0.464
Cue \times Source Interaction	0.292	0.162
Iraq has chemical weapons		
Racial Cue within Media Source Level 1	0.054	0.037
Racial Cue within Media Source Level 2	0.589	0.373
Cue \times Source Interaction	0.047	0.026
Approve George W. Bush		
Racial Cue within Media Source Level 1	0.051	0.012
Racial Cue within Media Source Level 2	0.585	0.361
Cue \times Source Interaction	0.082	0.015
Note: Approximate exact p -values based on 10,000 permutations of the data. Cell sizes range from 23 to 28.		

5 Conclusion

Our treatment of randomization tests has certainly not been exhaustive, but we have endeavored to cover tests for standard political science experimental designs. There is a large variety of tests for other designs, including block and within-subjects designs. While we have not explored the complete range of randomization inference possibilities in this paper, we hope we have sufficiently covered the basic logic behind randomization tests. We should also note that randomization tests do not preclude the need

for adjustment due to accidental imbalances. Analysts should check for imbalances across treatment and control groups and adjust if necessary. The randomization tests can be extended in many ways. Rosenbaum (2002a) outlines a method for covariance adjustment that is fully integrated with randomization tests and is easy to implement. Such covariate adjustment often helps increase the power of the test. There are also randomization tests for block designs and within-subjects experiments. Rosenbaum (1996) also develops randomization tests for intention to treat analyses when noncompliance occurs. Hansen and Bowers (2008) adapt the randomization framework to an experimental design with clustering and noncompliance.

While experiments offers unique and considerable leverage on questions of causal inference, the traditional statistical methods political science experimenters have had at their disposal are often inappropriate, or at least less than ideal, given the designs they employ. We have argued that randomization tests are better tools for the purposes of these experimenters. They offer increased conceptual coherence through their direct estimation of the uncertainty introduced by random assignment. The flexibility of the approach offers the experimenter a range of tests to meet their particular data needs, including rank-based tests that handle the problems of outliers and skewed distributions. And, as we have illustrated, these tests can make a practical difference. Applying these tests to existing experimental data, we showed how using the appropriate randomization test can mean a different inference might be made than would be with parametric alternatives on the same data—in some cases lowering p -values below standard levels of statistical significance, and raising them over the same threshold in other cases. Some of these differences were due to the use of ranks, which allowed for more robust testing of hypotheses in the presence of outliers or heavy-tailed distributions. Given the confluence in randomization tests of the conceptual appeal of the interpretation of the p -values, the freedom from parametric assumptions, confidence intervals that are informative about testing power, and a capacity to change substantive inferential conclusions, it seems the experimentalist would be well-served by the addition of randomization tests to his/her methodological toolkit. Modern computing capacity makes doing so far more feasible than in decades past (though we acknowledge that the average practitioner would be better served by greater integration of a full range of randomization tests into the statistical software with which they are already familiar); familiarity seems the only remaining impediment to the adoption of a randomization inference approach.

6 Appendix

6.1 Nonconstant Effects

The sharp null hypothesis and the model of constant-additive effects are predicated on the assumption that treatment effects are constant across units in the study. We might expect, instead, that for some

units the treatment produced a positive effect, while for other units the effect may be zero or possibly even negative. When the test statistic used is based on ranks, one can make probabilistic statements about the presence of nonconstant treatment effects (Rosenbaum 2001, 2003).¹⁶ Rosenbaum's procedure allows us to state the magnitude of positive effects due to treatment with a chosen confidence level. First, we let W be the test statistic for the sum rank test.¹⁷ Next we calculate, c_α , the value at which we would reject the null for a $1 - \alpha$ confidence level. This can be done with distributional tables such as those found in Hollander and Wolfe (1999) or with statistical software. We would then be $(1 - \alpha \times 100)$ percent confident that at least $W - c_\alpha + 1$ of the treated outcomes were positive due to treatment. This value is informative but must be adjusted for positive differences that might occur due to chance. Therefore, we next calculate the null expectation of positive differences. Let $E(W_0)$ be the expected number of positive differences under the null hypothesis of no treatment effect. For the sum rank test, the null expectation is $m(N - m)/2$, where N is the total number of units and m is the number of treated units. As such, $(W - E(W_0))/E(W_0)$ is the expected percentage of positive differences greater than would be produced by chance. Thus, we could be $1 - \alpha \times 100$ percent confident that $W - c_\alpha + 1/E(W_0) \times 100$ percent of the excess in positive differences were caused by treatment and not by chance fluctuations. These quantities reveal how plausible the sharp null is and as such serve as an important diagnostic.

As an example, we return to the simulated data in Section 2.3. The sum rank test statistic, W , in this example is 35.5; under a test of the sharp null hypothesis we would reject the null decisively (exact p -value = .002). We next need to calculate c_α . We would reject the sharp null hypothesis for any value of W greater than 28 at the 0.05 level, so $c_{0.05}$ is 28. In this experiment, we can be 95% confidence that at least $35.5 - 28 + 1 = 8.5$ of the differences were positive because of the treatment. In an experiment of this size, if there was no treatment effect, we would expect $7(14 - 7)/2 = 18$ positive differences to occur by chance. Therefore, the observed 35.5 positive differences are 97% larger than we would expect by chance. We can also be 95% confident that at least $8.5/18 = 47\%$ of this excess was caused by the treatment and not by chance fluctuations.

¹⁶It is also possible extend non-constant effects to the Fisher exact test, another nonparametric test. While it should be possible to extend these procedures to other test statistics, we are not aware of any work that does this.

¹⁷Specifically, this is the Mann Whitney U test statistic which differs by a constant from Wilcoxon's sum rank test statistic. Most software reports U. Details on the rank sum test are below. A large number of ties will result in a more conservative statement of positive effects due to treatment (Rosenbaum 2001).

References

- Brader, Ted. 2005. "Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions." *American Journal of Political Science* 49(April): 388–405.
- Brewer, Marilynn. 2000. "Research Design and Issues of Validity." In *Handbook of Research Methods in Social and Personality Psychology*, ed. Harry T. Reis, and Charles M. Judd. Cambridge: Cambridge University Press pp. 3–16.
- Clark, William, Michael Gilligan, and Matt Golder. 2006. "A Simple Multivariate Test for Asymmetric Hypotheses." *Political Analysis* 14(Summer): 311–331.
- Cobb, Michael D, and James H. Kuklinski. 1997. "Changing Minds: Political Arguments and Political Persuasion." *American Journal of Political Science* 41(January): 88–121.
- Cox, David R. 1958. *Planning of Experiments*. New York: Wiley.
- Diaconis, Persi, and Erich Lehmann. 2008. "Comment: On Student's 1908 Article "The Probable Error of a Mean"." *Journal of the American Statistical Association* 103(March): 16–18.
- Druckman, James N. 2001. "On the Limits of Framing Effects: Who Can Frame?" *The Journal of Politics* 63(November): 1041–1066.
- Druckman, James N., and Kjersten R. Nelson. 2003. "Framing and Deliberation: How Citizens' Conversations Limit Elite Influence." *American Journal of Political Science* 47(October): 2003.
- Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100(November): 627–635.
- Falk, Armin, and James J. Heckman. 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326(October): 535–538.
- Fisher, Ronald A. 1935. *The Design of Experiments*. London: Oliver and Boyd.
- Fowler, James H., and Cindy D. Kam. 2007. "Beyond the Self: Social Identity, Altruism, and Political Participation." *The Journal of Politics* 69(August): 813–827.
- Freedman, David A. 2008a. "On Regression Adjustments in Experimental Data." *Advances in Applied Mathematics* Forthcoming.
- Freedman, David A. 2008b. "On Regression Adjustments in Experiments with Several Treatments." *Annals of Applied Statistics* 2(March): 179–196.
- Freedman, David A. 2008c. "Randomization Does Not Justify Logistic Regression."
- Gelman, Andrew, John S. Carlin, Hal S. Stern, and Donald B. Rubin. 2003. *Bayesian Data Analysis*. 2 ed. Boca Raton, FL: Chapman and Hall.
- Gibson, James L. 2002. "Truth, Justice, and Reconciliation: Judging the Fairness of Amnesty in South Africa." *American Journal of Political Science* 46(July): 540–556.
- Gilliam, Franklin D., and Shanto Iyengar. 2000. "Prime Suspects: The Influence of Local Television News on the Viewing Public." *American Journal of Political Science* 44(July): 560–573.
- Golebiowska, Ewa A. 1996. "The "Pictures in Our Heads" and Individual-Targeted Tolerance." *The Journal of Politics* 58(November): 1010–1034.
- Green, Donald P., and Alan S. Gerber. 2002. "Reclaiming The Experimental Tradition in Political Science." In *Political Science: The State of the Discipline*, ed. Ira Katznelson, and Helen V. Milner. New York: W.W. Norton pp. 805–832.

- Hansen, Ben B., and Jake Bowers. 2008. "Covariate Balance in Simple, Stratified, and Clustered Comparative Studies." *Statistical Science* Forthcoming.
- Hansen, Ben B., and Jake Bowers. 2009. "Attributing Effects to A Clustered Randomized Get-Out-The-Vote Campaign." *Journal of the American Statistical Association* 104(September): 873–885.
- Hibbing, John R., and John R. Alford. 2004. "Accepting Authoritative Decisions: Humans as Wary Cooperators." *American Political Science Review* 48(January): 62–76.
- Ho, Daniel E., and Kosuke Imai. 2006. "Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 Election." *Journal of the American Statistical Association* 101(September): 888–900.
- Hodges, J. L., and E.L. Lehmann. 1963. "Estimates of Location Based on Ranks." *The Annals of Mathematical Statistics* 34(June): 598–611.
- Hodges, J.L. Jr., and E.L. Lehmann. 1956. "The Efficiency of Some Nonparametric Competitors to of the t -test." *The Annals of Mathematical Statistics* 27(June): 324–335.
- Hoeffding, W. 1952. "The Large Sample Power of Tests Based on Permutations of the Observations." *The Annals of Mathematical Statistics* 23(June): 169–192.
- Hollander, Myles, and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. 2nd ed. New York, NY: John Wiley and Sons.
- Imbens, Guido W. 2009. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).".
- Imbens, Guido W., and Donald B. Rubin. 2008. *Causal Inference in Statistics and the Medical and Social Sciences*. Vol. Forthcoming Cambridge, UK: Cambridge University Press.
- Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(March): 467–476.
- Imbens, Guido W., and Paul Rosenbaum. 2005. "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education." *Journal of The Royal Statistical Society Series A* 168(January): 109–126.
- Kam, Cindy D., Jennifer R Wilking, and Elizabeth J. Zachmeister. 2008. "Beyond the "Narrow Data Base": Another Convenience Sample for Experimental Research." *Political Behavior* Forthcoming.
- Kinder, Donald R., and Thomas R. Palfrey. 1993. "On Behalf Of An Experimental Political Science." In *Experimental Foundations of Political Science*, ed. Donald R. Kinder, and Thomas R. Palfrey. Ann Arbor: University of Michigan Press pp. 1–39.
- Lehmann, E. L. 1975. *Nonparametrics: Statistics Based on Ranks*. San Francisco, CA: Holden-Day.
- Lehmann, E. L., and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. 3rd ed. New York, NY: Springer.
- Maxwell, Scott E., and Harold D. Delaney. 2003. *Designing Experiments and Analyzing Data: A Model Comparison Approach*. 2nd ed. Philadelphia, PA: Lawrence Erlbaum.
- Miller, Joanne M., and Jon A. Krosnick. 2000. "News Media Impact on the Ingredient of Presidential Evaluation: Politically Knowledgeable Citizens Are Guided by a Trusted Source." *American Journal of Political Science* 44(April): 301–315.
- Mitchell, Gregory, Philip E. Tetlock, Daniel G. Newman, and Jennifer S. Lerner. 2003. "Experiments behind the Veil: Structural Influences on Judgments of Social Justice." *Political Psychology* 24(September): 519–547.

- Morton, Rebecca, and Kenneth Williams. n.d. "From Nature to the Lab: Experimental Political Science and the Study of Causality."
- Nelson, Thomas E., Rosalee Clawson, and Zoe M. Oxley. 1997. "Media Framing of a Civil Liberties Conflict and Its Effect on Tolerance." *American Political Science Review* 91(September): 567–583.
- Neyman, Jerzy. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5(November): 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed (1990).
- Patel, Kantilal, and David G. Hoel. 1973. "A Nonparametric Test of Interaction in Factorial Experiments." *Journal of the American Statistical Association* 68(343): 615–620.
- Pittman, E.J.G. 1948. "Lecture Notes on Nonparametric Statistics."
- Rosenbaum, Paul R. 1996. "Identification of Causal Effects Using Instrumental Variables: Comment." *Journal of the American Statistical Association* 91(June): 465–468.
- Rosenbaum, Paul R. 2001. "Effects Attributable To Treatment: Inference In Experiments And Observational Studies With A Discrete Pivot." *Biometrika* 88(March): 219–231.
- Rosenbaum, Paul R. 2002a. "Covariance Adjustment In Randomized Experiments and Observational Studies." *Statistical Science* 17(August): 286–387.
- Rosenbaum, Paul R. 2002b. *Observational Studies*. 2nd ed. New York, NY: Springer.
- Rosenbaum, Paul R. 2003. "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test." *The American Statistician* 57(May): 132–138.
- Rubin, Donald B. 1996. "Which Ifs Have Causal Answers." *Journal of the American Statistical Association* 81(December): 961–962.
- Sears, David O. 1986. "College Sophmores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature." *Journal of Personality and Social Psychology* 51(3): 515–530.
- Sekhon, Jasjeet S., and Rocio Titiumik. 2008. "Exploiting Tom Delay: A New Method for Estimating Incumbeny Advantage and the Effect of Candidate Ethnicity on Turnout."
- Sigelman, Carol K, Lee Sigelman, Barbara J. Walkosz, and Michael Nitz. 1995. "Black Candidates, White Voters: Understanding Racial Bias in Political Perceptions." *American Journal of Political Science* 39(February): 243–265.
- Sprent, Peter, and Nigel C. Smeeton. 2007. *Applied Nonparametric Statistical Methods*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC.
- Valentino, Nicholas A., Vincent L. Hutchings, and Ismail K. White. 2002. "Cues That Matter: How Political Ads Prime Racial Attitudes during Campaigns." *American Political Science Review* 96(March): 75–90.
- Walsh, A. H. 1996. *Aspects of Statistical Inference*. Hoboken, NJ: Wiley and Sons.
- White, Ismail K. 2003. "Racial Perceptions of Support for the Iraq War."