Running head: RANDOM ASSIGNMENT OF AVAILABLE CASES

Random Assignment of Available Cases:
Let the Inferences Fit the Design

Clifford E. Lunneborg

University of Washington

Abstract

For historic reasons, reviewed here, psychologists commonly use techniques of statistical inference, valid only when cases are randomly sampled from large populations, in the analysis of data obtained from the randomization of available, non-sampled cases. Modern computing power now permits more appropriate statistical inference for such studies, in the form of rerandomization analysis. The foundations of this approach are outlined and rerandomization illustrated for a range of randomized available case designs.

Author Footnotes

The author is Emeritus Professor of Psychology and Statistics at the University of Washington, Seattle. He may be reached by e-mail at cliff@stat.washington.edu or by post at the Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322. This paper grew out of a series of seminars presented to psychologists and statisticians in the course of a yearlong visit to several Australian universities. The author is indebted to his hosts for providing the opportunity to explore and refine the ideas presented here.

In a recent issue of *The American Statistician* two medical researchers (Ludbrook & Dudley, 1998) summarize and comment upon a survey they carried out of "252 prospective, comparative studies reported in five, frequently cited biomedical journals." Ludbrook and Dudley report that in these studies the "experimental groups were constructed by randomization in 96% of the cases and by random sampling in only 4%" and go on to express concern that the results of 84% of the randomization-design studies were analyzed by *t* or *F* tests. They explain these results as follows: "Statisticians appear to believe that biomedical researchers do most experiments by taking random samples, and therefore recommend statistical procedures that are valid under the population model of inference. In fact, randomization of a nonrandom sample, not random sampling, is more usual. Given this, it is our thesis that the randomization rather than population model applies, and that the statistical procedures best adapted to this model are those based on permutation." (p. 127). They call upon statisticians to provide more appropriate training and consulting.

A survey of recently published psychological research would almost certainly find the same, apparent mismatch between statistical analysis and study design. Psychologists rarely draw random samples of laboratory animals, clinical clients, or student volunteers from large or global populations; we generally employ local stocks of cases, those available to us. Like our medical counterparts we recognize the importance to scientific inference of randomizing these available cases among the treatments to be compared. And, with considerable regularity, we choose to ground our treatment comparisons statistically in an analysis of variance or its more focussed form, a *t*-test.

Is the mismatch between design and analysis real? If it is, how did it arise? What problems has it given rise to? How does the model for statistical inference that Ludbrook and Dudley refer to as randomization or permutation differ from a population inference model? Why is the randomization model more appropriate to

randomized available (i.e., nonrandom) case studies? How can we design studies that will support randomization inference?

These are the questions I'll answer. First, there is a mismatch. Briefly, we commonly over interpret the analysis of variance when it is applied to randomized available case studies. To appreciate the problem and its origin, it will help to review some statistical work of the 1930s.

The Permutation Test Defined and Approximated

In his *Design of Experiments* R. A. Fisher (1935) described, at least by example, a nonparametric alternative to the *t*-test for a paired sample of observations. The data analyzed by Fisher were collected, some sixty years earlier, by Charles Darwin to compare the growth of self-fertilized and cross-fertilized plant material. To appreciate the logic of Fisher's nonparametric test it is worth revisiting the example. I will take some liberties with the description of the study (as did Fisher), but only to emphasize the elements that Fisher felt were critical.

In my scenario Darwin visited his garden and at each of 15 different sites dug up some of the soil, enough when thoroughly mixed to fill two clay pots. He labeled one of each pair of pots *L* and the other *R* and set them side by side, left and right, in the garden. Darwin then returned to the prepared pots with two bags of plant material, one labeled *S* and one *C*, a coin, and an experimental protocol, written on a small slip of paper. The protocol read "At each pair of pots flip the coin. If it lands Heads, plant a randomly chosen specimen from bag *S* in pot *L* and a randomly chosen specimen from bag *C* in pot *R*. If the coin lands Tails, plant a randomly chosen specimen from bag *C* in pot *L* and a randomly chosen specimen from bag *S* in pot *R*. " Following this protocol, Darwin planted the 15 pairs of pots.

Insert Table 1 about here

After a suitable period of time Darwin returned once more to the garden and measured the heights of each of the 30 plants. These heights, in inches, are given in Table 1. The plant material in bag *S* was self-fertilized, that in bag *C* cross-fertilized, and the purpose of the study was to determine if, as Darwin believed, the cross-

fertilized material would show more vigorous growth. The final column of Table 1 gives the difference in heights of the two plants grown in each of the 15 matched pairs of pots, the height of the self-fertilized specimen subtracted from the height of the cross-fertilized one. For the most part, these differences are positive, consistent with Darwin's scientific hypothesis.

By making the height comparison at the paired pot level, Darwin was controlling for any differences in the favorability of growing conditions from one pair to another owing either to differences in the soil or in the subsequent locations of the pot pairs in the garden. As described in my scenario, Fisher would have had Darwin also randomize the plantings in each pair. In this way any remaining difference in growth favorability between the pots in a pair would not bias the outcome of the study. There is no evidence that Darwin did randomize the planting at each location. And the absence of the randomization does not affect the statistical analysis Fisher developed.

What Fisher notes as critical to his analysis is that the 15 self-fertilized specimens and the 15 cross-fertilized specimens were *random samples* from two large populations of those plant materials. Fisher proposed to test the null hypothesis of identical distributions of height for the two populations against the alternative of a greater mean height in the cross-fertilized population. It was to be a nonparametric test, an alternative to the paired observations *t*-test, as no assumption was to be made about the parametric form of the population distributions. Complicating the null and alternative hypotheses about the population distributions of height is the implicit assumption that they could be realized by fixing the growing environment at any one of the 15 employed by Darwin.

The average of the differences in height of the cross-fertilized and self-fertilized plants, over the 15 growing environments, has the height of the cross-fertilized specimen exceeding that of the self-fertilized one by 2.606667 inches. Does a difference that large provide evidence of the superior growth potential of the cross-fertilized material or should we expect differences of this magnitude, apparently

favoring cross-fertilized material, to occur fairly often when randomly sampling from identical growth potential populations?

Fisher found an answer by this logic. In the first pair of pots, the first growing environment, the randomly chosen cross-fertilized specimen achieved a height of 23.5 inches and the randomly chosen self-fertilized specimen a height of 17.4 inches. This gave rise to a cross-fertilized minus self-fertilized difference in heights of 6.1 inches. If the two population distributions of height had been identical, then it is equally likely that a randomly chosen cross-fertilized specimen would have achieved a height of 17.4 inches and a randomly chosen self-fertilized specimen would have achieved a height of 23.5 inches, when grown in this same environment. That outcome would have given a cross-fertilized minus self-fertilized difference in heights of of -6.1 inches. That is, under the null hypothesis the sign attaching to the 6.1 inch difference in heights is as likely to be negative as positive.

This null hypothesis result holds true, of course, for each of the 15 growing environments. And because of the random sampling, the random determination of sign takes place independently from environment to environment. Thus, Fisher was able to argue that under the null hypothesis there were 2^{15}or 32,768 equally likely patterns of signs for the 15 differences in heights.

Darwin observed one of these 2^{15} patterns of pluses and minus, the pattern given in the final column of Table 1. This particular pattern yielded an average difference in heights of $+2.606667$ inches. The probability under the null hypothesis of an average difference in heights of $+2.606667$ inches or larger is the proportion of the 32,768 sign patterns that produce average differences in height of $+2.606667$ inches or larger. How many are there? Fisher counted them and, as testimony to the difficulty of the task, had to correct his initial result (Edgington, 1995). There are 866 sign patterns yielding an average difference in heights as large or larger than the observed 2.606667 inches. Thus we have a *P*-value for the hypothesis test of $866/32,768 = 0.02643$.

We refer to this nonparametric test of Fisher's as a *permutation* test because the null hypothesis allows us to permute our randomly sampled observations among the population distributions from which they were sampled. In the Darwin example the distributions of the heights of a population of cross-fertilized plants and of a population of self-fertilized plants, when grown in the environment provided by the first pair of pots, were hypothesized to be identical. This allowed Fisher to permute the pair of randomly sampled heights, 23.5 inches and 17.4 inches, between the two population distributions.

In 1935 it was totally impractical to determine the *P*-value of the nonparametric permutation test by actually counting the number of outcomes under the null hypothesis that would be at least as favorable to the alternative as that observed. Fisher, though, provided a way to approximate this *P*-value by an easy computation. Continuing with the Darwin example, he computed the *t*-statistic

$$t = \frac{2.606667 - 0}{\sqrt{22.21067/15}} = 2.1422$$

where 22.21067 is the normal random variable variance estimate computed from the 15 height differences. Referring this *t*-statistic to the distribution of Student's *t*-random variable with 14 d.f. yields an upper tail *P*-value, $P(t_{14} \geq 2.1422)$ of 0.02512. For this example, the parametric *t*-test *P*-value provides a close approximation to that of the nonparametric permutation test. Fisher argued that the approximation would be a good one for other sets of data as well.

The Randomization Test Introduced

In a series of articles published in 1937-38, all bearing the general title "Significance tests which may be applied to samples from any populations," Pitman (1937a, 1937b, 1938) extended Fisher's permutation test to other experimental designs. The first of these (Pitman, 1937a) addressed what is probably the most common comparative treatment design, that involving two independent treatment groups.

Pitman assumes the design to be based on observations for *random samples* from two case populations. The two samples might be the result of independently sampling two distinct or *natural* case populations, e.g., male and female lecturers in a state university system, and then observing an attribute of interest on the sampled cases, e.g., years of teaching experience.

Alternatively, and this design was of greater interest to Pitman, a single random sample might be obtained from one natural case population, e.g., freshmen at one university, and then *randomly divided* to form two treatment groups. After exposing cases in the two groups to different treatments, e.g., an orientation program with no emphasis on alcohol moderation or one featuring alcohol moderation, a response to treatment is observed, e.g., end-of-freshman-year self-reports of binge drinking. As a result of the random division of the single random sample into two treatment groups, the resulting treatment responses actually are those of two random samples drawn from two different populations, one random sample is drawn from the population of freshmen all of whom would have received the no mention of moderation orientation and a second random sample is drawn from the population of freshmen all of whom would have received the alcohol moderation orientation. Populations of this kind I describe as *prospective* populations, they are the case populations that would result if all cases in some natural population were to be exposed to a particular treatment (Lunneborg, 1999).

The permutation test for two independent, randomly sampled treatment groups has come to be known as the Pitman test and is rather more easily described than Fisher's paired-observations test. The Pitman test takes as its null hypothesis that the two populations have identical distributions of response observations. The alternative is that the two population distribution means differ and often specifies the distribution with the higher mean.

An artificially small example will be used to demonstrate Pitman's test. Let's assume that our researcher's random sample of university freshmen was made up of just six students and that these students were then randomly divided into two

treatment groups of three students each. Table 2 reports the results of this study, the reports by the six students of the number of weekends in the freshman year during which they consumed more than eight alcoholic drinks, our researcher's definition of an alcoholic binge.

Insert Table 2 about here

The difference in the the mean number of binges for the two orientation groups is

$$\overline{Y}_{NM} - \overline{Y}_{M} = \left(\tfrac{13+9+11}{3}\right) - \left(\tfrac{8+12+7}{3}\right) = 11 - 9 = 2,$$

suggestive of less frequent bingeing in the alcohol moderation orientation population. Is this a large enough difference to convince us that the population mean would be smaller if the orientation featured alcohol moderation? Or would sample mean differences of this magnitude or greater be relatively common when drawing random samples of this size from population distributions that are identical?

Pitman's permutation test procedure follows from this line of reasoning. Randomly sampling the *No Moderation* population distribution produced three scores $(9, 11, 13)$ and randomly sampling the *Moderation* population distribution produced three scores $(7, 8, 12)$. Sampling three times from each of the two distributions yielded six scores $(7, 8, 9, 11, 12, 13)$.

Now let's assume that the *No Moderation* and *Moderation* population distributions were identical and that randomly sampling from each three times produced these same six scores, $(7, 8, 9, 11, 12, 13)$. Because the population distributions were identical, then any three of these scores could belong to the sample from the *No Moderation* population distribution and the remaining three to that from the *Moderation* population distribution. The obtained result, $(9, 11, 13)$ from the *No Moderation* distribution and $(7, 8, 12)$ from the *Moderation* distribution, was but one of the possible *permutations* of those six observations into two samples of three each. In how many different ways can the six observations be permuted, forming two subsets, three from *No Moderation* and three from *Moderation*? The answer is given by application of a familiar combinatorial formula:

$$\frac{(n_{NM}+n_M)!}{n_{NM}!\times n_M!}=\frac{6\times5\times4\times3\times2\times1}{(3\times2\times1)(3\times2\times1)}=20\,.$$

Under the null hypothesis, sampling from identical distributions, these 20 permutations of the observed set of six observation are *equally likely* outcomes.

How likely is that $\left(\overline{Y}_{NM}-\overline{Y}_M\right)$ would take a value of 2 or larger if the two population distributions were identical? The permutation test answer is the proportion of the 20 equally likely permutations that yield a value of $\left(\overline{Y}_{NM}-\overline{Y}_M\right)$ of 2 or greater. For this artificially small example it is convenient to display all 20 of the permutations and the differences in means associated with them. These results make up Table 3.

Insert Table 3 about here

It is easily established from Table 3 that for 4 of the 20 permutations $\left(\overline{Y}_{NM}-\overline{Y}_M\right)$ is 2 or greater. This gives a *P*-value of $(4/20)=0.20$, the probability under the null hypothesis of a result at least as favorable to the alternative hypothesis as the one obtained.

For realistic-sized samples, of course, the computation of the nonparametric, permutation test *P*-value was infeasible in 1937. Pitman showed how to approximate that *P*-value with the one provided by the corresponding parametric test, the independent groups *t*-test.

Pitman developed his two sample test having in mind that the natural population or populations from which the two random samples of cases were drawn would be large, perhaps infinite in size. In describing the test, however, he noted that it also was applicable to experiments in which the two samples *exhaust* the population. When do an experimenter's two random samples exhaust a population? The experiment referred to by Pitman most commonly takes the form of what I call the two-group randomized available case design. A *local population*, i.e., a set of available cases, is randomly divided into two treatment groups. Because of the randomization, the resulting two groups not only exhaust the local population but are random samples of that population. The permutation test inferences in this design are

inferences about the local population, i.e., they are restricted to the set of available cases.

The distinction in range of inference is illustrated by revisiting the example of the impact of a *Moderation* orientation on binge drinking. If the students randomized by our researcher to the two orientation treatments were a random sample from a much larger or *global population*, e.g., the 3,000 freshmen at Central State University, then the *permutation test* hypotheses would be hypotheses about binge drinking in that global population, as impacted by any *Moderation* orientation. Would the mean number of reported binges *in the freshman population* be smaller if all freshmen were to receive the *Moderation* orientation compared with what it would be if all received the *No Moderation* one?

However, if the students randomized between the two treatments were a set of available cases, e.g., six volunteers from one Introduction to Psychology quiz section, then the *randomization test* hypotheses would be restricted to that set of cases. Would the mean number of reported binges *among these six volunteers* be smaller if all were to receive the *Moderation* treatment, than it would be if all were to receive the *No Moderation* one?

The distinction in range of inference is reinforced by a change in the name of the hypothesis test. I follow here the practice of Edgington (1995) and refer to the permutation test when applied to a randomly divided available set of cases as a randomization test. Thus, Pitman's first article introduced the randomization test and, through its parentage in the permutation test, provided support for the use of the parametric population *t*-test as a way of approximating the null hypothesis test *P*-values of the randomization test. He also speculated in that article that the randomization test would be of greater importance than the large population version of the permutation test. In that randomized available case studies are much more common than those in which cases are randomly sampled from large populations, Pitman certainly was correct.

Local Causal Inference

The third of Pitman's articles on permutation tests (Pitman, 1938) appeared in the same issue of the journal *Biometrika* as one by Welch (1938). Both were concerned with the practicality of extending the nonparametric permutation test to studies in which cases, randomly sampled from one or more natural populations, are randomized among more than two levels of treatment. How well could *k*-group permutation (or, randomization) test *P*-values be approximated by those provided by a parametric analysis of variance?

For a certain class of designs, Welch compared two null sampling distributions for the *F*-ratio. One null distribution was based on permuting observations among *k* identical populations and the other was the parametric, *F*-random variable one, based on randomly sampling *k* identical normal population distributions. He determined that the two null distributions while quite similar would differ and that, translating into modern terminology, the *P*-values obtained from a normal-theory *F*-distribution appeared to provide better approximations to permutation test *P*-values than to randomization test *P*-values.

Noteworthy in Welch's article is the attention he gave to the ranges of *statistical inference* associated with the permutation and randomization tests and to the distinction between statistical inference and *scientific inference*. He recognized the limited range of statistical inference for the randomized available case design, i.e., to those available cases. But he stressed the importance to scientific inference of the case randomization that is characteristic of that design. Randomization of cases to treatments provides the basis for the belief that an observed difference in response to treatment is the result of differences in treatments, where that response difference is large enough not be a chance result of the randomization. That is, we may draw *causal inferences* as a result of the randomization of cases among treatments.

Welch made it clear that although these causal inferences could be only local from a statistical perspective, their scientific importance may have greater sway. The researcher may know enough about the science driving the research to argue successfully that the response difference caused by the differential treatment of these

particular cases, e.g., water-deprived-mice in a San Francisco laboratory, hypertensive patients from a Cleveland clinic, English-speaking sophomores enrolled in a first course in psychology at a Minnesota college, could be reproduced among similar cases, in other places at other times, given the same treatments. That is, *scientific inference* about the generalizability of the results may be feasible where statistical generalizability is not.

Local causal inference, perhaps buttressed by scientific generalizability, is the intent of randomized available case studies. Where the scientific argument is strong enough the randomized case study may be sufficient to establish a generalizable result. In other instances, establishing a causal effect in a local context may provide the evidence needed to justify the expense of mounting a study employing random samples of cases from carefully defined populations, to determine the generalizability of the causal effect. Whether confirmatory of a scientific conjecture or a pilot study, the randomized available case study and its associated potential for determining treatment-response causal relations, at least for a local population, has become the workhorse of empirical science.

Over interpretation of the Normal-Theory Approximation

By the late 1930s, then, a randomization model for the statistical analysis of randomized available case studies had been established. Computational difficulties precluded the direct tabulation of null hypothesis reference distributions for treatment comparison statistics but those distributions could be approximated, researchers were assured, by the probability distributions of the appropriate *t* and *F* random variables. It is important to remember that the randomization test reference distribution eschews sampling from any global population, not just those with normal response distributions.

As a result of the useful *P*-value approximations provided, the design and analysis of randomized available case studies have been incorporated into the analysis of variance. Randomization of cases is generally regarded as justifying the analysis of variance.

Some experimental design texts make explicit the approximation basis for the linkage. For example, Box, Hunter and Hunter (1978) wrote with respect to the two-treatment comparison, "… the randomization reference distribution is usually approximated reasonably well by the appropriately scaled *t* distribution. Hence, provided that we randomize, we can employ *t*-tests as approximations to exact randomization tests, and we will be free of the random sampling assumption as well as the assumption of exact normality." (p. 96) and they repeated this message with respect to multiple-treatment comparisons, "Thus, as before, the normal theory test can be regarded as an approximation for the randomization test, which is its ultimate justification." (p. 188).

More recently Maxwell and Delaney (1990), in a chapter titled *Introduction to the Fisher Tradition*, used data from a hypothetical randomized blocks study with twins to demonstrate a near-agreement in randomization test and *t*-test *P*-values and went on to conclude, more generally, that "… the close correspondence between the results of randomization and normal theory-based tests provides a justification for using the normal theory methods … regardless of whether subjects are in fact randomly *sampled* from a population." (p. 50).

Other experimental design texts, e.g., Cobb (1997), offer randomization as a justification for normal-theory analysis, without any reference to approximating a randomization test. Still others, e.g., Winer, Brown, and Michels (1991), treat randomization solely from the point of view of increasing experimental control, ignore it as the basis for statistical inference in available case studies, and justify the analysis of variance on the assumption that a group of available cases "is considered by the experimenter to be the equivalent of a random sample from the population of interest." (p. 74).

However randomization is regarded, experimental design texts encourage the over interpretation of the results of the application of normal-theory methods to the analysis of randomized available case studies. This is inadvertent, the result of offering a uniform approach to the analysis of two kinds of studies, those in which our

random samples of cases exhaust a local population, i.e., randomized available case studies, and those in which we randomly sample relatively small numbers of cases from very large, global populations. These over interpretations are implicit in the worked examples and in the problems set for readers and take three forms.

Global Population Hypotheses Tested. First, the distinction between local and global populations is nearly always ignored when statistical hypotheses are stated or tested. Analysis of variance hypotheses are hypotheses about the means of treatment response scores for very large, essentially infinite case populations. And the *P*-value for a treatment comparison is interpreted in terms of a *null sampling distribution* for the treatment comparison statistic, the distribution of values of that statistic over all possible samples from these very large populations when the distribution of response scores for those populations are restricted by a null treatment hypothesis. These samples would necessarily include a large number of cases who did not participate in our study and who would contribute treatment responses different from those actually observed.

The randomization test treatment comparison hypotheses that are most like those in the analysis of variance are hypotheses about the mean response to treatment for a relatively small number of cases, those who received one of the treatments being compared. And the *P*-value for a randomization test treatment comparison in the randomized available case design is based on a *null reference distribution* for the treatment comparison statistic, the distribution of values of that statistic if those responses to treatment actually observed were permuted among treatments, in all the ways possible under a null treatment hypothesis.

While the *P*-value for a randomization test may be approximated by that for a *t*-test or *F*-ratio it has different interpretations in the two traditions. The two interpretations are confused when the global population hypotheses of the analysis of variance are inappropriately applied to the randomized available case study.

Global Population Parameters Estimated. The typical analysis of variance yields more than *P*-values. Students, by their texts, and researchers, by their editorial

reviewers, are encouraged to report not only estimates of contrasts among means but estimates, as well, of the standard errors (*S.E.*s) for those estimates and of confidence intervals (*C.I.*s) for the contrasts. The conventional estimates of *S.E.*s and *C.I.*s, those taught by our experimental design texts and incorporated into popular statistical computing packages, assume random samples from very large populations with response distributions that are homoscedastic and normal. The robustness of the *P*-value, however, does not carry over to these accuracy assessments. The large population estimates of *S.E.*s and *C.I.*s are not appropriate to the randomized available case design.

Extension to Nonrandom Factors. The scope of analysis of variance hypothesis testing has been extended usefully, at least in *P*-value estimation, to include randomized treatment factors. Frequently, researchers attempt to extend the envelope of applicability even further, to the statistical assessment of factors which in the design have neither a random sampling nor a randomization basis. As an example, it is common to encounter tests of a sex-effect or of a sex-by-treatment interaction in research where the cases have not been randomly sampled from separate male and female populations nor, clearly, have they been randomized to the two sex levels. There is no probabilistic basis for such tests. Even where sex is used as a blocking factor, male and female cases having been randomized separately and independently among levels of a treatment factor, no sex-difference test is warranted. Not every factor in a randomized available case design supports statistical inference, only *treatment factors*, those with cases randomly distributed over the factor levels.

In importing wholesale those analysis of variance computations appropriate to large population samples researchers misinterpret the outcomes and implications of the randomized available case study. This is the misuse of the population model of inference addressed by Ludbrook and Dudley (1998).

Impact of ANOVA Distributional Assumptions. Before an ANOVA is carried out it is customary to verify that the observed data are consistent with the distributional assumptions of the ANOVA model. For between groups designs

apparent violations can lead to the transformation of the response variable. Subsequent hypothesis testing, of course, concerns the population distributions of the transformed variable. Such hypotheses may be a poor fit to those substantive hypotheses that motivated the study. The ANOVA for within case or repeated measures designs has its own distributional requirements and psychologists have been active participants in the quest for better ways of adapting the within-case ANOVA to apparent violations of those assumptions.

These distributional assumptions are grounded in randomly sampling global, i.e., normal or multivariate normal, population distributions. Their apparent violation is dealt with no differently in our texts and research literature when they are found in local population, available case studies. Otherwise the appropriateness of the ANOVA *P*-value as an approximation to the randomization *P*-value presumably would be compromised.

Adoption of the ANOVA model as a means of approximating a randomization test *P*-value leads to testing distributional assumptions that are irrelevant to local causal inference and may result in transforming response measures away from their natural, interpretable metrics.

Computational Feasibility of Randomization Inference

The computational difficulties that precluded the application of randomization inference to available case studies in the 1930s have been overcome. All researchers now have on their desktops computing power sufficient to create the null reference distributions needed for randomization tests. And the statistical software is available to make the task a relatively easy one. We no longer need to approximate *P*-values from normal-theory statistics and, in so doing, risk confusing global or population inference with local inference.

Exact Randomization Tests. It is customary to distinguish two approaches to obtaining a null reference distribution for a treatment comparison statistic. If we visit each randomization (or, rerandomization) of the observed responses that is possible under the null treatment hypothesis, recomputing the treatment comparison statistic

for each randomization, the result is an *exact* reference distribution. That is, when we compute a *P*-value by referring the observed value of the treatment comparison statistic to that distribution we obtain an exact, not an approximate, *P*-value for the hypothesis test.

As an example, we randomize ten available cases five apiece to two treatment groups, expose the members of the two groups to two different treatments (*A* and *B*), and measure their responses to treatment. The randomization test null treatment hypothesis is that the distribution of measured responses to treatment for our ten cases would be the same if all had received treatment *A* as it would be if all had received treatment *B*. Under that hypothesis all possible divisions of the ten observed response scores into five from distribution *A* and five from distribution *B* are equally likely. There are $M = 10!/(5! \times 5!) = 252$ such possibilities and it would require very little computing time to form all of them systematically and to compute the treatment comparison statistic in each. These would make up the exact null reference distribution against which we would evaluate our obtained treatment comparison statistic.

Doubling the size of the two treatment groups would increase substantially the number of possible rerandomizations, $M = 20!/(10! \times 10!) = 184,756$. Rather than inventory all possible rearrangements of the data, clever timesaving algorithms have been devised for some treatment comparison statistics that identify just those rerandomizations for which the treatment comparison statistic would take at least as extreme a value, relative to the null hypothesis, as that actually observed, e.g., Mehta and Patel (1995). These algorithms, in effect, produce exact *P*-values from just the tails of a null reference distribution.

Exact to Monte Carlo Accuracy Randomization Tests. If we double the treatment group sizes a second time the number of rerandomizations of the observed responses possible under a null treatment effect hypothesis is staggeringly large, $M = 40!/(20! \times 20!) = 137,846,528,820$. For moderate to large studies the computing time required to visit all rerandomizations or, our statistic permitting, even

to identify and inventory the tails of the reference distribution is likely to be longer than is acceptable. Where this is true we can approximate the null reference distribution by computing the treatment comparison statistic for a series of randomly chosen rerandomizations of the observed responses.

The *P*-value computed from this Monte Carlo approximated reference distribution can be made arbitrarily close to the exact *P*-value, that based on the exact reference distribution, by including enough randomly chosen rerandomizations in the sequence. How close to approximate the exact *P*-value will depend both on the magnitude of the *P*-value and the use to which it is to be put. One strategy (Good, 1999) is to obtain a preliminary estimate based on a random sequence of 500 to 1,000 rerandomizations and then to refine this estimate by increasing the sequence to 5,000 or 10,000 rerandomizations where greater accuracy is wanted. Edgington (1995) and Manly (1997) give tables relating the size of the Monte Carlo reference distribution to the accuracy of the resulting *P*-values. Roughly, if the exact *P*-value for a test is 0.05, then using a reference set based on 5,000 rerandomizations will yield a *P*-value that falls between 0.042 and 0.058, 99% of the time. And, if the exact *P*-value is 0.01, then a 10,000 element reference distribution will give a value between 0.007 and 0.013, again 99% of the time.

The creation of exact reference distributions requires specialized computational routines to carry out the systematic identification of all possible rerandomizations. Specialized routines are needed as well to implement tail-filling algorithms and these algorithms are limited to certain treatment comparison statistics. By contrast, Monte Carlo approximations to null reference distributions can be formed using only the standard tools available in most statistical computing packages. They can be formed whatever the researcher chooses as a treatment comparison statistic. For these reasons Monte Carlo approximated reference distributions, rather than exact ones, are of central importance to statistical inference for the randomized available case study.

Examples of Randomized Available Case Study Designs

In the balance of this paper I illustrate the application of rerandomization in drawing local causal inferences from a range of randomized available case designs. An emphasis in the illustrations will be the linkage between the randomization scheme used in the design and the subsequent testing of treatment effect hypotheses.

Completely Randomized Two Treatment Groups Designs

The completely randomized, two-groups design is the most straightforward randomized available case design, and perhaps the most common. Complete randomization (CR) is a strategy whereby the full or complete set of available cases is randomized among two or more treatment groups. The *CR* design contrasts, for example, with the randomized blocks design in which the available cases are first formed into two or more blocks, homogenous with respect to some characteristic, and each block of cases then independently randomized among treatments.

In a study reported by Regan, Williams and Sparling (1977) shoppers in a mall were recruited into either a *Guilt* or *Control* treatment group. In both instances the shoppers were approached and asked to use the experimenter's camera to photograph the experimenter. In the *Control* treatment the shopper was thanked but in the *Guilt* treatment the shopper was led to believe (only for the duration of the shopper's participation in the experiment) that the shopper damaged the camera. Shopper-subjects then encountered the experimenter's accomplice who was clearly having trouble with a disintegrating shopping bag. The shoppers either *Helped* or *Did not help* the accomplice.

The 40 shoppers were not a random sample from some larger population; they were the first 40 who, when approached, agreed to photograph the experimenter. They were, however, completely randomized into two groups of 20 by a strategy of this kind: In advance of approaching shoppers, the experimenter prepared a 40-element list consisting of a well-shuffled 20 *Guilts* and 20 *Controls*. As the *k*th shopper agreed to participate he or she was assigned to the treatment level identified by the *k*th element on the list.

The relevant data for this study are these. Among the 20 shoppers randomized to the *Control* treatment there were 3 helpers and 17 non-helpers while among those randomized to the *Guilt* treatment there were 11 helpers and 9 non-helpers. If all 40 shoppers (the local population) had been exposed to the *Control* treatment they would define one (local) population distribution of helpers and non-helpers and if all 40 had been exposed to the *Guilt* treatment they would define a second population distribution of helpers and non-helpers.

The randomization null treatment effect hypothesis is that the two 40-element population distributions, each a mixture of helpers and non-helpers, would be identical. The alternative, substantive hypothesis is that the *Guilt* population distribution would include more helpers than the *Control* distribution.

Under the null treatment hypothesis of identical population distributions the randomization of cases could result in any 20 of the observed 14 helpers and 26 non-helpers coming from the *Control* distribution and the remaining 20 from the *Guilt* distribution. Each of these outcomes would have the same chance of occurring. For example, to focus on a pair of possibilities, it is just as likely under the null hypothesis that the randomization would yield 4 helpers from the *Control* distribution and 10 from the *Guilt* distribution as that it would yield 10 helpers from the *Control* distribution and 4 from the *Guilt*. By contrast, under the alternative hypothesis the researcher would expect to see more helpers from the *Guilt* distribution than from the *Control* one.

The researcher's randomization of the shoppers, we noted earlier, yielded 3 helpers from the *Control* and 11 helpers from the *Guilt* population distributions. The result is in the direction of the alternative. Does the result provide strong evidence in support of the alternative or is there a good chance that evidence of the same or greater strength could emerge under the null hypothesis? To answer the question we need to examine the null reference distribution for an appropriate treatment comparison statistic. We could take as our statistic the difference in the number of helpers in the two treatment groups, $s = n(\textit{Helpers}|\textit{Guilt}) - n(\textit{Helpers}|\textit{Control})$, and

ask whether the observed difference, $s = 8$, is large enough to provide statistical support for the substantive hypothesis that the *Guilt* manipulation leads to more helpers among these shoppers. Our statistic might be more easily interpreted if we were to divide it by 20, converting it into a difference in the proportions of helpers in the two treatment groups, $s = p(\textit{Helpers}|\textit{Guilt}) - p(\textit{Helpers}|\textit{Control})$. The statistical consequences, it should be noted, of the two statistics would be identical. They are *reference distribution equivalent statistics*. That is, they would order any set of rerandomizations of the observed data in exactly the same way.

As noted above, the number of equally-likely-under-the-null rerandomizations of the observed responses to treatment, $40!/(20! \times 20!)$, is quite large and I will not attempt to form the exact null reference distribution. Rather, I'll use the ***S-Plus*** `permutationTest` function (MathSoft, 1998; Hesterberg, 1999) to generate a Monte Carlo approximation. Figure 1 is the log of the interactive computing.

Insert Figure 1 about here

The response vector `helpv` has as its sequential elements 3 '1's, 17 '0's, 11 '1's, and 9 '0's. The first 20 elements code the helpers and non-helpers in the *Control* treatment group and the final 20 elements do the same for the *Guilt* cases. The treatment vector, `trtv`, is aligned with the response vector, 20 'C's followed by 20 'G's.

The `permutationTest` function has two mandatory and several optional arguments. The mandatory arguments are a vector whose elements are to be randomly shuffled (or an array whose rows are to be randomly reordered) and a statistic to be computed before the vector is shuffled and again following each shuffling of the vector. Together they will comprise the Monte Carlo null reference distribution for the statistic. The shuffled vector, in my use of the function, codes the treatment group to which a case is randomized or, equivalently, the population distribution to which a response is to be attributed. In this first application the vector to be shuffled is `trtv` and the statistic to be computed is the difference between the proportion of helpers in the *Guilt* and *Control* samples: `mean(helv[trtv=="G"])-`

`mean(helv[trtv=="C"])`. Optional arguments are supplied here to specify that the test is not two-sided and that the number of rerandomizations of treatment assignments to be included in the Monte Carlo reference distribution is other than 1,000. My alternative hypothesis is that the proportion of helpers would be greater in the *Guilt* population distribution than in the *Control* one so I want a one-sided test. For the *P*-value to correspond to the proportion of the reference distribution that is at least as large as the observed value of the statistic I specify `alternative="greater"`. I want the reference distribution to consist of the observed value of the treatment comparison statistic plus the values of that statistic computed from `B=999` rerandomizations of the treatment assignments giving a total of 1,000 null hypothesis values for the statistic. A final optional argument `trace=F` asks that ***S-Plus*** not fill my computer screen with updated reports on just which of the rerandomizations it is currently processing!

The `permutationTest` returns its results as an object, in this example one named `hnh`. A summary of that object gives the `Observed` value of the treatment comparison statistic as well as some descriptive statistics for the reference distribution, including the *P*-value for the permutation test. The value of our treatment comparison statistic is 0.40, the difference between the proportion of helpers in the *Guilt* treatment group ($11/20 = 0.55$) and that in the *Control* group ($3/20 = 0.15$). Exactly 14 of the 999 rerandomizations of the treatment assignment vector resulted in values of the statistic of 0.40 or larger and, hence, to the reported *P*-value of 0.015.

My Monte Carlo approximated *P*-value was close to 0.01. Were the formal rejection of the null hypothesis to depend on the *P*-value for the hypothesis test falling below 0.01, a more accurate approximation might be desirable. I repeated the `permutationTest` command, increasing the number of rerandomizations to `B=5000` and obtained a *P*-value of $0.008798 = 44/5001$. As the two random sequences of rerandomizations were independent of each other I may as well pool the two results to get an approximation based on 5,999 randomly chosen rerandomizations, $(14 + 43 + 1)/6000 = 0.009667$.

Multiple Treatments and Multiple Comparisons

If cases are randomized to two treatments, there is but one treatment comparison to be made; however, multiple treatments invite multiple comparisons. Consider this example, reproduced as Data Set 50 in Hand, Daly, Lunn, McConway & Ostrowski (1994) and described therein as follows. “A double-blind experiment was carried out to investigate the effect of the stimulant caffeine on performance on a simple physical task. Thirty male college students were trained in finger tapping. They were then divided at random into three groups of 10 and the groups received different doses of caffeine (0, 100, and 200 mg). Two hours after treatment, each man was required to do finger tapping and the number of taps per minute was recorded. Does caffeine affect performance on this task? If it does, can you describe the effect?” (p. 40). The finger tapping rates are given in Table 4.

Insert Table 4 about here.

There is no suggestion that the 30 male college students were a random sample from a well-defined case population. Almost certainly they were available cases, student volunteers. Their randomization among treatment levels provides the mechanism, though, for local causal inference.

First treatment comparison: caffeine Vs no caffeine. I'll try to answer the question posed by making two treatment comparisons. These may not be the comparisons another analyst would choose. The first comparison is directed at providing one answer to the question “Does caffeine cause an increase in tapping speed?” The randomization null treatment hypothesis associated with this substantive hypothesis is that three local population distributions of tapping speeds ($N = 30$, in each) are identical. They are the tapping speed distributions of these 30 students, if all were to have received the 0 mg, the 100 mg, or the 200 mg caffeine treatment. The alternative is that, while the 100 mg and 200 mg distributions may be identical, the 0 mg distribution includes slower tapping speeds.

The test of this first null hypothesis will be based on a comparison of the performances of the 20 students receiving some amount of caffeine with those of the

10 students receiving none. How shall I do that? The natural thing to do, given our training in psychological statistics, is to compute the two means and take the difference between them. The mean of the 0 mg group estimates the mean of the population distribution at 0 mg caffeine treatment. The mean for the 20 students in the 100 mg and 200 mg groups pools observations from the other two population distributions to provide an estimate of their common mean. However, for purposes of illustration, I'll express a concern that, because either of the population response distributions might have at least one long tail, the mean may be an inappropriate assessment of the *typical magnitude* of response. So I choose as my treatment comparison statistic the difference in *20% trimmed means* for the *Caffeine* and *No Caffeine* groups. That is, I'll trim from each group of tapping speeds the smallest and largest 20% before computing the means.

Insert Figure 2 about here

The upper portion of Figure 2 reports the randomization analysis for the *Caffeine/No Caffeine* comparison. The treatment responses, tapping speeds, are arranged in the `tapspeed` vector to coincide with the treatment assignments coded in a second vector, `caff`. The 20% trimmed mean tapping speed for the *Caffeine* students (`caff>0`) is 2.667 taps/min faster than that for the *No Caffeine* students (`caff==0`). Is that a big difference? Under the null treatment hypothesis all divisions of the 30 observed tapping speeds, 10 from each of the three *Caffeine* population distributions, had the same chance of being observed under the randomization strategy employed in this study. The call to `permutationTest` provides a Monte Carlo approximation to the null reference distribution for my treatment comparison statistic by rerandomizing the treatment assignment vector, `caff`, 999 times and computing the difference in trimmed means for each rerandomization.

The probability of observing a difference in trimmed means favoring the combined *Caffeine* groups by at least 2.667 taps/min–if the local population tapping speed distributions for the three treatments are identical–is of the order of 0.004.

Many of us would conclude, on this evidence, that the administration of caffeine led to increased tapping speeds among these 30 male collegians.

Second treatment comparison: 100 mg Vs 200 mg caffeine: My second treatment comparison for this three-treatment design is intended to answer this question, “Does a larger dose of caffeine lead to faster tapping than a smaller dose?” That is, I'm interested in comparing the tapping rates among these students when they receive either 100 mg or 200 mg caffeine.

The second null treatment effect hypothesis is that the tapping speeds observed for the 100 mg treatment students and those for the 200 mg students are random samples from two identical (local) population distributions. The alternative is that tapping rates in the 200 mg population distribution are higher than those in the 100 mg population.

Again, I'll use a difference in 20% trimmed means as the treatment comparison statistic. The observed difference, as reported in the lower portion of Figure 2, is 1.833 taps/min higher for the 200 mg group, in the predicted direction. Does that difference provide strong evidence in support of the substantive hypothesis or is it a reasonable probability under the null hypothesis?

Under the null treatment hypothesis any 10 of the 20 tapping speeds observed for the 100 mg and 200 mg students could have come from the 100 mg population distribution and the other 10 from the 200 mg population distribution. I develop a (Monte Carlo) null reference distribution for my treatment comparison statistic by asking the `permutationTest` function to find the value of that statistic for 999 randomly chosen rerandomizations of the treatment assignments making up the `caff` vector. Our second null hypothesis, however, restricts these rerandomizations. Observations from the 100 mg and 200 mg population distributions are exchangeable between those two distributions, but those observations are no longer exchangeable with ones from the 0 mg distribution. The optional `group` argument is used by `permutationTest` to control exchangeability. It takes as its value a vector the same length as the treatment assignment vector with elements coding exchangeable

treatment assignments. In this example I use a vector `shufgrp` coded '1' for the 0 mg elements of `caff` and '2' for both 100 mg and 200 mg elements. This insures that treatment assignments can be exchanged, on rerandomization, between 100 mg and 200 mg, but not between either of those and 0 mg.

Based on the resulting reference distribution we see that there is a 5-6% chance that, when sampling from identical 100 mg and 200 mg local population distributions, the 20% trimmed mean for the 200 mg sample will exceed that for the 100 mg sample by at least 1.83 taps/min.

For the second hypothesis test, as for the first, the local population is defined by the 30 student subjects. The original randomization insured that the treatment groups were all random samples from those 30 cases.

My history of the randomization test emphasized its origins as a distribution-free alternative to the *t*- and *F*-tests. The *Guilt* and *Caffeine* examples serve to establish that the randomization test is not restricted to testing hypotheses about the population means of continuously-varying response measures.

Randomized Block Designs

The second *Caffeine* randomization test illustrated that the rerandomizations needed for the reference distribution had to be faithful to the particular null treatment hypothesis as well as to the original randomization strategy.

The *Guilt* and *Caffeine* studies are both examples of an all-cases-at-once or complete randomization strategy. This is not the only randomization strategy employed by available-case researchers. One important alternative is first to form the available cases into two or more blocks, on the basis of some attribute of the cases, and then independently to randomize each block of cases among treatment groups. The blocking attribute is one that is expected to influence the response attribute. Thus, in the *Caffeine* study the 30 student volunteers might have been formed into blocks based on their pretreatment tapping speeds prior to being randomized among

treatment levels. The randomized blocks (*RB*) design can give us increased experimental and statistical control.

At the experimental level, randomizing within blocks can insure that the blocking attribute is not confounded with treatment. Each treatment group can be made up of the same blend of cases with respect to their blocking attribute levels.

At the statistical level, randomizing within blocks can increase our sensitivity to a difference in response to treatments. Where blocking level influences the magnitude of treatment response, as it will if the blocking attribute is well chosen, then the treatment responses will be more homogeneous within a block than across blocks. For the rerandomization analysis of randomized blocks this is important; the null and substantive hypotheses take the blocks of cases as their local populations. For a block of cases, then, there is a population distribution of responses for each treatment level. Against the low variability within these population distributions any differences in, say, location among the distributions are more likely to be identified than they would against the more heterogeneous population distributions of a completely randomized study.

The *RB* design impacts how we carry out the rerandomizations needed to form reference distributions for our treatment comparisons. We must be careful to limit our rerandomizations to those that rerandomize cases within blocks and, hence, maintain the block structure within treatment groups. These restrictions on the rerandomizations reduce the sizes of null reference distributions. The use of 10 blocks of three students, for example, rather than a single block of 30 students would reduce the sizes of the two reference distributions to

$$\left(\frac{3!}{2! \times 1!}\right)^{10} = 59,049 \text{ and } \left(\frac{2!}{1! \times 1!}\right)^{10} = 1,024$$

for the *Caffeine/No Caffeine* and *200 mg/100 mg Caffeine* comparisons. Here the change in reference distribution sizes would not be critical. Each provides scope certainly for testing at the 0.01 level. When planning *RB* studies, however, it is

important to insure that the overall number of cases is large enough to support inference for focussed treatment comparisons which involve only some of the cases. A simulation, similar to a power analysis for random global population samples, can help to determine the number of cases to be recruited into a study. Lunneborg (1999) provides an example.

Insert Table 5 about here

In the *RB* study summarized in Table 5 each block consists of exactly two experimental units. The data are from a plant-growing study reported in Mead, Curnow and Hasted (1993) but, having already visited Darwin's garden, I've invented a psychologically more interesting scenario for the data. The cases are 10 pairs of identical twins, aged 4 years 6 months to 5 years 6 months. The response data are scores on an *Attention* task, higher scores indicating greater attention. Each pair of twins is randomized. One twin completes the task with Mother Absent, the other with Mother Present. Does Mother Present increase attention? Figure 3 is my ***S-Plus*** dialog.

Insert Figure 3 about here

The three 20-element vectors, `atn`, `mothr`, and `pairs` are aligned with one another and contain, respectively, the attention score, the randomized treatment level (*P* for Mother Present, *A* for Mother Absent), and the twin-pair identification (*1* through *10*) for each of the 20 children in the study. My treatment comparison statistic is the mean of the differences in attention scores for the ten pairs of twins, Mother Present minus Mother Absent: `mean(atn[mothr=="P"]-atn[mothr=="A"])`. Its observed value is 1.7.

To help decide whether that is a large enough difference in favor of Mother Present as to be unlikely solely because of the randomization of the twins within each pair, i.e., in the absence of any Mother Present effect, the `permutationTest` function is used once again to develop and evaluate a Monte Carlo approximation to the null reference distribution for my observed mean difference statistic.

In this application the elements of the `mothr` vector are shuffled repeatedly to form a random sequence of 999 vectors, each used in the computation of another value of the treatment comparison statistic. However, each shuffling or rerandomization of treatment assignment must be constrained to take place separately and independently within blocks. This is accomplished again with the `group` argument in the `permutationTest` call. Here, the elements of the `pairs` vector identify the blocks (twin pairs) to which the corresponding elements of `mothr` belong. By specifying `group=pairs` only those elements of `mothr` with a common block-identifier in `pairs` are shuffled together. In this way each twin-pair is independently rerandomized.

Including the obtained mean attention difference score, only 27 of the 1,000 rerandomizations appropriate under the null treatment hypothesis lead to test statistics with values of 1.7 or greater, $P = 0.027$.

Restricted randomization. Restricted randomization (*RR*) represents a special form of the *RB* design. Here some cases are eligible for randomization to some but not to all treatments. They can be thought to form blocks corresponding to the range of treatments available to them and would be rerandomized, under a null hypothesis, only among those treatments contributing to that null hypothesis and for which they were eligible at the actual randomization.

Independent randomization. Independent randomization (*IR*) is a limiting form of *RB*. Here each case is randomized independently, a block of size one. Independent randomization leads to variation in initial treatment group sizes and these initial group sizes will not be preserved from one rerandomization to another. Because of the lack of control over group sizes *IR* designs should be used only sparingly.

Within Case Designs

The designs illustrated thus far are between groups (of cases) designs. Each case is randomized to a single treatment and treatment comparisons, perforce, are comparisons between cases. Randomization and rerandomization also work for within case (WC) designs. In these designs each available case receives all treatments on

offer and we make our treatment comparisons within cases. Rerandomization tests for *WC* designs arise as the result of the randomization of cases among alternative sequences of treatments. Usually these sequences are temporal though not always, e.g., in agricultural field trials they may be spatial. For increased experimental control, the cases may be blocked before they are randomized to treatment sequence, e.g., cases recruited into a study over a long period of time may be blocked by time of entry to protect against a confounding of time of entry with treatment sequence assignment that might result from complete randomization.

As the case receives several treatments, its response to treatment(s) can be conceptualized as multivariate, a vector of responses, one element for each *position* in that case's treatment sequence. It is the local population distributions of these vector-valued treatment responses that are the subject of within case null and substantive hypotheses. I'll describe this first in the context of the simplest randomized available case *WC* design.

The two-treatment, two-period design. Cases in this design receive one of two treatments in the first of two periods and then are switched or *crossover* to the other treatment in the second period. Sometimes known as the *AB/BA* design, available cases are *randomized* between two treatment sequences, Treatment *A* followed by Treatment *B* or Treatment *B* followed by Treatment *A*.

Let's assume that, in a particular study, the twenty cases available were randomized, ten to the *AB* sequence and ten to the *BA* sequence. The 20 cases define a local population, as they would for a between groups study. At the completion of the study each case provides an ordered bivariate response, the response to the first period treatment followed by the response to the second period treatment. If all 20 cases were to receive the *AB* sequence they would generate one population distribution for the bivariate response. If all 20 cases were to receive the *BA* sequence they would generate a second population distribution. The study itself provides a random sample of 10 observations from each of these two population distributions.

The null treatment effect hypothesis is that the two bivariate population distributions are identical. Under this hypothesis the bivariate observations are *exchangeable* between the two distributions, subject only to the consideration that 10 were sampled from one distribution and 10 from the other. Our alternative hypothesis depends, of course, on the purpose of the study. We might predict that Treatment *B* will produce larger responses than Treatment *A*. Such a prediction serves as the basis for a treatment comparison statistic, perhaps

$$s = (1/20)\left[\sum(\textit{Period 2} - \textit{Period1}|AB) + \sum(\textit{Period 1} - \textit{Period 2}|BA)\right],$$

the average of the $(B - A)$ responses.

A null reference distribution for the treatment comparison statistic can be produced if we reattribute the bivariate responses to treatment sequence population distributions, either in all $20!/(10! \times 10!)$ possible rerandomizations or for a long sequence of randomly chosen rerandomizations, and recompute the statistic for each rerandomization.

The two-treatment, multi-period design. A common research design in psychology calls for subjects to respond to a sequence of stimuli belonging to two classes, the researcher's interest in contrasting the accuracy or latency of responses to the two classes. The sequence usually is of some pseudo-random form to preclude subjects developing response strategies. Randomizing available subjects among alternative sequences can provide the basis for rerandomization-based local causal inference.

For example, 30 subjects might be randomized 15 to each of two 96-stimuli sequences, perhaps *abbababa*... *baba* and *baababab*... *abab*, where *a* and *b* designate stimuli of two distinct classes. Our null treatment (stimulus) effect hypothesis would be that the two population distributions $(N = 30)$ of 96-element vector-valued responses are identical. We could rerandomize the 30 observed vectors under that hypothesis, recomputing the appropriate *a/b* comparison statistic for each

rerandomization to provide a null reference distribution against which to evaluate the statistical significance of an observed stimulus comparison statistic.

Multi-treatment comparisons. In the two-treatment design the null treatment hypothesis does not impose any constraints on the rerandomizations; they can follow the original randomization strategy. With the randomization of cases among sequences of three or more treatments, however, this will not be the case. Table 6 is taken from Data Set 11 in Hand et al. (1994) and gives the (average) Flicker Fusion Frequencies (*FFF*) for nine volunteer subjects in a three-period three-treatment study.

Insert Table 6 about here.

The nine subjects were randomly matched with nine predetermined three-drug sequences. As there are six possible sequences of three drugs the design was an unbalanced one; three sequences appeared twice and the other three only once each.

The purpose of the study was to evaluate the effectiveness of a new antihistamine drug, Meclastine. It was anticipated that Meclastine should produce less drowsiness than an older antihistamine, Promethazine. The design of the study suggests two treatment comparisons:

(1) Meclastine (A) vs. Promethazine (C). Meclastine should produce higher *FFF*s than Promethazine. The null hypothesis is no difference.

(2) Meclastine (A) vs. Placebo (B). The null hypothesis, again, is no difference. The alternative hypothesis, for me, is that Meclastine should produce lower *FFF*s as there is no suggestion that any antihistamine would result in increased alertness. But, some researchers would prefer a non-directional alternative.

What does the first null treatment hypothesis imply about the exchangeability of trivariate observations, $(Day\ 1, Day\ 2, Day\ 3)$, among the six population distributions, one for each treatment sequence? To answer the question let's focus on one particular observation and how it would enter into the computation of the first treatment comparison statistic. Subject 2 was randomized to the sequence Promethazine-Meclastine-Placebo or *CAB* and produced a response vector

$(25.87, 26.63, 26.00)$. From what population distributions, other than that for *CAB*, could this vector have been sampled? Subject 2's contribution to the comparison of Meclastine and Promethazine is the difference $(A - C) = (\textit{Day 2} - \textit{Day 1})$ in *FFF*s and, correctly, takes no account of the *Day 3 FFF* as the response to the Placebo (B) plays no role in the first hypothesis. However, if this vector were to be reattributed to any of the population distributions *BCA, CBA, ABC*, or *BAC* the *Day 3* (Placebo) response would be drawn, inappropriately, into a comparison of treatments *A* and *C*. For purposes of testing the first hypothesis the vector contributed by Subject 2 can be rerandomized only to the treatment sequences *CAB* or *ACB*.

Similar analyses of other treatment response vectors would establish that, for purposes of building a null reference distribution for the first treatment comparison, vectors are exchangeable between population distributions *CAB* and *ACB*, between distributions *ABC* and *CAB*, and between distributions *BAC* and *BCA*.

This restriction on available rerandomizations leads us to rerandomize cases in this design as if the cases had been blocked for the original randomization. Subjects 1, 2, and 5 form one block. They would be rerandomized among sequences *ACB, CAB*, and *CAB*. Subjects 3, 4, and 9 form a second block. They would be rerandomized among sequences *CBA*, *ABC*, and *ABC*. And subjects 6, 7, and 8 form the third block. They would be rerandomized among sequences *BAC*, *BCA*, and *BCA*. I refer to these blocks as a symmetric partitioning of the cases in that the cases within blocks are rerandomized among exactly those sequences to which they were originally randomized.

The `group` argument for the `PermutationTest` function in ***S-Plus*** can be used to control the rerandomization by blocks.

For the second comparison, Meclastine vs. Placebo, the associated null treatment effect hypothesis would restrict rerandomization in a similar way. The nine cases again would be blocked into three sets of three, although the cases blocked together would differ from the first comparison.

For this nine-case study, either reference set of rerandomizations is quite limited. For each block of three cases the number of different rerandomizations, two to one sequence and one to a second sequence, is just $[3!/(2! \times 1!)] = 3$. With the three blocks rerandomized independently of one another, the total number of possible rerandomizations is $3^3 = 27$. The smallest *P*-value possible is just $(1/27) = 0.037$. Had 12 cases been employed, two randomized to each of the six possible medication sequences, the reference sets for each of the two hypothesis tests would have contained $[4!/(2! \times 2!)]^3 = 6^3 = 216$ elements, providing a distinctly more powerful experiment.

Split Plot (*SP*) Designs: Mixed Within and Between Cases Treatment Factors

As for the random sample ANOVA, it is possible to design randomization studies with both between and within cases treatment factors. To provide for the rerandomization analysis, cases must be doubly randomized. Cases could first be randomized among sequences of levels of the Within Factor (B) and then randomized among levels of Between Factor (A).

Comparisons among the levels of Factor *A* would be carried out as for between groups designs. Typically, the treatment response score for a case is the sum (or, average) of the responses over the levels of Factor *B*.

Comparisons among the levels of Factor *B* are carried out as for within cases designs. When rerandomizing, however, the cases should be treated as blocked on the levels of Factor *A*, so as to avoid confounding the effects of the two treatment factors.

Interaction of Treatment Factors in Randomized Studies

In some studies carried out as factorial or mixed ANOVAs the researcher is interested in testing for an interaction between the effects of two treatment factors. For studies based on the randomization of available cases the locus of any interaction can not be in the pattern of means of global population distributions. There are three randomization designs, however, for which treatment interactions, though differently located, can be hypothesized.

Within case interaction hypotheses. In a (2×2) factorial within-cases design an interaction may be present in the responses of an individual case. The cases are randomized among sequences of the four treatments A_1B_1, A_1B_2, A_2B_1, and A_2B_1. Each case receives all four treatments, allowing us to compute an interaction score for the *i*th case:

$$s_i = [(A_1B_1)_i - (A_1B_2)_i] - [(A_2B_1)_i - (A_2B_2)_i] .$$

The average of these provides an interaction test statistic. Under the no interaction hypothesis it should be close to zero. How can we generate a null reference distribution to test that hypothesis?

As in the *FFF* study there are restrictions on how the four-element response vectors can be rerandomized. Under the null interaction hypothesis the $(B_1 - B_2)$ differences can be permuted between the two levels of *A*. This translates into exchangeability between pairs of population distributions. For example, the population distributions for the two treatment sequences **I**: $(A_1 B_1, A_1B_2, A_2B_1, A_2B_2)$ and **II**: $(A_2 B_1, A_2B_2, A_1B_1, A_1B_2)$ would be exchangeable, the two levels of *A* having been interchanged in the sequences.

The randomized within case factorial design should provide as well for testing for *A* and *B* effects. Under the null *A* effect hypothesis the $(B_1 + B_2)$ sums can be permuted between the two levels of *A*. Thus, the same exchangeability rules hold for the *A* effect as for the $A \times B$ interaction. Under the null *B* effect hypothesis the $(A_1 + A_2)$ sums can be permuted between the two levels of *B*, if there is no $A \times B$ interaction. In the event of an interaction, the null *B* effect hypothesis is evaluated separately at the two levels of *A*, i.e., the $A_1$and A_2 responses can be permuted between the two levels of *B* as two separate analyses. In either instance, two additional treatment sequences would be needed: under a null *B* hypothesis, observations from the sequence **III**: $(A_1 B_2, A_1B_1, A_2B_2, A_2B_1)$ would be exchangeable with those from sequence **I** above and observations from **IV**: $(A_2 B_2, A_2B_1, A_1B_2, A_1B_1)$ would be

exchangeable with those from sequence **II**, the levels of *B* having been interchanged to produce the two new sequences.

Thus, cases must be randomized among a minimum of four well-chosen sequences for interaction and main treatment effect hypotheses to be tested. Alternative or additional sets of four sequences, each set constructed in accordance with the same considerations, could be used as well.

Split plot design interaction hypotheses. Interaction can also be defined at the individual case level in the mixed between and within case treatment factor design. I'll assume, as described for the earlier *SP* design, that each case is doubly randomized, first to one of the sequences (B_1, B_2) or (B_2, B_1) of the within case factor and then to either level A_1 or level A_2 of the between groups factor. We can obtain for each case a Factor *B* difference score, $s_i = (B_2 - B_1)_i$, and take as our interaction test statistic the average of these Factor *B* difference scores for those cases randomized to level A_2 subtracted from the average Factor *B* difference score for cases randomized to level A_1.

A null reference distribution is generated for this test statistic quite easily by rerandomizing the Factor *B* difference scores among levels A_1 and A_2, according to the original between cases randomization strategy.

A randomized blocks between groups interaction test. The within cases and split plot factorial designs require that we administer all levels of at least one treatment factor to each case. This will not be possible in all studies. What can we do about testing for an interaction in the randomized between groups 2×2 factorial design? If randomization among the four treatment levels, A_1B_1, A_1B_2, A_2B_1, and A_2B_1, is completely random there are no exchangeabilities among population distributions that can be used to construct a null reference distribution for a test of treatment interaction. However, if cases are randomized by blocks among the four treatment levels a test of interaction, *at the block level*, is possible. How interpretable such an interaction proves to be will depend on how homogeneous the cases are within each block. This approach may be quite satisfactory, say, where the cases

within a block are rat pups from the same litter. It may be less satisfactory if they are, say, male college students matched only on grade-point-average.

I'll outline the general approach. It is based on what we have seen for the split plot design. Again, for ease of illustration, we take the two treatment factors, *A* and *B*, to each have two levels. Four homogeneous cases make up each block. Within each block we randomize two cases to level A_1 and two to level A_2. The pair randomized to level A_1 is randomized again, one case to level B_1, the other to B_2. A second randomization is applied as well to the pair of cases at level A_2. The response scores for the four cases permit the computation of an interaction score *for each block of cases*, $[(A_1B_1) - (A_1B_2)] - [(A_2B_1) - (A_2B_2)]$, and these can be averaged over blocks to produce an interaction test statistic. To develop a null reference distribution for this statistic we can rerandomize, between levels of *A*, the pair of $(B_1 - B_2)$ differences in each block and recompute each block's interaction score.

Randomized Points of Intervention

The randomized within case designs are predicated on the assumption that we can randomize the order in which treatments are presented to the individual case. There are instances in which that may not be true, or, where such a research strategy will not answer the question in which we are interested. For example, we may want to know explicitly what happens when a case is moved from Treatment *A* to Treatment *B*. Is there a statistically significant increase, decrease, or change in the case's response to treatment?

Can a randomization strategy be used to answer this question? Yes, and the strategy is to randomize the point at which the case is moved from one treatment to the other, the point of intervention.

Consider this example. An instructor is interested in the impact on student attendance at her lectures of publishing lecture outlines on the web a day in advance. Her course meets 40 times in the course of a term and she can capture attendance numbers for each class meeting. Her research strategy is this. She randomly chooses

one lecture between the 10th and 29th to introduce the publication. Her test statistic is the difference in attendance between that at the lecture immediately preceding the chosen one and that at the fifth lecture after starting publication. To be specific, I'll assume her random choice of lecture was number 14. For each lecture beginning with lecture 14 a lecture outline is published on the preceding day. As her test statistic she subtracts from the attendance at lecture 13 the attendance at lecture 19. Is the change in attendance significantly large, or does it just reflect the random choice of a pair of lectures at which to count heads?

The null hypothesis is that the intervention has no effect, that the observed vector of attendance figures would be just as likely an outcome for any of the other starting dates that might have been randomly chosen. To test the null hypothesis our instructor will compare her attendance change against a null reference distribution of changes, computed by looking at attendance differences for all of the intervention points that might have been chosen. Because the choice of intervention point was limited to 20 possibilities, the null reference distribution will consist of 20 elements, one for each possibility: the change in attendance from lecture 9 to lecture 14, from lecture 10 to lecture 15, . . . , from lecture 27 to lecture 32, and from lecture 28 to lecture 33. If her change is large when compared with the distribution she may have a case for deciding that outline publication did influence attendance.

Alas, the small reference distribution precludes a *P*-value smaller than $(1/20) = 0.05$ in this hypothetical study. How could the study be made more powerful? Enlist a colleague, converting it from a one-case to a two-case study. If the two instructors independently choose one of 20 possible points of intervention, the null reference set for the test statistic–now the average of two changes in attendance–will be made up of $20^2 = 400$ elements.

In the laboratory setting, of course, the range of points of intervention can be dramatically larger. Rather than choosing among one of 20 days, one may choose a stimulus onset time randomly from among a thousand or more possibilities.

A Final Note on *P*-values

I have emphasized the natural use of rerandomization techniques to establish *P*-values for tests of hypotheses in randomized available case studies. This should not be read as support for dependence on *P*-values as the sole indicators of differential response to treatment. It should be expected of researchers that they report as well both the magnitudes of treatment effects in metrics that are clearly interpretable, using standardized effect size measures where appropriate, and the results of sensitivity analyses, assessing the impact of individual cases on those treatment effects. The latter precaution is important particularly in randomized available case studies. The set of available cases is not a random sample and, depending on how constituted, it may include one or more cases that are distinctly different, contaminating the local population.

Summary

Modern computing power has brought about a resurgence of interest in resampling approaches to statistical inference. One result is to free psychologists from continuing reliance on infinite population inference approximations, notably in the form of the normal-theory *t* and *F* tests, to summarize statistically studies that are based on the randomization of available cases. The result of adherence to the normal-theory approximations has been the over- or misinterpretation of statistical findings. By contrast, rerandomization tests build directly on the experimental randomization of subjects, clients, or animals to alternative treatments to provide the appropriate local causal inference. Rerandomization analyses can be developed quite readily for both between groups and within cases research designs.

Further practical applications of rerandomization can be seen in the recent texts of Edgington (1995), Good (1994 and 1999), Lunneborg (1999), Manly (1997), and Sprent (1998).

References

Box, G. E. P., Hunter, W. G. & Hunter, J. S. (1978). *Statistics for experimenters: An introduction to design, data analysis and model building*. New York: J. Wiley & Sons.

Cobb, G. (1997). *Introduction to design and analysis of experiments.* New York: Springer.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Marcel Dekker.

Fisher, R. A. (1935). *Design of experiments.* New York: Hafner.

Good, P. (1994). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York: Springer-Verlag.

Good, P. (1999). *Resampling methods: A practical guide to data analysis.* Boston: Birkhäuser.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. & Ostrowski, E. (1994). *A handbook of small data sets*. London: Chapman & Hall.

Hesterberg, T. (1999). *Beta test versions of S-Plus bootstrap tilting and permutation functions.* Personal communication.

Ludbrook, J. & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *The American Statistician, 52*, 127-132.

Lunneborg, C. E. (1999) *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Brooks-Cole.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology* (2nd ed.). London: Chapman & Hall.

MathSoft (1998). *S-Plus user's guide, version 4.5*. Seattle: Data Analysis Products Division, MathSoft, Inc.

Maxwell, S. E. & Delaney, H. D. (1990). *Designing experiments and analyzing data.* Pacific Grove, CA: Brooks/Cole.

Mead, R., Curnow, R. N. & Hasted, A. M. (1993). *Statistical methods in agriculture and experimental biology.* (2d. Ed.) London: Chapman & Hall.

Mehta, C. R. & Patel, N. R. (1995). *StatXact 3 for windows*. Cambridge, MA: Cytel Software Corporation.

Pitman, E. J. G. (1937a). Significance tests which may be applied to samples from any populations. *J. Roy Statist Soc, Suppl., 4*, 119-130.

Pitman, E. J. G. (1937b). Significance tests which may be applied to samples from any populations: Part II. The correlation coefficient. *J. Roy Statist Soc, Suppl., 4*, 225-232.

Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any population: Part III. The analysis of variance test. *Biometrika, 29*, 322-335.

Regan, P., Williams, M. & Sparling, S. (1977). Voluntary expiation of guilt: a field experiment. *J. of Personality and Soc. Psychology, 24*, 42-45.

Sprent, P. (1998). *Data driven statistical methods*. London: Chapman & Hall.

Welch, B. L. (1938) On the z-test in randomized blocks and Latin squares. *Biometrika, 29*, 21-52.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design.* (3rd Ed.) New York: McGraw-Hill.

Table 1

Heights of Self- and Cross-fertilized Plant Pairs

Pot Pair	Left Plant Type	Left Plant Height	Right Plant Type	Right Plant Height	Diff C − S
1	C	23.5	S	17.4	+6.1
2	C	12.0	S	20.4	−8.4
3	S	20.0	C	21.0	+1.0
4	C	22.0	S	20.0	+2.0
5	S	18.4	C	19.1	+0.7
6	S	18.6	C	21.5	+2.9
7	S	18.6	C	22.1	+3.5
8	S	15.3	C	20.4	+5.1
9	S	16.5	C	18.3	+1.8
10	S	18.0	C	21.6	+3.6
11	C	23.3	S	16.3	+7.0
12	C	21.0	S	18.0	+3.0
13	C	22.1	S	12.8	+9.3
14	S	15.5	C	23.0	+7.5
15	S	18.0	C	12.0	−6.0

Table 2

Reported Alcoholic Binges, by Orientation Condition

Moderation Orientation:	8	12	7
No Moderation Orientation:	9	11	13

Table 3

Rerandomizations of the Binge Reports

Customary Orientation	Moderation Orientation	Difference in Means
7,8,9	11,12,13	−4
7,8,11	9,12,13	−2.67
7,8,12	9,11,13	−2
7,8,13	9,11,12	−1.33
7,9,11	8,12,13	−2
7,9,12	8,11,13	−1.33
7,9,13	8,11,12	−0.67
7,11,12	8,9,13	0
7,11,13	8,9,12	0.67
7,12,13	8,9,11	1.33
8,9,11	7,12,13	−1.33
8,9,12	7,11,13	−0.67
8,9,13	7,11,12	0
8,11,12	7,9,13	0.67
8,11,13	7,9,12	1.33
8,12,13	7,9,11	2
9,11,12	7,8,13	1.33
9,11,13	7,8,12	2
9,12,13	7,8,11	2.67
11,12,13	7,8,9	4

Table 4

Finger Tapping Rates, College Males Randomized to Treatments

Treatment Group	Taps per Minute
0 mg Caffeine:	242, 245, 244, 248, 247, 248, 242, 244, 246, 242
100 mg Caffeine:	248, 246, 245, 247, 248, 250, 247, 246, 243, 244
200 mg Caffeine:	246, 248, 250, 252, 248, 250, 246, 248, 245, 250

Table 5

Attention Scores for Twins, by Mother Condition

Twin Pair:	1	2	3	4	5	6	7	8	9	10
Mother Present:	7	10	9	8	7	6	8	9	12	13
Mother Absent:	4	6	10	8	5	3	10	8	8	10

Table 6

Average Flicker Fusion Frequencies by Subject by Treatment

Subject No.:	Day 1	Day 2	Day 3
1	31.25 (A)	31.25 (C)	33.12 (B)
2	25.87 (C)	26.63 (A)	26.00 (B)
3	23.75 (C)	26.13 (B)	24.87 (A)
4	28.75 (A)	29.63 (B)	29.87 (C)
5	24.50 (C)	28.63 (A)	28.37 (B)
6	31.25 (B)	30.63 (A)	29.37 (C)
7	25.50 (B)	23.87 (C)	24.00 (A)
8	28.50 (B)	27.87 (C)	30.12 (A)
9	25.13 (A)	27.00 (B)	24.63 (C)

Medications: A = meclastine, B = placebo, C = promethazine

Figure 1

Analysis of the Guilt/Control Comparison of Helper Behavior

```
> helpv<- c(rep(1,3),rep(0,17),rep(1,11),rep(0,9))
> trtv<- c(rep("C",20),rep("G",20))

> hnh<- permutationTest(trtv,mean(helpv[trtv=="G"])
+ -mean(helpv[trtv=="C"]),alternative="greater",
+ B=999,trace=F)

> summary(hnh)

Summary Statistics:
      Observed      Mean     SE alternative p-value
Param      0.4 -0.003504 0.1594     greater   0.015

Empirical Percentiles:
       2.5%  5.0% 95.0% 97.5%
Param  -0.3  -0.3   0.3   0.3
>
```

Figure 2

Analysis of Finger Tapping Rates among Caffeine Groups

```
> mg0<- c(242,245,244,248,247,248,242,244,246,242)
> mg100<- c(248,246,245,247,248,250,247,246,243,244)
> mg200<- c(246,248,250,252,248,250,246,248,245,250)

> tapspeed<- c(mg0,mg100,mg200)

> caff<- c(rep(0,10),rep(100,10),rep(200,10))

> cnc<- permutationTest(caff,mean(tapspeed[caff>0],trim=0.20)-
+ mean(tapspeed[caff==0],trim=0.20),B=999,alternative="greater",
+ trace=F)

> summary(cnc)

Summary Statistics:
      Observed     Mean    SE alternative p-value
Param    2.667 -0.02986 1.018     greater   0.004

Empirical Percentiles:
        2.5%   5.0% 95.0% 97.5%
Param -1.917 -1.667 1.833 2.083

> shufgrp<- c(rep(1,10),rep(2,20))

> c12<- permutationTest(caff,mean(tapspeed[caff==200],trim=0.20)-
+ mean(tapspeed[caff==100],trim=0.20),group=shufgrp,B=999,
+ alternative="greater",trace=F)

> summary(c12)

Summary Statistics:
      Observed       Mean    SE alternative p-value
Param    1.833 -0.0008342 1.122     greater   0.055

Empirical Percentiles:
        2.5%   5.0% 95.0% 97.5%
Param -2.167 -1.833 1.833 2.167

>
```

Figure 3

Analysis of Mother Present Effect, Twin Attention Study

```
> atn<- c(7,10,9,8,7,6,8,9,12,13,4,6,10,8,5,3,10,8,8,10)
> mothr<- c(rep("P",10),rep("A",10))
> pairs<- rep(1:10,2)

> tmp<-permutationTest(mothr,mean(atn[mothr=="P"]-atn[mothr=="A"]),
+ group=pairs,B=999,alternative="greater",trace=F)

> summary(tmp)

Summary Statistics:
      Observed    Mean     SE alternative p-value
Param      1.7 0.01311 0.8125     greater   0.027

Empirical Percentiles:
       2.5%  5.0% 95.0% 97.5%
Param  -1.7  -1.3   1.3   1.7
>
```