TITLE: Bootstrap Inference for Local Populations

AUTHOR: Clifford E. Lunneborg, Ph D

AFFILIATION: Professor Emeritus, Departments of Statistics and Psychology, University of Washington, Seattle, WA

WORKSHOP PRESENTATION: This paper was presented at the 2000 FDA/Industry Workshop on Statistics, 14-15 September 2000, Bethesda, MD

REQUESTS: Prof C. E. Lunneborg, Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322 or e-mail: cliff@ms.washington.edu

RUNNING HEAD: Local Bootstrap Inference

Bootstrap Inference for Local Populations

Abstract

The randomized available case study, in which a nonrandom set of cases (patients, animals, laboratory runs) is randomized among two or more treatments, is a staple of biomedical research. Traditionally, such studies have been analyzed as though the cases were a random sample from an infinitely large population (Ludbrook and Dudley, 1998). The resulting statistical inferences address incorrect populations. More importantly, in the presence of response measurement error these inferences are inappropriate for the correct populations, understating the differential impact of treatment (Reichardt and Gollob, 1999). In this paper I develop and illustrate a nonparametric bootstrap approach to inference in such studies, an approach that is faithful to the local origins of the randomized cases and can account for the influence of measurement error.

KEYWORDS: available cases, bootstrap inference, measurement error, randomization, resampling

Introduction

Biomedical and biobehavioral scientists have long acknowledged the importance to good experimentation of randomizing the assignment of patients or animals to competing treatments or levels of treatment. Randomization is essential to causal inference, to our ability to argue that the observed differences in response to treatment owe to differences in treatment and not to some source confounded with treatment. The span of this inference is determined by how chance enters the design of the study. If the cases to be randomized are a random sample from a larger well-defined population of cases, the inference extends to that population. If, on the other hand, the cases are not a random sample, but those available to the scientist, e.g., patient volunteers or laboratory mice of an appropriate age and lineage, then the inference extends only to the local population made up of those available cases.

Scientists, and their statistical advisors, have been slow to recognize the restricted span of causal inference for the randomized available case study. For example, Ludbrook and Dudley (1998) summarized a survey of research reported in prominent biomedical journals. They report that fully 96% of the studies in those journals were randomized available case studies. Of these local population studies, however, 84% had been analyzed as if the cases had been random samples from infinitely large populations. The authors decried the use of large population analyses and advocated the use of randomization tests (Edgington, 1995; Good, 1999; and Lunneborg, 2000) as more appropriate to local causal inference.

Randomization tests are justified fully by the treatment randomization of cases and inferences based on those statistical tests are directed at the local population of cases, those randomized among competing treatments. While a shift in the analysis of randomized available case studies—away from, as examples, the analysis of variance (ANOVA) or the *t*-test and towards randomization tests—certainly should be encouraged, randomization tests have significant limitations. Here are three:

(1) Randomization tests rely on what has been called a "sharp null hypothesis" (Rubin, 1990a, 1990b, 1991), that any particular case would have responded identically to each of the treatments under comparison. This is a much sharper hypothesis, for example, than that the average response over all cases would be the same to each treatment and may be too sharp for many scientists.

(2) Not all hypotheses of interest are amenable to randomization tests. For example, in the two-way factorial design there is not a randomization test of the hypothesis of no interaction between the two experimental factors.

(3) The focus of the randomization test is null hypothesis testing. That is, the result of a randomization test is either a decision to reject or not the null hypothesis or a *p*-value assessing the strength of the experimental evidence against the null hypothesis. Randomization tests do not make it easy to report "effect sizes", estimates of treatment effect magnitudes in the population. Many scientists and statisticians would either complement or replace the *p*-value with a confidence interval (C.I.) for the population treatment effect (Wilkinson, et al. 1999).

In this paper I develop an alternative to randomization tests in support of local causal inference, inference from and appropriate to randomized available case studies. This alternative applies the logic of nonparametric bootstrap inference (Efron and Tibshirani, 1993; Davison and Hinkely, 1997; Chernick, 1999; and Lunneborg, 2000) to the randomized study and, thus, surmounts the shortcomings mentioned above. My approach will provide C.I.s for any desired treatment comparison, not just main effects. These C.I.s are specific to the local population, the relevant span of inference. And the scientist is given wide scope in defining treatment comparisons. Where required, the bootstrap C.I.s may be used to test null hypotheses or to generate *p*-values.

A distinct advantage of my bootstrap approach is that C.I.s and the related estimates of the standard errors (S.E.s) of sample treatment comparisons can be corrected for the effect of errors of measurement of the response to treatment. For the randomized available case study it is known (Neyman, 1990, and Reichardt and

Gollob, 1999) that errors of measurement lead to C.I.s and S.E.s that are too wide when estimated in the usual way, i.e., by assuming samples from infinite populations.

In the following sections I provide brief reviews of statistical approaches to causal inference and of bootstrap inference before describing bootstrap randomization. I finish the paper with an application of bootstrap randomization to a randomized case study.

Statistical Models for Randomized Case Studies

The problem of choosing a statistical basis for making causal inferences has been studied extensively by Rubin (1990a, 1990b, 1991) and my approach owes much to his formulation of the randomized case study.

Consider what is, perhaps, the simplest design: A set of n cases, randomly sampled from a population of size N, is randomly allocated by a researcher, n_A to an active treatment and n_P to a placebo treatment. Following treatment a response to treatment is measured, $y_{ij},\ i = 1, \ldots, n,\ j = A$ or P depending on treatment assignment.

For this case population and the researcher's choice of treatments, Rubin defines a treatment response population distribution, one that I represent as an array Y with N rows and two columns. The rows correspond to the cases and the columns to the two treatments. The entries are responses to treatment and, for the ith case, $i = 1, \ldots, N,$ the ith row of the population distribution gives a measure of the differential effect of the active treatment on that case: $(Y_{iA} - Y_{iP})$. The scientist is interested in the magnitude of this treatment effect at the population level. An often compelling choice for the population level effect is the average of the individual case effects: $\mu_{A-P} = (1/N)\sum_{i=1}^{N}(Y_{iA} - Y_{iP})$.

Although the randomized case study does not provide the values of both Y_{iA} and Y_{iP} for any case, the design does allow us to estimate μ_{A-P} unbiasedly. Here is a summary of the rationale.

1. The population level treatment effect can be re-expressed as

$$\mu_{A-P} = (1/N)\sum_{i=1}^{N} Y_{iA} - (1/N)\sum_{i=1}^{N} Y_{iP} = \mu_A - \mu_P,$$

the difference in the mean responses to treatment over the population.

2. When a random sample of n cases from some case population is randomly split to form two treatment groups, those treatment groups are both random samples from that case population.

3. As a result, the treatment responses of the n_A cases randomly assigned to the active treatment are a random sample of the population of responses to that treatment. The mean of those n_A responses provides an unbiased estimate of the mean of the population of responses: $\widehat{\mu}_A = (1/n_A)\sum_{i=1}^{n_A} y_{iA} = \overline{y}_A$. Similarly, the mean of the responses of the n_P cases randomized to the placebo treatment, $\overline{y}_P$, provides an unbiased estimate of the mean of the population of placebo treatment responses.

4. The difference between these two unbiased estimates of population means provides, in turn, an unbiased estimate of the difference in population means or of the population mean of differences in treatment response: $\widehat{\mu}_{A-P} = (\overline{y}_A - \overline{y}_P)$.

The estimate of the population mean, μ_{A-P}, is unbiased—it is as likely to be too large as too small—but the scientist needs to know as well how close to the population mean the estimate is likely to be. How accurate is this method of estimation? And, knowing the accuracy of estimation, can the scientist be confident, for example, that μ_{A-P} is positive rather than negative? These are questions posed of statistical inference.

Rubin (1990a, 1990b, 1991) distinguishes among four statistical models for answering such questions from randomized case studies, where those answers take on causal significance. Briefly, I characterize these models as follows.

Superpopulation Repeated Samples Inference

This is the near-universal model of choice of researchers and their statistical advisors. A basis for statistical inference is created by modeling the distribution of the value of our active-placebo treatment comparison statistic, $(\overline{y}_A - \overline{y}_P)$, when computed for each possible random sample of cases from the infinitely large

population that, it is assumed, was sampled randomly to provide cases for the actual study. The standard deviation of the resulting sampling distribution is known as the standard error of the treatment comparison statistic, *S.E.*$(\overline{y}_A - \overline{y}_P)$, and assesses the sample to sample variability in that statistic. How the contents of this sampling distribution are spread out around the value of the population mean treatment effect tells us how to construct a confidence interval for that population effect, *C.I.*(μ_{A-P}), based on the results of a single study. S.E.s and C.I.s are used widely to assess the accuracy of our estimate of a treatment effect.

When the model used to generate the sampling distribution of $(\overline{y}_A - \overline{y}_P)$ includes a null hypothesis of equal treatment effectiveness—i.e., that $\mu_{A-P} = 0$—we refer to the sampling distribution as a null sampling distribution and use it to test that null hypothesis or to assign a *p*-value to the observed value of $(\overline{y}_A - \overline{y}_P)$.

The infinite population (or, superpopulation) basis of the random samples facilitates the modeling of sampling distributions, providing the statistician with knowledge of two important consequences:

1. The two samples of response to treatment—the responses of the n_A cases randomized to the active treatment and the responses of the n_P cases randomized to the placebo treatment—are statistically independent of one another (e.g., if the responses of those randomized to the active treatment are high, this has no implication for the level of response of those randomized to the placebo treatment).

2. Within each of the treatment groups, the responses are independent of one another (e.g., the responses of the n_A cases randomized to the active treatment are, essentially, random selections from an unchanging distribution of responses).

The sampling distribution for $(\overline{y}_A - \overline{y}_P)$ is estimated, almost always, with the help of the assumptions that the distribution of responses to treatment in the population follows that of a normal random variable for both the active and placebo treatment and, further, that both distributions have the same variance. These assumptions underly the two-sample *t*-test and its extension to the multi-treatment analysis of variance (ANOVA) and to normal linear regression.

The assumptions of normality of distribution and of homogeneity of variance can be avoided in nonparametric bootstrap inference for the infinite population (Efron & Tibshirani, 1993 and Davison & Hinkley, 1997).

It remains, of course, the apparently unwarranted assumption that the researcher's sample of convenience (or local population) is a random sample from an infinite population of cases that is at the heart of the call by Ludbrook & Dudley (1998) for the replacement of superpopulation models of inference, whether parametric or nonparametric, in the analysis of the randomized available case study.

Randomization Tests of Sharp Null Hypotheses

A second of Rubin's (1991) models for inference is the one advanced as an alternative to superpopulation inference by Ludbrook & Dudley (1998). A null randomization distribution for the treatment comparison statistic, $(\overline{y}_A - \overline{y}_P)$, is developed by computing the value of that statistic for all possible randomizations of the $(n_A + n_P)$ cases, n_A to the active treatment and n_P to the placebo treatment, under the null hypothesis that the response of the *i*th case would be exactly the same to either treatment (Edgington, 1995). A small *p*-value (or null hypothesis rejection) is associated with the observed $(\overline{y}_A - \overline{y}_P)$ taking an extreme value in that distribution.

The sharpness of this null hypothesis, as noted earlier, can limit the typical researcher's interest in this mode of inference. As well, the randomization test provides no direct evidence of the accuracy of the randomization study estimate of the treatment effect over the local population. These two factors may help explain the limited implementation of randomization hypothesis tests in the analysis of randomized available case studies.

Bayesian Predictive Inference

In a third, Bayesian paradigm, a posterior probability distribution is developed for the range of candidate values for the (local) population treatment effect, μ_{A-P}. The more concentrated this distribution, the more certainty we have about the magnitude of this effect. This posterior distribution depends not only on the responses to treatment collected in the randomized available case study, but on the researcher's

prescription of an initial or prior probability distribution for μ_{A-P}; what values, in advance of the study, are likely ones for μ_{A-P}. The indeterminacy of this prior specification, perhaps, is what leads Rubin (1991) to worry that the approach may be more easily "abused in practice" (Rubin, 1991, p. 1214).

Randomization Repeated Samples Inference

A fourth and final model for causal inference is the one I exploit in this paper. This model is based on the same concepts described for superpopulation repeated samples inference—i.e., sampling distributions, S.E.s, and C.I.s. The distinction is that the sampling distribution of $(\overline{y}_A - \overline{y}_P)$ is one based on repeated samples that are obtained now by randomly randomizing the cases making up a local case population, rather than by randomly sampling an infinitely large case population. In this way, the sampling distribution reflects the actual source of randomness in the randomized available cases study.

Superpopulation repeated samples inference requires that we estimate the distribution of responses to each treatment in the infinitely large case population. For the *t*-test or ANOVA these distribution estimates have a "normal" shape, a common variance (estimated from the sample responses), and, for purposes of hypothesis testing, a common mean as well. The population response distributions underlying bootstrap nonparametric superpopulation inference, by comparison, are each estimated by (an infinite number of) copies of the samples of responses from those distributions (Lunneborg, 2000).

Randomization repeated samples inference similarly depends on an estimate of the population response distribution. This estimation task is constrained by two factors:

1. As the local case population is completely exhausted by the randomization to treatment $(N = n_A + n_P)$, it is not realistic to estimate the active and placebo response distributions separately. To accommodate the lack of independence of the randomization samples, we should estimate the joint distribution in the population of the responses to the two treatments.

2. As the local case population may be decidedly finite—e.g., $N = 100$ when $n_A = n_P = 50$, a parametric estimate, particularly one assuming the population distribution of responses to be continuous (e.g., bivariate normal) may be unrealistic as well.

In view of these restrictions, the estimates of the population response distributions that I develop below will be bivariate and they will take the form of nonparametric maximum likelihood (NPML) estimates (Efron & Tibshirani, 1993). The NPML estimate is widely used as a basis for bootstrap inference and I provide an introduction to this in the following section.

Nonparametric Bootstrap Inference

Applications of what has come to be known as the bootstrap approach to statistical inference have increased impressively since Efron (1979) introduced the name. The hallmark of bootstrap inference is that it provides a more widely applicable basis for estimating sampling distributions. Previously, statisticians were dependent upon mathematical analysis—and, frequently, parametric or sample size assumptions essential to those analyses—to estimate the mathematical characteristics of the sampling distribution of a sample-based estimate or test statistic. The bootstrap substitutes computing power for mathematical analysis, freeing the statistician of both analysis and assumptions and opening up the range of statistical inferences for which sampling distributions can be estimated (Chernick, 1999; Lunneborg, 2000).

Bootstrap inference rests on the following ideas, illustrated for simplicity by a single treatment experiment:

1. A random sample of n cases from a case population of size N is administered treatment.

2. The n measured responses to treatment, y, constitute a random sample from the N responses making up the otherwise unobservable population response distribution, Y.

3. The population response distribution, Y, is characterized by a parameter θ, unknown and to be estimated (e.g., θ might be the population distribution mean).

4. The sample of responses to treatment, y, provides an estimate, t, of θ. (If t is computed from y by the same rule as would yield θ from Y, it is termed a plug-in estimate. The sample mean is a plug-in estimate of the population mean.)

5. Let $F(t|Y,\theta,n)$ denote the unknown sampling distribution of t, based on all possible random samples of size n from the population Y with parameter θ. Knowledge of this distribution would allow us to report the accuracy of our estimate, t, of θ in terms of the S.E. of t or a C.I. for θ.

6. Let $\widehat{Y}$ be a numeric estimate of Y. This estimate is a collection of N estimated responses to treatment.

7. The estimated population distribution, $\widehat{Y}$ is characterized by a parameter $\widehat{\theta}$, calculated from $\widehat{Y}$ by the same rule we would use (if we could) to compute θ from Y.

8. Let y_1^* be a random sample of size n from $\widehat{Y}$ and t_1^* be an estimate of $\widehat{\theta}$ computed from that random sample. This estimate is computed by the same rule as used to compute t from y. y_1^* is known as a (first) bootstrap sample.

9. As $\widehat{Y}$ is a collection of values, we can draw not just one, but many bootstrap samples, the number limited only by computational speed.

10. Let y_b^* and t_b^* denote the bth in a random sequence $b = 1, 2, \ldots, B$ of bootstrap samples and estimates computed from those samples. B might be on the order of $5,000$.

11. Let $F_B^*\left(t_b^*|\widehat{Y},\widehat{\theta},n\right)$ designate the bootstrap sampling distribution of t_b^*. This is the distribution of B values of t_b^*, each computed from a randomly chosen sample of n observations from $\widehat{Y}$. Each y_b^* is to be sampled from $\widehat{Y}$ by the same random process as was used to sample y from Y.

The central tenet of bootstrap inference is that, after centering the distributions about their parameters, the computable bootstrap sampling distribution,

$$F_B^*\Big[(t_b^* - \widehat{\theta}) \mid \widehat{Y},\widehat{\theta},n\Big],$$

provides an appropriate basis for estimating *S.E.*(t) and the limits of a C.I. for θ, quantities that would be defined exactly by the unknowable

$$F[(t-\theta)\,|\,Y,\theta,n].$$

Experience has taught that somewhat better estimates can be obtained, where θ is a location parameter such a mean or difference between means, if the Studentized bootstrap sampling distribution,

$$F_B^*\left[\frac{(t_b^*-\widehat{\theta})}{\widehat{S.E.}_b^*(t^*)}\right],$$

is used as an estimate of the Studentized sampling distribution,

$$F\left[\frac{(t-\theta)}{\widehat{S.E.}(t)}\right],$$

(Efron & Tibshirani, 1993; Davison & Hinkley, 1977). In these expressions, $\widehat{S.E.}(t)$ is an estimate of *S.E.*(t) computed from y while $\widehat{S.E.}_b^*(t^*)$ is an estimate of *S.E.*(t^*) computed from y_b^*. The Studentizations are most easily accomplished where a closed form expression is available for the estimation of the S.E.s. As noted below, I take advantage of such an expression for *S.E.*$(\overline{y}_A-\overline{y}_P)$ in the randomized available case study. The term Studentization honors the statistician Gossett who, writing under the name Student, derived the family of ***t***-distributions to describe the sampling distribution of $\boldsymbol{t}=(\overline{y}-\mu)/\sqrt{(\widehat{\sigma}^2/n)}$ where $\overline{y}$ is the mean of n randomly chosen observations from a normal distribution with mean μ and $\widehat{\sigma}^2$ is an unbiased estimate of the variance of that normal distribution, computed from those same sample observations.

All bootstrap inference rests, of course, on having $\widehat{Y}$, the numeric estimate of the unknown population distribution Y. Most applications utilize a NPML estimate. A small example illustrates the principle. Let $y=(3,9,7)$ be a random sample of $n=3$ observations from a population distribution, Y, containing a total of $N=6$ observations. How should we estimate Y? In fact, we know n of the values, from the sample. How should we estimate the remaining $(N-n)$ values?

$$\widehat{Y}=(3,9,7,?,?,?)$$

NPML estimation is based on the following idea. We seek that $\widehat{Y}$ from which it is most likely that we would obtain $y = (3, 9, 7)$ when drawing three observations at random and without replacement from among a total of six. There are exactly

$$N!/[n! \times (N - n)!] = 6!/(3! \times 3!) = (6 \times 5 \times 4)/(3 \times 2 \times 1) = 20$$

distinct random samples of size three. How many of the 20 samples can contain the values 3, 9, and 7? If, for example, $\widehat{Y} = (3, 9, 7, 4, 1, 2)$ there is only once chance in twenty of obtaining a sample consisting of a "3", a "9", and a "7." However, if $\widehat{Y} = (3, 9, 7, 3, 1, 2)$ the chances of a sample with those three values becomes two in twenty and if $\widehat{Y} = (3, 9, 7, 3, 9, 7)$ those chances increase to eight in twenty. This is a maximum. There is no set of $N = 6$ values for which the chances of a random sample of $n = 3$ consisting of the values in y can exceed eight.

These results generalize beyond $N = 6$ and $n = 3$. The NPML estimate of Y consists of $k = (N/n)$ copies of y. If k is not an integer, there is no unique $\widehat{Y}$ and we are advised to cycle among alternative estimates (Booth, Butler, & Hall, 1994). If k is large enough—i.e., greater than 20, it is practicable to replace $\widehat{Y}$ with a single copy of y, drawing bootstrap samples by sampling with replacement from the n observations making up y rather than by sampling without replacement from the N observations making up $\widehat{Y}$. This is equivalent to assuming the population to be infinitely large.

I turn now to the focus of this paper, adapting bootstrap inference to the random sampling associated with the randomized available case design.

Randomization NPML Estimation and Bootstrap Inference

I begin with the problem of defining an NPML estimate of the bivariate population response distribution for the randomized available case design.

NPML Estimation

Again, I use a small example to illustrate my adaptation of bootstrap inference. Assume I have randomized six patients, three apiece, between an active and a placebo treatment. Before assessing their response to treatment we have no information about the contents of the local bivariate population response distribution, Y, as reflected in the last two columns of the following matrix:

Patient	Treatment	Active Response	Placebo Response
1	P	?	?
2	A	?	?
3	A	?	?
4	A	?	?
5	P	?	?
6	P	?	?

Once the patients' responses to the treatments to which they were randomized are known, we have partial knowledge of Y:

Patient	Treatment	Active Response	Placebo Response
1	P	*a*	3
2	A	4	*b*
3	A	6	*c*
4	A	11	*d*
5	P	*e*	9
6	P	*f*	7

To fully define $\widehat{Y}$, our estimate of Y, we must estimate the six remaining unknown responses, *a*, *b*, *c*, *d*, *e*, and *f*. How can we insure that the resulting $\widehat{Y}$ is an NPML estimate of Y?

In this example, there are

$$\frac{(n_A + n_P)!}{n_A! \times n_P!} = \frac{6!}{3! \times 3!} = 20$$

different ways in which the six patients can be randomized between the two treatments. The NPML goal is a set of (6×2) responses that maximize the chances among these 20 that we observe responses of 3, 9, and 7 for the patients randomized to the placebo treatment and responses of 4, 6, and 11 for the patients randomized to the active treatment.

Our NPML estimate of the bivariate Y must satisfy the univariate rule introduced earlier:

1. Replace *a*, *e*, and *f* with a second copy of the values 4, 6, and 11 and replace *b*, *c*, and *d* with a second copy of the values 3, 9, and 7.

As well, the estimate must insure that:

2. The replacement values are consistently aligned with the observed values.

That is, if we replace *a* with a 4, we must also replace *b* with a 3. The resulting NPML estimate will consist, in this case, of two copies of three bivariate observations. Here is an NPML $\widehat{Y}$, estimated values in italics :

Patient	Treatment	Active Response	Placebo Response
1	P	*4*	3
2	A	4	*3*
3	A	6	*7*
4	A	11	*9*
5	P	*11*	9
6	P	*6*	7

And, here is a second NPML $\widehat{Y}$:

Patient	Treatment	Active Response	Placebo Response
1	P	*6*	3
2	A	4	*9*
3	A	6	*3*
4	A	11	*7*
5	P	*4*	9
6	P	*11*	7

There are, in fact, six NPML estimates of Y. For each, the chances of a randomization of patients resulting in "active" responses of 4, 6, and 11 and "placebo" responses of 3, 7, and 9 are eight in twenty. These chances cannot be bettered.

The Importance of the Active-Placebo Response Correlation

The six NPML estimates are alike in their distribution of active response and in their distribution of placebo responses. That is, they have the same marginal distributions. They differ in how the two distributions are aligned with one another. Or, in perhaps more familiar terms, they differ in the degree of correlation between the two sets of responses, their estimates of ρ_{AP}. For our first example, the estimated correlation between the two is almost perfect, $\widehat{\rho}_{AP} = 0.9398$, while for the second it nearly vanishes, $\widehat{\rho}_{AP} = -0.0908$.

The multiplicity of NPML estimates and the variation in the between treatments correlation of responses to treatment comes about because the randomized available case study provides no information about the magnitude of the correlation. No cases receive both treatments and, hence, there can be no estimate of that

correlation. Does the multiplicity of estimates matter? And, if it does, how should we proceed?

That the value of this correlation does matter to statistical inference was recognized in the 1920s by the eminent statistician Jerzy Neyman (Neyman, 1990) and was verified independently by psychological scientists Reichardt and Gollob (1999). From the latter paper, we have this formula for the sampling variance (or, square of the S.E.) of the randomized study estimate of an active treatment effect:

$$S.E.^2(\overline{y}_A - \overline{y}_P) = \left(\frac{\sigma_A^2}{n_A} + \frac{\sigma_P^2}{n_P}\right) - \frac{(\sigma_A - \sigma_P)^2}{n_A + n_P} - \frac{2(1-\rho_{AP})\sigma_A\sigma_P}{n_A+n_P}$$

where σ_A^2 and σ_P^2 are the variances of the responses to the two treatments in the population distribution. The first term on the right of this equation provides the basis for the classical, infinite population, estimate of the S.E. The second and third terms tell us that this classical estimate will be too large in the randomized available case study if the two variances differ or if the correlation between the two treatments is less than $+1$. In fact, $S.E.^2(\overline{y}_A - \overline{y}_P)$ is a decreasing linear function of ρ_{AP}. When $n_A = n_P$ and $\sigma_A^2 = \sigma_P^2$ the S.E. falls by 29% as ρ_{AP} decreases from $+1$ to 0 and continues to fall to 0 as ρ_{AP} reaches -1. The widths of C.I.s would shrink at essentially the same rate.

Estimating ρ_{AP}

Clearly, the magnitude of ρ_{AP} makes a difference to the sizes of S.E.s and C.I.s in the randomized available cases study. What can the researcher do to take into account the size of this correlation?

1. The correlation is almost certain to be positive and substantial in magnitude. Indeed, if we had reason to suspect that the responses to the active and placebo treatments would order the cases in our local population in substantially different ways, we would choose a research design that permits us to do more than estimate a mean effect of the active treatment in the population. Response of a case to the active treatment would be dependent on more than the active treatment and the response of

that case to a placebo treatment. Identifying what that "more" could be would necessarily become an important goal of the study.

2. In the absence of an estimate of this positive correlation, a reasonable strategy, the one originally proposed by Neyman (see Neyman, 1990, and the accompanying commentary by Rubin, 1990a), is to accept the more conservative results associated with using the maximum correlation NPML estimate as the basis for generating a bootstrap sampling distribution and from that, S.E.s and C.I.s.

3. Earlier research in which both of our treatments—or, two similar treatments—were administered to the same cases might give us an estimate of the size of the correlation, $\widehat{\rho}_{AP}$. Care is required here. The correlation based on observed responses is almost certain to underestimate the (hypothetical) ρ_{AP}. But, by what amount?

4. An indirect estimate of ρ_{AP}, or of an upper bound to its value, may be available, even though no cases will have responded to both treatments. The correlation of two sets of measures is restricted by the accuracy of those measures. An upper bound for the value of ρ_{AP} is given by the psychometric reliability of the treatment response measure in the local population (Reichardt & Gollob, 1999). Basically, the psychometric reliability is the correlation between two sets of measurements, the two collected for the population under near-identical circumstances (e.g., Anastasi, 1982). There are at least three instances in which this response measure reliability for the local population might be estimated.

5. In our randomization study, the treatment response measure for a case could be the result of summing (or, averaging) responses to a sequence of, for example, stimuli, trials, or items. Let r_{oe} be the correlation, say, for the placebo treatment cases, between the sums of the odd-numbered and even-numbered responses in the sequence. This correlation between half-length measure measures can be promoted to a full-length reliability estimate—and, hence, to an estimated upper bound for ρ_{AP}—through application of what is known as the Spearman-Brown prophecy formula (Anastasi, 1982) :

$$\widehat{max}(\rho_{AP}) = \frac{2\, r_{oe}}{1 + r_{oe}} \, .$$

6. The measurement accuracy of the response to treatment may be known. In particular, we may know the variability of the measure for a typical case of the kind in our local population. For example, we might know from an earlier investigation the variance of systolic blood pressure when measured repeatedly on the same patient. A related definition of psychometric reliability is that is the (estimated) proportion of the overall population variance of a response measurement that is true between-case variance, eliminating within-case variance (Anastasi, 1982). Let S_W^2 be the within-case variance of the response measure, averaged over a set of cases not unlike those used in our randomized treatment study. Let S_P^2 be the observed between-cases variance of the response measure for the cases randomized to the placebo treatment. Then, we have as a second estimate of the upper bound of ρ_{AP}:

$$\widehat{max}(\rho_{AP}) = \frac{S_P^2 - S_W^2}{S_P^2} \, .$$

7. The within-case variance, S_W^2,—or, its square root, the standard error of measurement—is held to be relatively constant from one population to another, provided the two are made up of the same kind of cases (Anastasi, 1982). This permits us to estimate the reliability of a response measure in a new, local population from a reliability study carried out on a sample of cases from a second, similar population. Let S_P^2 be defined as above and let r_{XX} and S_X^2 be the reliability estimate and observed between-cases variance of the response measure obtained from an earlier reliability study. Thus, a third estimate of a maximum value for ρ_{AP} is:

$$\widehat{max}(\rho_{AP}) = \frac{S_P^2 - (1 - r_{XX})S_X^2}{S_P^2} \, .$$

If, for example, we have a reliability estimate for a measure of cognitive ability (preferably of the split-half or single-administration variety, rather than one based on measures of the performance obtained at two different times) based on one sample of male college sophomores, we could use that to estimate the reliability in a new sample

of male college sophomores, so long as we have estimates of the variances in that cognitive performance for the two samples. Typically, reliability studies are carried out on samples with wider variability than is found in a sample of experimental volunteers and the two reliability estimates will differ.

Interpolating a S.E. and C.I. Limits

How can we use an estimate of, or limiting value for, ρ_{AP} once we have found such a value? We would be exceedingly fortunate to find exactly that correlation between the two sets of responses making up any one of the NPML estimates of the bivariate Y. Rather, I propose using $\widehat{\rho}_{AP}$ or $\widehat{max}(\rho_{AP})$—where such an estimate is available—to interpolate the value of the S.E. or of limits to a C.I. I proceed as follows:

1. Form the NPML bivariate population estimate with maximum between-treatments correlation. To do this I create a basic building block of bivariate observations by aligning the active and placebo responses by magnitude. Let r_{max} be the between-treatments correlation in that estimated population.

2. Draw bootstrap samples from this estimated population, create a bootstrap sampling distribution and use this to estimate a maximum S.E. for the treatment comparison statistic and outer limits for a C.I. for the population treatment effect.

3. Form a NPML bivariate population estimate for which the between-treatments correlation is near-zero. To do this I create a second basic building block of bivariate observations, this time randomly pairing the active responses with the placebo responses. Let r_{min} be the between-treatments correlation in this estimated population.

4. Draw bootstrap samples from this estimated population, create a second bootstrap sampling distribution and use this to estimate a minimum S.E. for the treatment comparison statistic and inner limits for a C.I. for the population treatment effect.

5. The Reichardt & Gollob (1999) formula establishes that the square of $S.E.(\overline{y}_A - \overline{P})$ and, hence, the square of the width of $C.I.(\mu_{A-P})$ are both linear

functions of ρ_{AP}. Thus, I can linearly interpolate between the minimum values of these squares, those associated with the r_{min} bootstrap sampling distribution, and the maximum values of these squares, those associated with the r_{max} bootstrap sampling distribution, finding a $S.E.^2(\overline{y}_A - \overline{P})$ and squared $C.I.(\mu_{A-P})$ width appropriate to our $\widehat{\rho}_{AP}$ or $\widehat{max}(\rho_{AP})$. These are then translated into S.E. and C.I. limit estimates. As a detail, I advocate interpolating the two halves of the confidence interval separately as the interval may not be symmetric or, more likely, a researcher's interest may center on a lower or upper confidence bound for the population treatment effect, rather than on an interval.

Forming Randomization Bootstrap Sampling Distributions

Steps 2 and 4 in the above flow of analysis call for forming bootstrap sampling distributions and, from these, estimates of $S.E.(\overline{y}_A - \overline{y}_P)$ and of the limits to a $C.I.(\mu_{A-P})$. Here I review the construction of these sampling distributions.

It is important to note that bootstrap samples are to be drawn from the NPML estimated bivariate distribution in exactly the same manner as the observed response samples were drawn from the, otherwise unobservable, "real world" bivariate distribution. That is, each bootstrap sample is the result of randomizing the N $(N = n_A + n_P)$ cases making up the local population, n_A to an "active" treatment and n_P to a "placebo" treatment, and then recording their responses from the "active" or "placebo" columns of the estimated response distribution, as appropriate.

The responses of the bth bootstrap sample provide what is needed to compute both $t^*_{1b} = (\overline{y}_A - \overline{y}_P)^*_b$ and $t^*_{2b} = (t^*_{1b} - \widehat{\mu}_{A-P})/\widehat{S.E.}(\overline{y}_A - \overline{y}_P)^*_b$. The first statistic is the difference in treatment means for that bootstrap sample. The second statistic is the Studentized form of this difference, the result of dividing the deviation of this bootstrap estimate from the parameter it estimates by an estimate of the S.E. of the bootstrap estimate.

The bootstrap population parameter, $\widehat{\mu}_{A-P}$, is the average of the differences in response to the two treatments as given in the NPML estimate, $\widehat{Y}$. The S.E. estimate, $\widehat{S.E.}(\overline{y}_A - \overline{y}_P)^*_b$, is computed from the bth bootstrap sample using one of two versions

of the Reichardt & Gollob formula given earlier, in which population variances are replaced with their bootstrap randomization sample estimates, $\widehat{\sigma}^{*2}_{bA}$ and $\widehat{\sigma}^{*2}_{bP}$. For bootstrap samples drawn from the maximum-correlation NPML $\widehat{Y}$, ρ_{AP} is assumed to be $+1$ in the formula. As a result, the S.E. estimate takes this form:

$$\widehat{S.E.}(\overline{y}_A - \overline{y}_P)^*_b = \sqrt{\left(\frac{\widehat{\sigma}^{*2}_{b\,A}}{n_A} + \frac{\widehat{\sigma}^{*2}_{b\,P}}{n_P}\right) - \frac{\left(\widehat{\sigma}^{*}_{b\,A} - \widehat{\sigma}^{*}_{b\,P}\right)^2}{n_A + n_P}} \ .$$

Where interpolation of results is required, bootstrap samples must be drawn as well from the minimum-correlation NPML $\widehat{Y}$. Then, ρ_{AP} is assumed to be 0, and the S.E. estimate then takes this form:

$$\widehat{S.E.}(\overline{y}_A - \overline{y}_P)^*_b = \sqrt{\left(\frac{\widehat{\sigma}^{2}_{bA}}{n_A} + \frac{\widehat{\sigma}^{2}_{bP}}{n_P}\right) - \frac{\left(\widehat{\sigma}_{bA} - \widehat{\sigma}_{bP}\right)^2}{n_A + n_P} - \frac{2\widehat{\sigma}_{bA}\widehat{\sigma}_{bP}}{n_A+n_P}} \ .$$

The bootstrap sampling distributions of both t^*_1 and t^*_2 are built up over $b = 1, 2, \ldots, B$ randomly chosen randomizations of the local population, a process that may need to be repeated, drawing treatment responses from the minimum-correlation $\widehat{Y}$ the second time around.

Bootstrap S.E. and Percentile-*t* C.I. Estimates

The bootstrap sampling distributions for t^*_1 and t^*_2 can then be used to estimate a S.E. for the treatment comparison statistic, $(\overline{y}_A - \overline{y}_P)$, and to estimate a C.I. for the local population treatment effect, μ_{A-P}. The bootstrap estimate of $S.E.(\overline{y}_A - \overline{y}_P)$ is given by

$$\widehat{S.E.}(\overline{y}_A - \overline{y}_P) = \sqrt{\left(\frac{1}{B-1}\right)\sum_{b=1}^{B}(t^*_{1b} - \widehat{\mu}_{A-P})^2} \ ,$$

effectively the standard deviation of the B values of $t^*_{1b} = (\overline{y}_A - \overline{y}_P)^*_b$ (e.g., Efron & Tibshirani, 1993).

This estimated S.E. provides a measure of the variability in the value of $(\overline{y}_A - \overline{y}_P)$, variability resulting from the randomization of cases between the two treatments. The estimated S.E. also plays a role in our estimation of a C.I. for μ_{A-P}.

The C.I. estimates described here are what are called percentile-*t* or bootstrap-*t* C.I.s (e.g., Lunneborg, 2000). Such estimates are justified by the finding noted earlier that, for estimates of location parameters, estimates, the bootstrap sampling distribution

$$F_B^*(t_2^*) = F_B^*\left[\frac{(\overline{y}_A - \overline{y}_P)_b^* - \widehat{\mu}_{A-P}}{\widehat{S.E.}(\overline{y}_A - \overline{y}_P)_b^*}\right],$$

can be expected to provide a good estimate of the Studentized sampling distribution,

$$F\left[\frac{(\overline{y}_A - \overline{y}_P) - \mu_{A-P}}{\widehat{S.E.}(\overline{y}_A - \overline{y}_P)}\right].$$

Suppose we want to estimate a $(1-2\alpha)100\%$ C.I. for μ_{A-P} . Let $F_B^{*-1}(\alpha)$ be the value below which lie the smallest $100\alpha\%$ of the bootstrap sampling distribution of t_2^* and let $F_B^{*-1}(1-\alpha)$ be the value above which lie the largest $100\alpha\%$ of the bootstrap sampling distribution of t_2^*, the α and $(1-\alpha)$ quantiles of that sampling distribution. Then, the lower limit to the estimated C.I. is given by

$$(\overline{y}_A - \overline{y}_P) - \left[\widehat{S.E.}(\overline{y}_A - \overline{y}_P) \times F_B^{*-1}(1-\alpha)\right]$$

and the upper limit is given by

$$(\overline{y}_A - \overline{y}_P) - \left[\widehat{S.E.}(\overline{y}_A - \overline{y}_P) \times F_B^{*-1}(\alpha)\right],$$

where $\widehat{S.E.}(\overline{y}_A - \overline{y}_P)$ is the bootstrap estimate of S.E. described earlier.

Example

A collection of reusable S-Plus (MathSoft, 1999) functions to carry out the required computations are available from the author. These functions were used in the analysis described here. The data are taken from a secondary source (Hand, et al., 1994).

Forty laboratory mice are randomized into four treatment groups, ten to each. One group is designated a placebo group. Animals in the other three groups are administered either a low, medium, or high dose of an investigative substance. Following administration of placebo or investigative substance, response time in milliseconds to an electric stimulus delivered to the tail is measured. The reaction times are these:

```
 [1] 2.4 3.0 3.0 2.2 2.2 2.2 2.2 2.8 2.0 3.0
[11] 2.8 2.2 3.8 9.4 8.4 3.0 3.2 4.4 3.2 7.4
[21] 9.8 3.2 5.8 7.8 2.6 2.2 6.2 9.4 7.8 3.4
[31] 7.0 9.8 9.4 8.8 8.8 3.4 9.0 8.4 2.4 7.8
```

Response times for the 10 placebo animals are reported in the first line, followed by those for the low, medium, and high dose animals.

For purposes of illustration, I find S.E. and C.I. estimates for each of six treatment comparisons, placebo vs. low, placebo vs. medium, placebo vs. high, low vs. medium, low vs. high, and medium vs. high. The sample values of these comparisons, differences in sample mean response times, are listed here.

```
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 2.28 3.32 4.98 1.04  2.7 1.66
```

All treatment comparisons reflect increased response times for increased dosages.

Estimated S.E.s for the six sample comparisons and estimated 90% bootstrap-*t* C.I. limits for the corresponding local population parameters were obtained, first under the assumption of maximum between treatment correlations:

```
               se         lb           ub
[1,] 0.9475473  0.1298728 2.28 4.268236
[2,] 0.9445015  1.1932715 3.32 5.125745
[3,] 0.9400225  2.8534152 4.98 6.812919
[4,] 0.9861751 -1.0333153 1.04 3.119057
[5,] 1.0119251  0.5454304 2.70 4.765690
[6,] 0.9802569 -0.4535278 1.66 3.625892
```

S.E. estimates are given in the first column, lower and upper bound parameter estimates in columns two and four, and the point estimate—the sample mean difference—in column three.

These are the corresponding estimates based on the assumption of minimum between-treatment correlations:

```
             se         lb            ub
[1,] 0.8069315  0.7943806 2.28 3.651933
[2,] 0.8360426  1.7413901 3.32 4.808211
[3,] 0.8121788  3.6136318 4.98 6.369008
[4,] 0.8385232 -0.3527204 1.04 2.533939
[5,] 0.8352721  1.2869342 2.70 4.120851
[6,] 0.8348318  0.1920332 1.66 3.122305
```

This second set of S.E. estimates is smaller and the C.I. widths are shorter than they were for maximum treatment correlations. Finally, these are the interpolated S.E.s and C.I. limits, assuming each of the six between-treatment correlations takes the value $\rho = 0.80$.

```
        se.int     lb.int        ub.int
[1,] 0.9240941  0.2322750 2.28 4.173308
[2,] 0.9206761  1.3061427 3.32 5.057823
[3,] 0.9196315  2.9618685 4.98 6.745638
[4,] 0.9632128 -0.9365398 1.04 3.033774
[5,] 0.9806463  0.6660628 2.70 4.658995
[6,] 0.9556651 -0.3524599 1.66 3.545035
```

These final results are purely illustrative. I have no basis for the assumption of this particular correlation between treatments. The value, $\rho = 0.80$, however, may be reasonably close to the proportion of the observed between animals response time variance that is "true" inter-animal variability. Or, roughly 20% of the observed variability may be error variability.

References

Anastasi, A. (1982). *Psychological testing*. London: Collier MacMillan.

Booth, J. G., Butler, R. W., & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, *89*, 1282-1289.

Chernick, M. R. (2000). *Bootstrap methods: a practitioner's guide*. New York: Wiley.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge: Cambridge University Press.

Efron, B. & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

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

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. (2000). *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Brooks-Cole.

Lunneborg, C. E. (in press). Random assignment of available cases: Bootstrap nonparametric standard errors and confidence intervals. *Psychological Methods*.

MathSoft (1999). *S-Plus user's guide.* Seattle: Data Analysis Products Division, MathSoft, Inc.

Neyman, J. (1923, English translation, 1990). On the application of probability theory to agricultural experiments: essay on principles. Section 9. *Statistical Science*, *5*, 465-472.

Reichardt, C. S., & Gollob, H. F. (1999). Justifying the use and increasing the power of a *t* test for a randomized experiment with a convenience sample. *Psychological Methods*, *4*, 117-128.

Rubin, D. R. (1990a). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472-480.

Rubin, D. R. (1990b). Formal models of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*, 279-292.

Rubin, D. R. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, *47*, 1213-1234.