Interference Between Units in Randomized Experiments

Paul R. ROSENBAUM

In a randomized experiment comparing two treatments, there is interference between units if applying the treatment to one unit may affect other units. Interference implies that treatment effects are not comparisons of two potential responses that a unit may exhibit, one under treatment and the other under control, but instead are inherently more complex. Interference is common in social settings where people communicate, compete, or spread disease; in studies that treat one part of an organism using a symmetrical part as control; in studies that apply different treatments to the same organism at different times; and in many other situations. Available statistical tools are limited. For instance, Fisher's sharp null hypothesis of no treatment effect implicitly entails no interference, and so his randomization test may be used to test no effect, but conventional ways of inverting the test to obtain confidence intervals, say for an additive effect, are not applicable with interference. Another commonly used approach assumes that interference is of a simple parametric form confined to units that are near one another in time or space; this is useful when applicable but is of little use when interference may be widespread and of uncertain form. Exact, nonparametric methods are developed for inverting randomization tests to obtain confidence intervals for magnitudes of effect assuming nothing at all about the structure of the interference between units. The limitations of these methods are discussed. To illustrate the general approach, two simple methods and two simple empirical examples are discussed. Extension to randomization based covariance adjustment is briefly described.

KEY WORDS: Attributable effect; Causal effect; Interference; Randomized experiment; SUTVA.

1. INTRODUCTION: EXAMPLES AND REVIEW

1.1 What Is Interference Between Units?

"There is no 'interference' between different units," wrote Cox (1958a, p. 19), if "the observation on one unit [is] unaffected by the particular assignment of treatments to the other units." Rubin (1986) called this the "stable unit-treatment value assumption" or SUTVA; it is key in saying, with Neyman (1923) and Rubin (1974), that the effect of a treatment compares two potential responses that the unit would exhibit under treatment and under control. If there is interference, then the unit has not two but many potential responses, depending on the treatments assigned to other units. Rubin (1990, p. 475) noted that: "interference between units can be a major issue when studying medical treatments for infectious disease... or educational treatments given to children who interact with each other."

Interference between units is a precisely defined statistical issue (see Sec. 2.1), but it is not always clearly distinguished from other quite different issues that also may arise in contexts that produce interference. There is interference if the treatment given to a person as an experimental subject does not affect just that person, but also other experimental subjects, or if the treatment given to a person in one location at one time affects that person's response at other locations or times. Interference is distinct from statistical dependence produced by pretreatment clustering, although both may be present. People in the same family may tend to exhibit similar responses to a viral infection because of shared genes; this is clustering. Vaccinating one child may prevent her from contracting a viral infection and spreading it to her unvaccinated brother; this is interference. Interference between units in the effects of vaccination is not limited to well-defined clusters of people; it may be unlimited in extent and impossible to specify in form.

It is commonly and reasonably argued that interference between units should be avoided whenever possible, perhaps by isolating experimental units from one another, so interference does not occur. Although this is good advice in many settings, it is highly impractical or not logically possible in many common situations; see the examples in Section 1.4. What analytical options are currently available for randomized experiments with interference?

1.2 Existing Methods for Interference in Experiments

Besides preventing inference by isolating experimental units, the most common tactics in randomized experiments with interference are limiting attention to a randomization test of no effect, and modeling interference that is local in time or space. Fisher's null hypothesis of no effect says that every unit would exhibit the same response under all treatment assignments, so if this hypothesis is true, then there is no effect and no interference in effect (see Sec. 2.2). Consequently, in a randomized experiment, Fisher's permutation test of no effect has the correct level, even if, under the alternative hypothesis, it is plausible that treatment effects exhibit interference. Moreover, Fisher's test uses randomization as the sole basis for inference, so careful and appropriate application of the method yields a valid test in the presence of dependence of various forms, such as clustering; see, for instance, Fisher's (1935, sec. 2) careful discussion of the single-subject experiment involving the lady tasting tea. In many randomized experiments with interference, formal statistical inference begins and ends with a test of no effect. In most contexts, it is useful to have confidence statements about the magnitude of effect, not just a test of no effect. Can valid confidence statements be created in randomized experiments when interference is possible?

There is a large, diverse literature about methods to address the distinct problem of clustering, but there are no general methods of inference about the magnitude of effect when interference between units may be widespread with unknown form. David and Kempton (1996) offered interesting and sensible advice about a local form of interference in which the treatment given to one plot in an agricultural experiment can also affect

Paul R. Rosenbaum is Professor, Statistics Department, Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail:*rosenbaum@stat.wharton.upenn.edu*). This work was supported by a grant from the National Science Foundation.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000001112

adjacent plots. Although the terms "interference" and "SUTVA" are not typically used, interference that is local in time is widely discussed in crossover studies and is addressed by models for "residual effects," say the lingering effect of the treatment given in the immediately preceding time interval (see, e.g., Grizzle 1965; Kershner and Federer 1981). Methods designed to model local interference are quite useful in some settings, but interference produced by social interaction, the spread of disease, or permanent change due to learning is not typically limited to adjacent or nearby units.

1.3 A Caution: The Key Role of Random Treatment Assignment in Randomization Inference With Interference

Randomization and permutation tests are used in a wide variety of contexts, not exclusively in randomized experiments. The reader is cautioned that the discussion of randomization inference with interference in this article is limited to randomized experiments in which the randomization distribution created by the experimental design is the basis for statistical inference. This brief section elaborates and clarifies this important issue.

Many theoretical discussions and applications of permutation tests begin not with a randomized experiment, but rather with an assumption that the observations are statistically independent, perhaps with certain additional symmetries. When there is interference between units, assumptions of independence will rarely, if ever be appropriate. Recall that interference between units means that the treatment given to one unit affects other units as well, so interference tends to generate certain forms of dependence, not independence. If a permutation test is derived from an assumption of independence, and if that assumption is false, then the nominal level of the test may be substantially in error. With different types of dependence, the permutation test may have a true level that is either systematically higher or lower than the nominal level deduced from assuming independence. Gastwirth and Rubin (1971) considered the behavior of the sign test and Wilcoxon's signed rank test when derived from assumptions of independence but applied mistakenly to a stationary Gaussian process with positive autocorrelation, finding that both tests will reject true hypotheses at a rate above the nominal level; that is, .05-level tests reject $>5\%$ of the time. Hollander, Pleger, and Lin (1974) considered Wilcoxon's rank sum test when derived from assumptions of independence but applied mistakenly to data that has some pairing that induces positive dependence, finding that the test will reject true hypotheses at a rate less than or equal to the nominal level. If a permutation test is derived from an assumption of independence and that assumption is false, then both the test and confidence statements may be invalid in the most basic sense that the reported level is wrong.

The situation is different when a randomization test is correctly and solely derived from randomization actually performed in an experiment, that is, when randomization forms the basis for inference, in Fisher's (1935) phrase. In this case, potential responses under alternative treatments may exhibit arbitrary dependence or even may be fixed numbers, but the conditional distribution of treatment assignments given the potential responses is known by design to be the randomization distribution. Tests derived solely from the actual randomization have the correct level whether or not potential responses under competing treatments are independent. This has been discussed in detail by Fisher (1935, sec. 2), Welch (1937), Wilk (1955), and Cox and Reid (2000, pp. 24–27), among others, who make no assumption that potential responses are independent (see also Rosenbaum 2002a, sec. 2.4). In such an experiment, the inference is *from* the observed responses that units exhibited under the treatments to which they were randomly assigned, and the inference is *to* the unobserved responses that these same units would have exhibited under the treatments that they were randomly denied; that is, it is an inference about the effects caused by the treatments.

To emphasize, the methods in this article concern randomization inference in randomized experiments.

1.4 Some Experiments in Which Interference Is Possible

Some common types of randomized experiments with interference will now be described. A "unit" is, by definition, an opportunity to apply or withhold the treatment.

Many experiments exploit symmetries within an organism, such as two eyes, randomly assigning treatment or control to symmetric locations. As discussed in Section 4.3, Alam, Dover, and Arndt (2002) contrasted in 15 patients the pain produced by two versions of Botox, a treatment for facial wrinkles, by randomly assigning the two versions to the two sides of the face. In experiments such as this, the hope is that the treatment will affect mostly the treated location and not affect the control location, but it is certainly possible that pain experienced at one location will alter the perception of pain experienced at another location, which is interference between units. In another example, Figueroa, Schocket, Dupont, Metelitsina, and Grunwald (2004) randomly selected one eye for laser treatment, with the other eye serving as a control.

Isolating units to prevent interference is often not possible. In a justly famous experiment, in one school, Rosenthal and Jacobsen (1968, sec. 6) gave 370 students the Harvard Test of Inflected Acquisition, which was said to identify students who would experience a "significant inflection or spurt in their learning within the next year or less," but which was actually an IQ test. Then 20% of the students were selected using random numbers, and their teachers were told that these students had excelled on the test, while being "cautioned not to discuss the test finding with their pupils or the children's parents." The experiment manipulated teacher's expectations for randomly selected students. The students were retested a year later, and the randomly labeled 20% of students experienced greater gains in IQ test performance compared with the remaining 80%. The main interest is the effect of labeling a student on the student herself, but interference is possible because a change in a teacher's behavior toward one student may affect other students. To teach students in isolation to prevent interference is to change the subject matter of the experiment.

In neuroscience, the performance of cognitive tasks is related to brain activity, randomizing a few subjects to many repetitions of certain tasks. Olson, Gatenby, and Grove (2002) studied integration of auditory and visual information in 10 subjects by varying the coordination of auditory and visual speech while

monitoring brain activity using functional magnetic resonance imaging. The intent is to study the effects of the current stimulus on the current response, but in any area involving human cognition, there is the possibility of learning or adaptation, so the application of a treatment at a certain time may affect responses recorded much later.

In some cases, the concern is with the effects of a treatment on a single person. Is one person allergic to a certain food or drug? Which of two drugs is more effective for a specific person? (See McLeod, Taylor, Cohen, and Cullen 1986, Weiss et al. 1980; and Sec. 5.2). One person is observed at a sequence of times, which are the units, with treatment or control randomly assigned to each time, and interference is possible. Edgington (1987, 1996) discussed randomized single-subject experiments in psychology, emphasizing tests of no effect.

The legendary Hawthorne effect (e.g., Jones 1992) is an extreme form of interference, in which the mere knowledge a treated group affects the responses of controls.

1.5 Outline

Interference between units means that the application of the treatment to one unit affects the responses of other units. A notation that permits arbitrary interference between units is discussed in Section 2.1. Although many, if not most, randomized experiments with interference provide substantial information about treatment effects, there is one key element not identified with arbitrary interference that would have been identified without interference, namely the distinction between "no effect" and "no primary effect," which is developed in Section 2.2. There is no effect if the treatment benefits no one, whereas there is no primary effect if the treatment benefits everyone, both treated and control subjects, to exactly the same degree. In many experiments with interference, arguably including all of the experiments described in Section 1.4, the treatment is hoped to affect treated units differently than it affects control units; thus the unidentified distinction between "no effect" and "no primary effect" is a minor issue—the treatment "did not work as hoped" if it has either no effect or no primary effect, and distinguishing these hypotheses is not the central concern. Whether of concern or not, this aspect is not identified, as defined precisely in Section 2.2. The general idea is sketched in Section 3 and made tangible in two specific instances in Sections 4 and 5, with each instance illustrated with an example. The example in Section 4.3 concerns perceptions of pain, where the two units in a block are the two sides of a person's face. The example in Section 5.2 concerns randomized single-subject experiments to determine individual responsiveness to drug treatment for autism. Randomization inference with covariance adjustment is discussed in Section 6.

In a randomized experiment without interference, treated units exhibit responses typical of units exposed to the treatment, and controls exhibit responses typical of units spared exposure to the treatment; thus it is natural to think of comparing treated units and controls. Interference changes the situation. Stated informally, the responses of controls in an experiment with interference are not typical of the responses of untreated units, but are typical of untreated units who are among treated units. In this case, it is natural to ask: How is the entire scene observed in the experiment different from the scene that would have been observed had the treatment been withheld from all subjects? For instance, in the study by Rosenthal and Jacobsen mentioned in Section 1.4, we observe classrooms in which the teacher has been misled to expect improved performance from some students, and this false expectation may lead the teacher to behave in ways that affect all students. How does the scene in the actual classroom compare to the scene that would have been observed in the same classroom had no information been provided to the teacher? Stated informally, how does the entire scene in this experimental classroom differ from the entire scene in the same classroom under normal conditions, without interventions by social psychologists? In point of fact, we have no data describing the scene in which no one is treated, so the comparison seems at first infeasible. This first impression is incorrect, however. For certain statistics, specifically distribution-free statistics, the null distribution of the statistic is known a priori, with no reference to any data, so for such a statistic, we do know how it would tend to behave in the scene in which units are randomized but none are actually treated, without having to observe that scene. This intuition is formalized and demonstrated in later sections. An argument of this sort applies to distribution-free statistics but is not applicable to, say, the treated-minus-control difference in means, because although we know that its expectation would be zero in the scene in which units were randomized but never actually treated, we have no information about the variability of the responses in that same scene, so we are missing the information needed to construct the relevant null distribution. Distribution-free statistics play a unique role in randomized experiments with interference between units.

2. RANDOMIZED EXPERIMENTS WITH INTERFERENCE

2.1 Response Depends on Treatments Received by Other Units

There are $B \geq 1$ blocks, $b = 1, \ldots, B$, and $I \geq 2$ units in each block, $i = 1, \ldots, I$. In each block, J units are picked at random for treatment, $1 \leq J < I$, with the remaining $I - J$ units receiving the control, with independent assignments in distinct blocks. Write $Z_{bi} = 1$ if unit i in block b is assigned to treatment and $Z_{bi} = 0$ if this unit is assigned to control; thus $J = \sum_{i=1}^{I} Z_{bi}$ for $b = 1, \ldots, B$. Write $\mathbf{Z}_b = (Z_{b1}, \ldots, Z_{bI})$, so that there are $\binom{I}{J}$ possible values $\mathbf{z}_b$ of the random vector $\mathbf{Z}_b$. Randomization ensures that each possible $\mathbf{z}_b$ has $\Pr(\mathbf{Z}_b = \mathbf{z}_b) = \binom{I}{J}^{-1}$ and that the $\mathbf{Z}_b$, $b = 1, \ldots, B$, are mutually independent. Write $\mathbf{Z}$ for the $B \times I$ matrix containing the Z_{bi}, so the bth row of $\mathbf{Z}$ is $\mathbf{Z}_b$. There are $W = \binom{I}{J}^{B}$ possible values of the matrix $\mathbf{Z}$; collect these in a set Ω, so that $\Pr(\mathbf{Z} = \mathbf{z}) = W^{-1}$ for each $\mathbf{z} \in \Omega$.

In Fisher's (1935) theory of randomization inference, the only probability distribution used in the inference is the known random assignment of treatments $\mathbf{Z}$ created by the experimenter. Quantities that depend on the random variable $\mathbf{Z}$ are random variables. Quantities that do not depend on $\mathbf{Z}$ are fixed features of the finite population of BI experimental units. In this way, randomization forms the "reasoned basis for inference" in randomized experiments, in Fisher's phrase.

If randomization select $\mathbf{Z} = \mathbf{z}$, with $\mathbf{z} \in \Omega$, then the ith unit in block b exhibits response $r_{bi\mathbf{z}}$, so this unit has W potential responses. Only one of these W potential responses is ob-

served: the one for the actual, randomly selected treatment assignment $\mathbf{Z}$; that is, only $r_{bi\mathbf{Z}}$ is observed. Here, as in other applications of Fisher's method of randomization inference (e.g., Welch 1937, sec. 2; Wilk 1955, sec. 2.2; Cox 1958b, sec. 5), the *potential* responses, $r_{bi\mathbf{z}}$ for $\mathbf{z} \in \Omega$, do not depend on the realized random assignment of treatments, $\mathbf{Z}$, and so they are fixed features of the finite population of BI units, whereas the *observed* response from the ith unit in block b, namely $r_{bi\mathbf{Z}}$, does depend on the random variable $\mathbf{Z}$, and so $r_{bi\mathbf{Z}}$ is a random variable. The situation is simpler if units in different blocks do not interfere with one another. There is no interference between units in different blocks if $r_{bi\mathbf{z}}$ may vary with the bth row $\mathbf{z}_b$ of $\mathbf{z}$ but not with other rows of $\mathbf{z}$; in this case, there are only $\binom{I}{J}$ possible values of $r_{bi\mathbf{z}}$, not W possible values. The situation is simplest if no units interfere with one another. There is no interference between units if $r_{bi\mathbf{z}}$ may vary with z_{bi} but not with other elements of $\mathbf{z}$; in this case, there are only two possible values of $r_{bi\mathbf{z}}$, yielding the notation for causal effects discussed by Neyman (1923) and Rubin (1974).

Consider the study by Rosenthal and Jacobsen (1968, sec. 6) mentioned in Section 1.4. If there had been $B = 12$ classes, each with $I = 30$ students, $J = 6$ of whom were picked at random and labeled as about to exhibit "significant inflection" in learning, then there would be $B \times I = 12 \times 30 = 360$ students, and $W = \binom{30}{6}^{12} \doteq 1.9 \times 10^{69}$ possible ways, $\mathbf{z} \in \Omega$, to assign the treatments to these students. If there were no interference between students, then each of the 360 students would have two potential responses or final test scores, one response if labeled as ready for "significant inflection," the other if not labeled. If the labeling of one student in a classroom could affect other students in the same classroom—that is, if there may be interference between students in the same class—then each of the 360 students would have not two but $\binom{30}{6} = 593{,}775$ potential responses, depending on which students in that class are labeled. If the $B = 12$ classes were in the same school, and students in different classes talked or studied together, then perhaps the labeling of one student would affect that student's friends in other classes. If there were such interference between students in different classes, then each of the 360 students would have not two potential responses, but $W = 1.9 \times 10^{69}$ potential responses. The interference might reflect friendships and rivalries that would be difficult to measure with accuracy.

It is convenient to assume that the responses of the I distinct units within each block are free of ties; that is, for any one block b and any one treatment matrix $\mathbf{z} \in \Omega$, the I responses $r_{b1\mathbf{z}}, r_{b2\mathbf{z}}, \ldots, r_{bI\mathbf{z}}$ of *different* units are distinct, so they can be ranked 1, 2, …, I. Many other types of ties are permitted. For instance, the responses of one unit, say unit i in block b, may exhibit any pattern of ties as the treatment $\mathbf{z}$ varies over Ω. Similarly, units in different blocks may be tied.

2.2 Two Null Hypotheses: No Effect and No Primary Effect

In an experiment with interference between units, it is useful to distinguish two null hypotheses. The first is the *null hypothesis of no primary effect*, which asserts that for every unit i in every block b, the response $r_{bi\mathbf{z}}$ does not vary with $\mathbf{z}$, that is, $H_0 : r_{bi\mathbf{z}} = r_{bi\mathbf{z}'}$ for every $\mathbf{z}, \mathbf{z}' \in \Omega$, $b = 1, \ldots, B$, $i = 1, \ldots, I$. Imagine performing the randomization, selecting $\mathbf{Z}$ at random from Ω, but then covertly withholding the treatment from all units; for example, all units receive a double-blind placebo, even though random treatment assignments $\mathbf{Z}$ are made and recorded in an office. Experiments of this sort were once called "uniformity trials," and were common in the early days of randomized experimentation as aids to designing of efficient experiments (see, e.g., Cochran 1937). Then the ith unit in block b has a response, $\widetilde{r}_{bi}$, that would have been observed had the treatment been withheld from all units, that is, this unit's response in the uniformity trial. The *null hypothesis of no effect* asserts that for every unit i in every block b, the response $r_{bi\mathbf{z}}$ under every treatment $\mathbf{z} \in \Omega$ equals this unit's response in the uniformity trial, $\widetilde{H}_0 : r_{bi\mathbf{z}} = \widetilde{r}_{bi}$ for every b, i, $\mathbf{z} \in \Omega$. Both null hypotheses imply no interference between units.

If there is no effect, then there is no primary effect, but the converse is not necessarily true; that is, $\widetilde{H}_0 \Rightarrow H_0$, but $H_0 \not\Rightarrow \widetilde{H}_0$. For instance, suppose that giving the treatment to any J units in a block raises the responses of all units in the block, treated or not, by a constant, say $\tau \neq 0$; then $r_{bi\mathbf{z}} = r_{bi\mathbf{z}'} = \widetilde{r}_{bi} + \tau$ for every $\mathbf{z}, \mathbf{z}' \in \Omega$, so H_0 is true but $\widetilde{H}_0$ is false. For example, if the treatment involved exposing to a contagious agent J randomly selected members of an interacting group of I animals, then it might be the case that the contagion would spread quickly to all I members of the group, regardless of which J members were exposed. In this case, exposing any J animals might increase by $\tau > 0$ the antibody response of all I animals, so H_0 would be true but $\widetilde{H}_0$ would be false.

To illustrate the distinction between H_0 and $\widetilde{H}_0$, consider an experiment with one block. If there is no effect, if $\widetilde{H}_0$ is true, then treatment confers no benefit and does no harm to any unit. If there is no primary effect, if H_0 is true, then the treatment confers no more benefit or harm to treated units then it confers to untreated controls, but its application to any J units may affect all I units. If H_0 is true, then no benefit is gained from receiving treatment. If $\widetilde{H}_0$ is true, then no advantage is gained from being one of the J treated units rather than one of the $I - J$ controls, but benefits may be shared by all I units.

Fisher's randomization test has the correct level for testing either H_0 or $\widetilde{H}_0$; that is, an α-level test will falsely reject in at most $100\alpha\%$ of experiments if either H_0 or $\widetilde{H}_0$ is true. However, the test cannot distinguish H_0 and $\widetilde{H}_0$; that is, if $\widetilde{H}_0$ is false but H_0 is true, then the power of an α-level test of $\widetilde{H}_0$ is at most α against the alternative H_0.

In later sections, Fisher's randomization test is inverted to provide confidence statements about measures of effect in the presence of interference, but the test still cannot distinguish H_0 and $\widetilde{H}_0$. The measures ask to what extent did the J treated units receive benefits or harms not shared with the $I - J$ untreated control units? The study by Rosenthal and Jacobsen (1968, sec. 6) asked to what extent did labeling $J = 6$ students in a class as ready for "significant inflection" cause these $J = 6$ labeled students to experience greater gains in IQ than did the $I - J = 24$ students who were not so labeled. If labeling $J = 6$ students in one class benefits all $I = 30$ students in the class equally, then there is no way to tell that this has happened or to measure the extent to which it has happened. However, if the $J = 6$ labeled students benefit more than the $I - J = 24$ unlabeled students, then the confidence statements can measure this.

3. CONFIDENCE STATEMENTS FOR EFFECTS WITH INTERFERENCE

I sketch and develop the general approach in two specific forms in Sections 4 and 5. Certain statistics, D, such as the Mann–Whitney–Wilcoxon statistic, have randomization distributions under the null hypothesis that may be determined without reference to the data (e.g., Lehmann 1998, sec. 1). This is not true of all randomization tests; for instance, it is not true of the permutational t-test. Statistics with this property are essentially the distribution-free statistics, although technically the definitions diverge slightly (cf. Randles and Wolfe 1979, p. 30), because one refers to probability distributions and the other refers to randomization and the hypothesis of no effect. With this caution stated, I call statistics D with this property *distribution-free* statistics.

In a uniformity trial in Section 2.2, randomization covertly identifies certain units as treated and others as control, but in fact no treatment of any kind is applied to any unit, so the null hypothesis of no effect is true. We know the distribution of a distribution-free statistic D in the uniformity trial in Section 2.2 without having to perform the uniformity trial. This will turn out to be convenient. It is possible to compare the behavior of D actually observed in an experiment with an active treatment to its behavior in the uniformity trial while performing only the experiment with an active treatment.

Certain statistics, such as the Mann–Whitney–Wilcoxon statistic, are easy to interpret as measures of the magnitude of effect, perhaps after dividing the statistic by a function of the sample size; see Section 4. Wolfe and Hogg (1971) discussed a variety of distribution-free statistics with this property. Whether or not there is interference between units, we know how D behaved in the actual experiment, and because D is distribution-free, we know how it would have behaved in the uniformity trial, and D is easy to interpret as a measure of magnitude of effect. Only a few details separate us from a confidence statement for magnitude of the effect of the treatment despite interference between units.

4. MANN–WHITNEY STATISTICS WITH INTERFERENCE

4.1 Comparing an Experiment and a Uniformity Trial

The Mann–Whitney statistic counts the number of times that a treated unit had a higher response than a control in the same block. In simple problems involving independent and identically distributed samples from two continuous populations, the statistic may be rescaled to estimate the probability that a treated response will exceed a control response (Lehmann 1998, sec. 2). In randomized experiments without interference, the null randomization distribution may be inverted to obtain exact randomization inferences about the number of such exceedances caused by effects of the treatment (Rosenbaum 2001). This approach extends to studies with interference, as we now discuss. The statistic F, defined later, is the difference of two Mann–Whitney statistics, one for the actual experiment and the other for the uniformity trial with no effect; thus it asks: In an experiment with interference, is there a greater tendency for treated subjects to have higher responses than controls than would have been seen in a uniformity trial with no effect?

For each $\mathbf{z} \in \Omega$, write $f_{bij\mathbf{z}} = 1$ if $r_{bi\mathbf{z}} > r_{bj\mathbf{z}}$ and $f_{bij\mathbf{z}} = 0$ otherwise, noting that this definition implies that $f_{bii\mathbf{z}} = 0$ for every i, so that $q_{bi\mathbf{z}} = 1 + \sum_{j=1}^{I} f_{bij\mathbf{z}}$ is the rank of $r_{bi\mathbf{z}}$ among $r_{b1\mathbf{z}}, r_{b2\mathbf{z}}, \ldots, r_{bI\mathbf{z}}$. The $f_{bij\mathbf{z}}$ and the $q_{bi\mathbf{z}}$ depend on the potential responses, $r_{bi\mathbf{z}}$ for $\mathbf{z} \in \Omega$, but not on the realized random assignment, $\mathbf{Z}$, so $f_{bij\mathbf{z}}$ and $q_{bi\mathbf{z}}$ are fixed features of the finite population of BI units. Of course, randomization picks at random a value $\mathbf{Z} \in \Omega$ yielding realized values, $f_{bij\mathbf{Z}}$ and $q_{bi\mathbf{Z}}$, of $f_{bij\mathbf{z}}$ and $q_{bi\mathbf{z}}$; these realized values, $f_{bij\mathbf{Z}}$ and $q_{bi\mathbf{Z}}$, are random variables because they do depend on $\mathbf{Z}$.

The statistic

$$T = \sum_{b=1}^{B}\sum_{i=1}^{I}\sum_{j=1}^{I} Z_{bi}\,(1-Z_{bj})\,f_{bij\mathbf{Z}} = \left(\sum_{b=1}^{B}\sum_{i=1}^{I} Z_{bi}\,q_{bi\mathbf{Z}}\right) - BJ(J+1)/2$$

is the number of times that a treated unit had a higher response than a control in the same block; that is, T is the sum of B Mann–Whitney statistics, one for each block, and $T + BJ(J+1)/2$ is the sum of B Wilcoxon rank sum statistics. The statistic T takes integer values from 0 to $BJ(I-J)$. If the treatment had no effect or no primary effect, then $E(T) = BJ(I-J)/2$. If the treatment increased the responses of all treated units above the responses of all controls, then $T = BJ(I-J)$.

In parallel, for the uniformity trial, write $\widetilde{f}_{bij} = 1$ if $\widetilde{r}_{bi} > \widetilde{r}_{bj}$ and $\widetilde{f}_{bij} = 0$ otherwise, and write $\widetilde{q}_{bi} = 1 + \sum_{j=1}^{I} \widetilde{f}_{bij}$. Then $\widetilde{T} = \sum_{b=1}^{B}\sum_{i=1}^{I}\sum_{j=1}^{I} Z_{bi}(1-Z_{bj})\widetilde{f}_{bij} = (\sum_{b=1}^{B}\sum_{i=1}^{I} Z_{bi}\widetilde{q}_{bi}) - BJ(J+1)/2$ is the sum of the B Mann–Whitney statistics that would have been observed in the uniformity trial. Even though the uniformity trial was never performed, the randomization distribution of $\widetilde{T}$ is known, because the null hypothesis of no effect is true in the uniformity trial, and $\widetilde{T}$ has the familiar null randomization distribution of the sum of B independent Mann–Whitney statistics. The statistic $\widetilde{T}$ also takes integer values from 0 to $BJ(I-J)$, with expectation $E(\widetilde{T}) = BJ(I-J)/2$ and variance $\text{var}(\widetilde{T}) = BJ(I-J)(I+1)/12$. Let c_α be the integer such that $\Pr(\widetilde{T} \geq c_\alpha) = \alpha$.

Consider $f_{bij\mathbf{z}} - \widetilde{f}_{bij}$, contrasting experimental assignment $\mathbf{z} \in \Omega$ to the uniformity trial,

$$f_{bij\mathbf{z}} - \widetilde{f}_{bij} = \begin{cases} 1 & \text{if } r_{bi\mathbf{z}} > r_{bj\mathbf{z}}, \widetilde{r}_{bi} \leq \widetilde{r}_{bj} \\ -1 & \text{if } r_{bi\mathbf{z}} \leq r_{bj\mathbf{z}}, \widetilde{r}_{bi} > \widetilde{r}_{bj} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

In the first line of (1), in block b, unit i would have a higher response than unit j under treatment $\mathbf{z} \in \Omega$ but i would have a response that was not higher than j in the uniformity trial in which the treatment was withheld. Conversely, in the second line of (1), in block b, unit i would have a response not higher than j under treatment assignment $\mathbf{z} \in \Omega$, but i would have a higher response than j in the uniformity trial. In all other cases, in block b, the relative order of the responses of i and j would be the same under treatment $\mathbf{z}$ as in the uniformity trial. Trivially, $f_{bii\mathbf{z}} - \widetilde{f}_{bii} = 0$. If the treatment has no effect, then $f_{bij\mathbf{z}} - \widetilde{f}_{bij} = 0$ for all b, i, j.

The quantity

$$F = \sum_{b=1}^{B}\sum_{i=1}^{I}\sum_{j=1}^{I} Z_{bi}\,(1-Z_{bj})(f_{bij\mathbf{Z}} - \widetilde{f}_{bij}) = T - \widetilde{T}$$

is an unobservable random variable comparing the relative position of responses of treated units and controls under the randomly selected treatment $\mathbf{Z} \in \Omega$ and in the uniformity trial. Specifically, F is the number of times that treated responses exceeded control responses in the actual experiment but not the uniformity trial, less the number of times that treated responses exceeded control responses in the uniformity trial but not in the actual experiment. In general, $F = T - \widetilde{T}$ can take integer values from $-BJ(I-J)$ to $BJ(I-J)$. If the treatment has no effect, then $F = 0$.

Although F is an unobserved random variable, a confidence statement about its magnitude may be made using observed results (cf. Weiss 1955; Rosenbaum 2001, sec. 4). Specifically, $F \geq T - c_\alpha + 1$ with confidence $1 - \alpha$, as demonstrated in Proposition 1.

Proposition 1. $\Pr(F \geq T - c_\alpha + 1) = 1 - \alpha$.

Proof. Because $\widetilde{T}$, F, and c_α are integers and $F - T = -\widetilde{T}$,

$$\Pr(F \geq T - c_\alpha + 1) = \Pr(-\widetilde{T} \geq -c_\alpha + 1) = \Pr(\widetilde{T} \leq c_\alpha - 1)$$
$$= 1 - \Pr(\widetilde{T} \geq c_\alpha) = 1 - \alpha.$$

The quantity F depends on the sample size, so it may aid interpretation to scale it by dividing by the constant $BJ(I-J)/2$, so that $V = 2F/\{BJ(I-J)\}$ has minimum value -2 and maximum value 2. Then V is 0 if the treatment has no effect, and V has expectation 0 if the treatment has no primary effect. If the treatment increases the responses of all treated units above the responses of all controls, then V has expectation 1, and $V \xrightarrow{P} 1$ as $B \to \infty$, whereas if the treatment decreases the responses of all treated units below all controls, then V has expectation -1 and $V \xrightarrow{P} -1$ as $B \to \infty$. From Proposition 1, with confidence $1 - \alpha$, assert that, $V \geq 2(T - c_\alpha + 1)/\{BJ(I-J)\}$.

4.2 Numerical Illustration

To clarify notation, consider just one block, $B = 1$, $I = 3$ units, with one unit picked at random for treatment, $J = 1$, so that $1 = \sum_{i=1}^{3} Z_{1i}$. These three units would yield responses $(\widetilde{r}_{11}, \widetilde{r}_{12}, \widetilde{r}_{13}) = (7, 2, 1)$ in the uniformity trial no matter which unit was randomly picked. The actual treatment benefits all three units, but the largest benefits accrue to the one unit which received the treatment. If the first unit, $i = 1$, is treated, $\mathbf{z} = (1, 0, 0)$, then the responses are $(r_{11\mathbf{z}}, r_{12\mathbf{z}}, r_{13\mathbf{z}}) = (15, 3, 2)$, so the treated unit benefits by $15 - 7 = 8$ and both controls benefit by $3 - 2 = 2 - 1 = 1$; however, in this case, $f_{bij\mathbf{z}} - \widetilde{f}_{bij} = 0$ for all i, j, because unit $i = 1$ would have had the highest response in both the uniformity trial and the actual experiment. If the second unit, $i = 2$, is treated, $\mathbf{z} = (0, 1, 0)$, then the responses are $(r_{11\mathbf{z}}, r_{12\mathbf{z}}, r_{13\mathbf{z}}) = (9, 25, 2)$, so the treated unit benefits by $25 - 2 = 23$, and the two controls benefit by $9 - 7 = 2$ and $2 - 1 = 1$; then $f_{b21\mathbf{z}} - \widetilde{f}_{b21} = 1 - 0 = 1$, $f_{b12\mathbf{z}} - \widetilde{f}_{b12} = 0 - 1 = -1$, and the other $f_{bij\mathbf{z}} - \widetilde{f}_{bij} = 0$, because the treatment reverses the ordering of responses of units $i = 1$ and $j = 2$. If the third unit, $i = 3$, is treated, $\mathbf{z} = (0, 0, 1)$, then the responses are $(r_{11\mathbf{z}}, r_{12\mathbf{z}}, r_{13\mathbf{z}}) = (8, 6, 22)$, so the treated unit benefited by $22 - 1 = 21$, and the two controls benefit by $8 - 7 = 1$ and $6 - 2 = 4$; then $f_{b31\mathbf{z}} - \widetilde{f}_{b31} = f_{b32\mathbf{z}} - \widetilde{f}_{b32} = 1 - 0 = 1$, $f_{b13\mathbf{z}} - \widetilde{f}_{b13} = f_{b23\mathbf{z}} - \widetilde{f}_{b23} = 0 - 1 = -1$, and the other $f_{bij\mathbf{z}} - \widetilde{f}_{bij} = 0$, because the treatment makes the treated response the highest but it would have been the lowest in the uniformity trial. The r's and f's are fixed potential responses under the three possible treatment assignments. Random assignment picks $\mathbf{Z} = (1, 0, 0)$ or $\mathbf{Z} = (0, 1, 0)$ or $\mathbf{Z} = (0, 0, 1)$ at random, each with probability $\frac{1}{3}$, so F takes values 0, 1, or 2, each with probability $\frac{1}{3}$, saying that if $\mathbf{Z} = (1, 0, 0)$, then the response ordering is unchanged; if $\mathbf{Z} = (0, 1, 0)$, then the treatment causes one inversion; and if $\mathbf{Z} = (0, 0, 1)$, then the treatment causes two inversions. Because $B = 1$, $I = 3$, $J = 2$, the scaled version of F, namely $2F/\{BJ(I-J)\}$, equals F itself, and its expectation is 1, which is the largest possible expectation for $2F/\{BJ(I-J)\}$, signifying that the treatment always yields treated responses above all control responses. In other words, in the actual experiment, the Mann–Whitney statistic is $T = 2$ for all three treatment assignments, but in the uniformity trial the Mann–Whitney statistic $\widetilde{T}$ has its usual null distribution, taking values $\widetilde{T} = 2$ if $\mathbf{Z} = (1, 0, 0)$, $\widetilde{T} = 1$ if $\mathbf{Z} = (0, 1, 0)$, or $\widetilde{T} = 0$ if $\mathbf{Z} = (0, 0, 1)$, so $F = T - \widetilde{T}$ takes values 0, 1, or 2 each with probability $\frac{1}{3}$.

4.3 Example: Wrinkles, Preservatives, and Pain

Botulinum A exotoxin (or botox) is an injected drug used to treat wrinkles. The treatment is apparently quite painful. The manufacturer, Allergan, is said to recommend reconstituting the desiccated form of the drug using a preservative-free saline solution because this may make treatment less painful. The addition of a preservative might allow unused reconstituted drug to be stored for weeks rather than have to be used within a day or discarded, possibly reducing waste and cost. Alam et al. (2002) investigated this in a randomized, double-blind experiment comparing two methods of reconstituting the drug: a preservative-free saline solution and a saline solution containing benzyl alcohol as a preservative. Benzyl alcohol also may have anesthetic properties. In this study, 15 patients gave informed consent, and each patient received one solution on the left side of the face and the other solution on the right side, with solutions randomly assigned to sides. Neither the patients nor the investigators directly involved with the patients knew which solution was which. Patients were asked to evaluate the level of pain experienced with injections on each side of the face.

In this study, the $B = 15$ blocks are the patients and the $I = 2$ units in each block are the two sides of the patient's face, with $J = 1$ unit assigned to treatment and $I - J = 1$ unit assigned to control. To be definite, "treatment" and "control" refer to the solutions without and with benzyl alcohol. Interference between units in the same block—different sides of the same face—is very likely. A painful injection on one side of the face is quite likely to alter the perception of a pain on the other side.

All $B = 15$ patients reported less pain on the side that included the preservative benzyl alcohol, possibly because of its anesthetic properties. With $J = I - J = 1$, the blocked Mann–Whitney statistic T of Section 4 equals the sign statistic, and $T = 15$. The unobserved $\widetilde{T}$ is the sign statistic that would have been observed in the uniformity trial in which face sides were randomized but the same solution was applied on both sides; that is, face sides were randomly labeled in an office without in any way altering the treatments actually applied to patients. The randomization distribution of $\widetilde{T}$ is binomial with sample size

$B = 15$, probability of success $\frac{1}{2}$, and $\Pr(\widetilde{T} \geq 15) = .000031$, so the hypotheses of no effect and of no primary effect are both rejected with one-sided significance level .000031.

The quantity $F = T - \widetilde{T}$ compares the observed experiment and the unobserved uniformity trial. Informally, F is a measure of the effect on pain comparisons of assigning some units to treatment and others to control, rather than letting the randomization remain a mere label in an office and assigning all units to control. More precisely, F is the number of patients who experienced more pain on the side of the face without benzyl alcohol because of effects caused by removing benzyl alcohol from the "control" solution. Using $\Pr(\widetilde{T} \geq 11) = .0176$ from the binomial distribution and Proposition 1 yields $\Pr(F \geq T - c_\alpha + 1) = \Pr(F \geq T - 11 + 1) = 1 - \Pr(\widetilde{T} \geq 11) = 1 - .0176 = .982$; thus, with 98.2% confidence, $F \geq 15 - 11 + 1 = 5$. Although all $B = 15$ patients reported less pain for the face side with alcohol, so $T = 15$, this would be expected for $BJ(I-J)/2 = 15 \times 1 \times 1/2 = 7.5$ faces by chance; however, we are 98.2% confident that 5 of the favorable results were effects of alcohol. Alternatively, with 98.2% confidence, $V = 2F/\{BJ(I-J)\} = F/7.5$ is at least $5/7.5 = 2/3$. These confidence statements are correct whether or not there is interference between sides of the same face (a likely possibility) or between faces of different people (an unlikely possibility).

5. EXCEEDANCE MEASURES WITH INTERFERENCE

5.1 Treated Responses Above the Control Median

The statistic of Mathisen (1943), Gart (1963), and Gastwirth (1968) counts the number responses observed under treatment that exceed the median observed under control; it is a distribution-free statistic that is not a linear rank statistic. Gastwirth and Wang (1987) proposed a symmetric version of this statistic. (See Randles and Wolfe 1979 for a textbook discussion and Chakraborti and van der Laan 1996 for a recent survey.) The randomization test of no effect using this statistic, and similar statistics for other quantiles, also may be inverted to yield confidence statements about the magnitude of effect when units interfere with one another.

Write $\widetilde{r}_b = (\widetilde{r}_{b1}, \ldots, \widetilde{r}_{bI})$ for responses in block b in the uniformity trial, and for each $\mathbf{z} \in \Omega$, write $r_{b\mathbf{z}} = (r_{b1\mathbf{z}}, \ldots, r_{bI\mathbf{z}})$ for the responses in block b that would be observed with assignment $\mathbf{z} \in \Omega$ in the actual experiment. Pick an integer k with $1 \leq k \leq I - J$ and define $q(\cdot, \cdot)$ to be the function that returns the kth-order statistic of the $I - J$ control responses observed in a block, so that for each $\mathbf{z} \in \Omega$, the quantities $q(\mathbf{z}_b, \widetilde{r}_b)$ and $q(\mathbf{z}_b, r_{b\mathbf{z}})$ are the kth largest control response in block b in the uniformity trial and the actual experiment with assignment $\mathbf{z} \in \Omega$. For instance, if $I - J$ is odd and $k = (I - J + 1)/2$, then $q(\mathbf{z}_b, \widetilde{r}_b)$ and $q(\mathbf{z}_b, r_{b\mathbf{z}})$ are the medians of the $I - J$ control responses in block b under treatment assignment $\mathbf{z} \in \Omega$. Write

$$s_{bi\mathbf{z}} = \begin{cases} 1 & \text{if } z_{bi} = 1 \text{ and } r_{bi\mathbf{z}} > q(\mathbf{z}_b, r_{b\mathbf{z}}) \\ & \quad \text{and } \widetilde{r}_{bi} < q(\mathbf{z}_b, \widetilde{r}_b) \\ -1 & \text{if } z_{bi} = 1 \text{ and } r_{bi\mathbf{z}} < q(\mathbf{z}_b, r_{b\mathbf{z}}) \\ & \quad \text{and } \widetilde{r}_{bi} > q(\mathbf{z}_b, \widetilde{r}_b) \\ 0 & \text{otherwise.} \end{cases}$$

In words, with treatment assignment $\mathbf{z} \in \Omega$, a treated unit, say the ith unit in block b with $z_{bi} = 1$, has a step up, $s_{bi\mathbf{z}} = 1$, if this unit's response in the actual experiment would exceed the median of control responses in the same block in the experiment but the unit's response in the uniformity trial would have been below the control median in block b in the uniformity trial. A step down, $s_{bi\mathbf{z}} = -1$, is defined analogously. One measure of the magnitude of the treatment effect is the net number, S, of steps up minus steps down, $S = \sum_{b=1}^{B} \sum_{i=1}^{I} s_{bi\mathbf{Z}}$. Here S is an unobserved random variable, because the results in the uniformity trial are not observed. Nonetheless, a confidence statement about S is possible.

Define the function $h(\cdot, \cdot)$ to count the number of times that treated responses in block b exceed the kth order statistic of control responses in block b. That is, in the actual experiment, $h(\mathbf{z}_b, r_{b\mathbf{z}})$ is the number of times that a treated response, $r_{bi\mathbf{z}}$ with $z_{bi} = 1$, exceeds $q(\mathbf{z}_b, r_{b\mathbf{z}})$. Then $h(\mathbf{z}_b, r_{b\mathbf{z}})$ is the statistic used in the control quantile tests of Mathisen (1943), Gart (1963), and Gastwirth (1968). In parallel, in the uniformity trial, $h(\mathbf{z}_b, \widetilde{r}_b)$ is the number of times that a treated response, $\widetilde{r}_{bi}$ with $z_{bi} = 1$, exceeds $q(\mathbf{z}_b, \widetilde{r}_b)$. Write $H = \sum_{b=1}^{B} h(\mathbf{z}_b, r_{b\mathbf{z}})$ and $\widetilde{H} = \sum_{b=1}^{B} h(\mathbf{z}_b, \widetilde{r}_b)$.

Note that $S = H - \widetilde{H}$, where H may be calculated from the observed data in the experiment but $\widetilde{H}$ is from the uniformity trial and is not observed. The randomization distribution of $h(\mathbf{Z}_b, \widetilde{r}_b)$ is given by

$$\Pr\{h(\mathbf{Z}_b, \widetilde{r}_b) = \ell\} = \frac{\binom{I-J-k+\ell}{\ell}\binom{k-1+J-\ell}{J-\ell}}{\binom{I}{J}} \quad \text{for } \ell = 0, \ldots, J \qquad (2)$$

[see Fligner and Wolfe 1976, cor. 4.1, for expression (2)]. Using Fligner and Wolfe's theorem 4.3, the expectation and variance of $h(\mathbf{Z}_b, \widetilde{r}_b)$ are

$$E\{h(\mathbf{Z}_b, \widetilde{r}_b)\} = J\left(1 - \frac{k}{I - J + 1}\right) = \mu, \quad \text{say,}$$

and

$$\operatorname{var}\{h(\mathbf{Z}_b, \widetilde{r}_b)\} = \frac{Jk(I - J - k + 1)(I + 1)}{(I - J + 1)^2 (I - J + 2)} = \nu, \quad \text{say.}$$

Let d_α be the integer such that $\Pr(\widetilde{H} \geq d_\alpha) = \alpha$. Because randomization ensures that the $\mathbf{z}_b$, $b = 1, \ldots, B$, are mutually independent, the randomization distribution of $\widetilde{H}$ is the B-fold convolution of (2). For small B, the distribution of $\widetilde{H}$ may be determined exactly using (2) with, for instance, the aid of the Bth power of the probability-generating function (Feller 1968, sec. XI), whereas for large B, the central limit theorem and a continuity correction yield the approximation $d_\alpha \doteq B\mu + \frac{1}{2} + \Phi^{-1}(1 - \alpha)\sqrt{B\nu}$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution. Proposition 2 provides a confidence statement for the unobserved random variable $S = H - \widetilde{H}$. The proof parallels the proof of Proposition 1 and is omitted here.

Proposition 2. $\Pr(S \geq H - d_\alpha + 1) = 1 - \alpha$.

5.2 Example: Treating Autism

Linday, Tsiouris, Cohen, Shindledecker, and DeCresce (2001) examined possible effects of the drug famotidine in randomized, double-blind, single-subject experiments on nine autistic boys, all living at home with their families. On entry

Table 1. Affection Scores for the First Autistic Boy in a Randomized Single-Subject Trial

	Periods	Minimum	Median	Maximum
Famotidine	22	65.0	70.0	77.5
Placebo	19	55.5	63.0	78.0

Source: Linday et al. (2001).

into the study, primary caregivers (mostly mothers, but in one case an aunt) completed an "Aberrant Behavior Checklist" and a "Clinical Global Impressions Scale," and on this basis, one symptom for each child was designated as the target symptom. For the first boy, this symptom was "affection;" for the second boy it was "calmness." Primary caregivers (e.g., mothers) kept a daily diary recording symptom levels on visual analog scale, which is similar to placing a mark on a printed image of a ruler, with the response recorded in millimeters. Each child was randomized to double-blind periods of placebo or famotidine, and randomization inferences were performed. Specifically, as described in Section 5.1, Linday et al. reported the number of times that each the boy's target symptom was better with famotidine than his median response under placebo. (Although the study was randomized and performed a randomization inference, their randomization inference is not quite the appropriate one for the randomization performed. As emphasized in Section 1.3, this is not a trivial issue either in principle or in practice. However, for the purpose of an illustrative example, I ignore this discrepancy between design and analysis in the example that follows.)

In his first double-blind trial, the first boy was observed for $I = 41$ periods, with $J = 22$ periods under famotidine and $I - J = 19$ periods under placebo. All of his targeted "affection" scores were better under famotidine than the median score under placebo, so $H = 22$ (Table 1). With a single subject, $B = 1$, and with $k = 10$ for the median of the $I - J = 19$ control periods, expression (2) yields $\Pr(\widetilde{H} \geq 17) = .0484$, so $d_{.0484} = 17$. Then Proposition 2 yields $\Pr(S \geq H - 17 + 1) = 1 - .0484 = .9516$, so, with 95% confidence, $S \geq 22 - 17 + 1 = 6$. By chance, 11 of the $J = 22$ treated responses are expected to exceed the median of the $I - J = 19$ control responses, so $H = 22$ is 11 more than expected by chance, and with 95% confidence, at least 6 are among the S effects of the treatment not present in the uniformity trial. To repeat, this inference makes no assumption that there is no interference between treatments given to this boy in different time periods.

The exact calculation of $d_{.0484} = 17$ may be compared with the large-sample normal approximation. Here $B = 1$, $\mu = 11$, and $v = 11$, so $d_{.05}$ is approximated as $B\mu + \frac{1}{2} + \Phi^{-1}(1 - \alpha)\sqrt{Bv}$ or $11 + \frac{1}{2} + 1.645\sqrt{11} = 16.96$, whereas $d_{.0484} = 17$ is approximated by $11 + \frac{1}{2} + 1.661\sqrt{11} = 17.01$.

6. COVARIANCE ADJUSTMENT

A covariate $\mathbf{x}_{bi}$ describing unit i in block b is a variable measured before treatment, and hence a variable unaffected by the treatment. For instance, in Section 5.2 the units of the trial are the $I = 41$ time periods for boy $b = 1$, and the period number, $x_{1i} = i, = 1, 2, \ldots, 41$, is a covariate. Random assignment of treatments tends to balance covariates, but some imbalances may occur by chance; for instance, by the luck of the randomization in Section 5.2, the $J = 22$ treated periods for the first boy, $b = 1$, might have tended to be slightly earlier than average, so $(\sum_{i=1}^{I} z_{1i} x_i)/22 < (\sum_{i=1}^{I} x_i)/41 = (41 + 1)/2 = 22$. Covariance adjustment is often used in randomized experiments to eliminate or reduce the impact of chance imbalances in covariates. Can covariance adjustment be used in randomized experiments with interference?

Figure 1 is a small simulated illustration, solely intended to make the following discussion more tangible. The experiment has $I = 100$ time periods, $i = 1, \ldots, I$, with $J = 50$ selected at random for treatment, $Z_i = 1$, and the remaining $I - J = 50$ periods receiving control, $Z_i = 0$. There is the simplest form of interference: Responses are enhanced by treatment in the current period and depressed by treatment in the previous period, so the effect in period i is $(3 \cdot Z_i) - (1.5 \cdot Z_{i-1})$, with Z_0 defined as 0. The covariate is time, $x_i = i$, and the responses are a quadratic in time, plus standard normal errors, plus the treatment effect $3Z_i - 1.5Z_{i-1}$. The curve in the upper left panel in Figure 1 is the lowess smooth of responses on x_i ignoring treatment group (Cleveland and Devlin 1988, as implemented in R), and the plot in the upper right panel displays the residuals from the smooth. The boxplots compare the responses and residuals in the treated and control groups. For a given x, treated units tend to have higher responses than controls, but when x is ignored, this is somewhat obscured.

In a randomized experiment, an exact covariance-adjusted randomization test of the null hypothesis of no effect is obtained by applying a randomization test of no treatment effect to the residuals of the responses obtained from some form of regression of the responses on the covariates (see Rosenbaum 2002b for general discussion and Raz 1990 for randomization inferences with a smoother). This test is a randomization test; it requires no distributional assumptions for the responses and no assumption that the regression model generated the data, but it does require the actual use of randomization in assigning treatments. In brief, whether on the right side or the left side of Figure 1, if the treatment had no effect, then randomization will simply pick 50 points of 100 points at random and call them treated points; thus randomization creates the null distribution for the test statistic whether applied to responses or residuals. For instance, one could apply either the stratified Mann–Whitney test T in Section 4 or the control median statistic H in Section 5 to the residuals to test the null hypothesis of no effect, comparing these with their usual randomization distributions. If this were done in the uniformity trial, then $\widetilde{T}$ and $\widetilde{H}$ would have their usual null randomization distribution. Confidence statements about $F = T - \widetilde{T}$ and $S = H - \widetilde{H}$ are constructed as in Sections 4 and 5, and these statements now describe the degree to which the residuals of the response track the treatment.

In Figure 1, for the responses themselves in the left panel, $H = 42$ of the $J = 50$ treated units had responses above the control median, defined as the 25th of the $I - J = 50$ order statistics for controls. In contrast, for the residuals in the right panel, all $H = 50$ of the treated units had residuals above the control median. To apply the control median procedure in a randomized experiment to either the responses or the residuals, note that from (2), $\Pr(\widetilde{H} \geq 35) = .0328$, so $\Pr(S \geq H - 35 + 1) = 1 - .0328 = 97\%$. That is, $H = 42$ treated units had responses

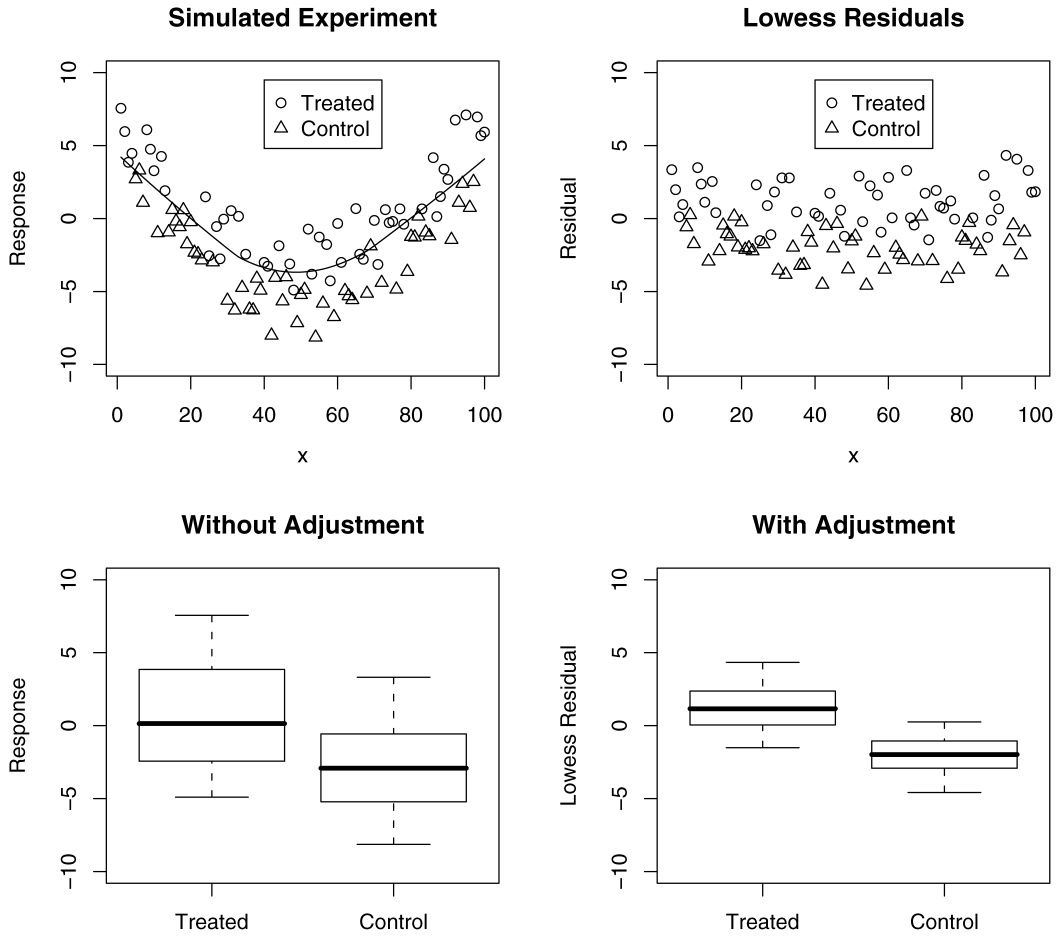

Figure 1. The Simulated Experiment and Its Lowess Residuals. The responses in the experiment are (a) a quadratic in time $x = i$, $i = 1, \ldots, I = 100$, (b) standard normal errors, (c) $J = 50$ treated periods, $Z_i = 1$, picked at random, and (d) a treatment effect that adds $3Z_i - 1.5Z_{i-1}$ to the ith response, with $Z_0 = 0$.

above the control median, and we are 97% confident that this is a net increase of at least $8 = 42 - 35 + 1$ above what would have been observed in the uniformity trial with no effect, or an increase of $8/25 = 32\%$ over the number, 25, expected by chance. For the residuals, all $H = 50$ of the treated units had residuals above the control median, and we are 97% confident that this is a net increase of at least $16 = 50 - 35 + 1$ above what would have been observed in the uniformity trial with no effect, or an increase of $16/25 = 64\%$ over the number, 25, expected by chance. In a randomized experiment, these exact confidence statements are correct as randomization-based confidence statements about $S = H - \widetilde{H}$, for responses or residuals, despite the interference between units and despite the use of lowess rather than the correct quadratic to make the adjustment.

7. SUMMARY

If there is interference between units in a randomized experiment, then a treatment applied to one unit may affect the responses of other units; see Section 2.1. Interference between units does not invalidate the level of a randomization test of the null hypothesis of no treatment effect, but interference does limit what can be said about the magnitude of the treatment effect, in the following specific sense. If all units, both treated and control, benefit equally from applying the treatment to the treated units, then there is no primary effect, and it is not possible to distinguish no primary effect from no effect at all, unless additional information is somehow brought in from outside the randomized experiment; see Section 2.2.

In many contexts, however, the main interest lies is benefits that accrue to treated units that are not shared by control units even though, to some degree, the treatment affects both treated and control units; see the examples in Sections 1.4, 4.3, and 5.2. As shown in Sections 4, 5, and 6, it is possible to invert distribution-free randomization tests of no effect to obtain confidence statements about the magnitude of benefit's that accrue to treated units that are not shared with controls. Inverting the randomization test yields a confidence interval for an unobserved random variable whose value measures the magnitude of the treatment effect. In this way, randomization forms the basis for inference in randomized experiments with interference, with no assumptions about the form of the interference required.

[Received March 2005. Revised July 2006.]

REFERENCES

Alam, M., Dover, J. S., and Arndt, K. A. (2002), "Pain Associated With Injection of Botulinum A Exotoxin Reconstituted Using Isotonic Sodium Chloride With and Without Preservative," *Archives of Dermatology*, 138, 510–514.

Chakraborti, S., and van der Laan, P. (1996), "Precedence Tests and Confidence Bounds for Complete Data: An Overview and Some Results," *The Statistician*, 45, 351–369.

Cleveland, W. S., and Devlin, S. J. (1988), "Locally Weighted Regression," *Journal of the American Statistical Association*, 83, 596–610.
Cochran, W. G. (1937), "A Catalogue of Uniformity Trial Data," *Supplement to the Journal of the Royal Statistical Society*, 4, 233–253.
Cox, D. R. (1958a), *Planning of Experiments*, New York: Wiley.
——— (1958b), "The Interpretation of the Effects of Nonadditivity in the Latin Square," *Biometrika*, 45, 69–73.
Cox, D. R., and Reid, N. (2000), *Theory of the Design of Experiments*, New York: Chapman & Hall/CRC.
David, O., and Kempton, R. A. (1996), "Designs for Interference," *Biometrics*, 52, 597–606.
Edgington, E. S. (1987), "Randomized Single-Subject Experiments and Statistical Tests," *Journal of Counseling Psychology*, 34, 437–442.
——— (1996), "Randomized Single-Subject Experimental Designs," *Behaviour Research and Therapy*, 34, 567–574.
Feller, W. (1968), *Introduction to Probability and Its Applications*, Vol. I, New York: Wiley.
Figueroa, M., Schocket, L. S., Dupont, J., Metelitsina, T. I., and Grunwald, J. E. (2004), "Effect of Laser Treatment for Dry Age-Related Macular Degeneration on Foveolar Choroidal Haemodynamics," *British Journal of Opthalmology*, 88, 792–795.
Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.
Fligner, M. A., and Wolfe, D. A. (1976), "Some Applications of Sample Analogues to the Probability Integral Transformation and a Coverage Property," *The American Statistician*, 30, 78–85.
Gart, J. J. (1963), "A Median Test With Sequential Application," *Biometrika*, 50, 55–62.
Gastwirth, J. L. (1968), "The First-Median Test: A Two-Sided Version of the Control Median Test," *Journal of the American Statistical Association*, 63, 692–706.
Gastwirth, J. L., and Rubin, H. (1971), "Effect of Dependence on the Level of Some One-Sample Tests," *Journal of the American Statistical Association*, 66, 816–820.
Gastwirth, J. L., and Wang, J.-L. (1987), "Nonparametric Tests in Small Unbalanced Samples," *Canadian Journal of Statistics*, 15, 339–348.
Grizzle, J. E. (1965), "Two-Period Change-Over Design and Its Use in Clinical Trials," *Biometrics*, 21, 467–480.
Hollander, M., Pledger, G., and Lin, P.-E. (1974), "Robustness of the Wilcoxon Test to a Certain Dependency Between Samples," *The Annals of Statistics*, 2, 177–181.
Jones, S. R. G. (1992), "Was There a Hawthorne Effect?" *American Journal of Sociology*, 98, 451–468
Kershner, R., and Federer, W. (1981), "Two-Treatment Crossover Designs for Estimating a Variety of Effects," *Journal of the American Statistical Association*, 76, 612–619.
Lehmann, E. L. (1998), *Nonparametrics: Statistical Methods Based on Ranks* (rev. ed.), Upper Saddle River, NJ: Prentice-Hall.
Linday, L. A., Tsiouris, J. A., Cohen, I. L., Shindledecker, R., and DeCresce, R. (2001), "Famotidine Treatment of Children With Autistic Spectrum Disorders," *Journal of Neural Transmission*, 108, 593–611.
Mathisen, H. C. (1943), "A Method of Testing the Hypothesis That Two Samples Are From the Same Population," *The Annals of Mathematical Statistics*, 14, 188–194.
McLeod, R. S., Taylor, D. W., Cohen, Z., and Cullen, J. B. (1986), "Single-Patient Randomized Clinical Trial: Use in Determining Optimum Treatment for Patient With Inflammation of Kock Continent Ileostomy Reservoir," *Lancet*, 1, 726–728.
Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," *Roczniki Nauk Roiniczych*, tom X, 1–51 [in Polish]; reprinted in English with discussion in *Statistical Science*, 5, 463–480.
Nichols, T. E., and Holmes, A. P. (2001), "Nonparametric Permutation Tests for Functional Neuroimaging: A Primer With Examples," *Human Brain Mapping*, 15, 1–25.
Olson, I. R., Gatenby, J. C., and Grove, J. C. (2002), "A Comparison of Bound and Unbound Audio-Visual Information Processing in the Human Cerebral Cortex," *Cognitive Brain Research*, 14, 129–138.
Randles, R. H., and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: Wiley.
Raz, J. (1990), "Testing for no Effect When Estimating a Smooth Function by Nonparametric Regression: A Randomization Approach," *Journal of the American Statistical Association*, 85, 132–138.
Rosenbaum, P. R. (2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies With a Discrete Pivot," *Biometrika*, 88, 219–231.
——— (2002a), *Observational Studies*, New York: Springer-Verlag.
——— (2002b), "Covariance Adjustment in Randomized Experiments and Observational Studies" (with discussion), *Statistical Science*, 17, 286–327.
Rosenthal, R., and Jacobson, L. (1968), *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development*, New York: Holt, Rinehart and Winston.
Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.
——— (1986), "Which Ifs Have Causal Answers?" *Journal of the American Statistical Association*, 81, 961–962.
——— (1990), "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies," *Statistical Science*, 5, 472–480.
Weiss, B., Williams, J. H., Abrams, B., Caan, B., Citron, L. J., Cox, C., McKibben, J., Ogar, D., and Schultz, S. (1980), "Behavioral Responses to Artificial Food Colors," *Science*, 207, 1487–1489.
Weiss, L. (1955), "A Note on Confidence Sets for Random Variables," *The Annals of Mathematical Statistics*, 26, 142–144.
Welch, B. L. (1937), "On the z-Test in Randomized Blocks and Latin Squares," *Biometrika*, 29, 21–52.
Wilk, M. B. (1955), "The Randomization Analysis of a Generalized Randomized Block Design," *Biometrika*, 42, 70–79.
Wilks, S. S. (1962), *Mathematical Statistics*, New York: Wiley.
Wolfe, D. A., and Hogg, R. V. (1971), "On Constructing Statistics and Reporting Data," *The American Statistician*, 25, 27–30.