Monte Carlo Evaluation of Resampling-Based Hypothesis Tests

Dennis D. BOOS and Ji ZHANG

Monte Carlo estimation of the power of tests that require resampling can be very computationally intensive. It is possible to reduce the size of the inner resampling loop as long as the resulting estimator of power can be corrected for bias. A simple linear extrapolation method is shown to perform well in correcting for bias and thus reduces computation time in Monte Carlo power studies.

KEY WORDS: Bootstrap; Extrapolation; Monte Carlo test; Permutation test; p value; Power function; SIMEX.

1. INTRODUCTION

The explosion of computing capabilities has greatly facilitated the use of both classical permutation tests and modern bootstrap methods. Statistical evaluation of these methods, however, often requires another level of computing (an outer Monte Carlo loop) that can still strain the fastest computers. In this article we introduce methods that can reduce the total computing time when estimating the power of resampling-based tests via Monte Carlo sampling. The basic tool that we use is an extrapolation method similar to SIMEX, the bias reduction method introduced by Cook and Stefanski (1994) for measurement error problems. We also give recommendations for allocation of computing effort when extrapolations are not used. These latter recommendations are essentially an empirical update of work of Oden (1991).

Suppose that we want to analyze a test procedure that produces a p value p and then rejects the null hypothesis at level α if $p \leq \alpha$. Monte Carlo estimation of the power function at a particular alternative would proceed simply by generating many independent datasets and computing the proportion of rejections. At each alternative, this Monte Carlo estimate will be unbiased for the true power function, and one simply chooses a large enough Monte Carlo sample size to obtain the desired accuracy.

In the situations that we have in mind, however, the p value for each dataset is computationally difficult, and an estimated p value $\hat{p}$ is used in place of p (at least within the Monte Carlo loop). The Monte Carlo estimate of the power function based on the proportion of times $\hat{p} \leq \alpha$ will be biased for estimating the power function of the test procedure defined by $p \leq \alpha$ (except possibly at the null hypothesis). This bias results from the fact that although typically $E\hat{p} = p$ for the inner loop sampling,

$$\begin{aligned} &E(\text{power estimate}) \\ &\quad = P(\hat{p} \leq \alpha) = EE\delta(\hat{p} \leq \alpha) \neq E\delta(E\hat{p} \leq \alpha) \\ &\quad = E\delta(p \leq \alpha) = \text{ true power,} \end{aligned}$$

because of the nonlinearity of the function $\delta(\cdot)$, where $\delta(A) = 1$ if A is true and $= 0$ otherwise. The connection to measurement error methods arises because we use a sample of $\hat{p}$'s (p's measured with error) in the estimation procedure.

If one replaces the original test procedure $p \leq \alpha$ by $\hat{p} \leq \alpha$, then this modified test procedure is called a Monte Carlo test (introduced first in Barnard 1963), and the foregoing Monte Carlo estimate of the power function for this modified procedure will be unbiased if the number of resamples used in the actual procedure is the same as in the Monte Carlo study to analyze the procedure. Hope (1968), Jockel (1986), and Hall and Titterington (1989), among others, have analyzed Monte Carlo tests. An alternative approach using a sequential approximation to the full permutation test was given by Lock (1991).

In this article we focus on estimation of the power function for the original test based on p. For permutation tests, this means that we are referring to the power function of the test based on the full set of permutations M, and for parametric bootstrap tests, we mean the test based on an infinite number of bootstrap samples. Our thinking is that for a particular dataset, one typically can make $\hat{p}$ as close to p as desired by taking a large number of resamples, in effect using the true p. For Monte Carlo analysis of this procedure, however, the resample size becomes an issue, because the number of test statistic evaluations is then the resample size times the Monte Carlo sample size.

Example. For illustration, let us consider the simple two-sample location shift problem where we have available iid samples $X_1, \ldots, X_m$ from $F(x)$ and $Y_1, \ldots, Y_n$ from $G(x) = F(x - \Delta)$. The null hypothesis is $H_0 : \Delta = 0$, and for simplicity we consider the one-sided alternative $H_a : \Delta > 0$. Suppose that we choose T to be the usual two-sample t statistic

$$T = \frac{\bar{Y} - \bar{X}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) s_p^2}},$$

Dennis D. Boos is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *boos@stat.ncsu.edu*). Ji Zhang is Director, Clinical Biostatistics, Merck Research Laboratories, Rahway, NJ 07065 (E-mail: *zhangj@merck.com*). The authors thank Len Stefanski, Lauren McIntyre, and the Statistical Genetics group at North Carolina State University for helpful discussions over a number of years. Special thanks to Andrew Gelman for noting that the estimators $\widehat{\text{pow}}_{I_j}, j > 1$, could be computed simply.

Journal of the American Statistical Association
June 2000, Vol. 95, No. 450, Theory and Methods

where s_p^2 is the pooled sample variance. We consider the following three methods of constructing rejection regions and p values, where T_0 is the value of T for our given sample:

1. Standard parametric approach. If we assume that F and G are normal distributions with the same variance, then we could compute the p value $p = 1 - t_{m+n-2}(\cdot)$, where t_{m+n-2} is a central t-distribution function with $m+n-2$ degrees of freedom, and reject H_0 if $p \le \alpha$.

2. Parametric bootstrap. Suppose now that we do not have available $t_{m+n-2}(\cdot)$ or even know the distribution of T under H_0 and normality. We could then generate I sets of iid samples of size m and n from the standard normal distribution, compute T for each set (say $T_1^*, \dots, T_I^*$), and compute an estimated p value

$$\hat{p} = \frac{1}{I}\sum_{i=1}^{I} \delta(T_i^* \ge T_0).$$

As $I \to \infty, \hat{p}$ approximates the p value in method 1 obtained from the t distribution. The related test procedure may be defined as reject H_0 if $\hat{p} \le \alpha$.

3. Permutation (distribution-free) approach. A permutation p value can be obtained by computing T for all possible $M = \binom{m+n}{m}$ partitions (permutations) of the pooled data $\{X_1, \dots, X_m, Y_1, \dots, Y_n\}$ into two samples and counting the proportion of these T that are greater than or equal to T_0 calculated from the original data. Because M is usually a very large number, standard practice is to draw only a simple random sample of size I of the permutations with replacement and compute T for each permutation, $T_1^*, \dots, T_I^*$, and obtain $\hat{p}$ as in method 2.

The ideas in this article are useful for estimating the power of methods 2 or 3 when very large I is to be used in practice. The Monte Carlo power estimate would be based on simulations with much smaller I values.

Following Oden's (1991) notation, we have an outer Monte Carlo loop of size O where datasets are generated under the null hypothesis or under some alternative. For each of these datasets, there will also be an inner resampling loop of size I (often called B in the bootstrap literature) where further datasets are generated to find the p value (or rejection region). The goal is then to estimate the power function of the test procedure in an optimal way under the restriction that the total number of datasets OI generated is fixed.

Let $H_\Delta(\alpha) = P(p \le \alpha)$ be the true power function of the test procedure at an alternative Δ. For each Monte Carlo sample, the estimated p value $\hat{p}$ based on I resamples (using sampling with replacement) will be such that $I\hat{p}$ has a binomial (I, p) distribution conditional on the Monte Carlo sample (and thus conditional on the p value p). Unconditionally, the estimated Monte Carlo power function

$$\hat{H}_{\Delta,O,I}(\alpha) = \frac{1}{O}\sum_{i}^{O} \delta(\hat{p}_i \le \alpha) \tag{1}$$

has mean

$$\begin{aligned} H_{\Delta,I}^*(\alpha) &\equiv E[\hat{H}_{\Delta,O,I}(\alpha)] \\ &= \sum_{k=0}^{[\alpha I]} \binom{I}{k} \int_0^1 t^k (1-t)^{I-k}\, dH_\Delta(t), \end{aligned} \tag{2}$$

where $[\alpha I]$ is the greatest integer part of αI. If H_Δ is a beta(a,b) distribution function, then (2) becomes (Oden 1991, eq. 4)

$$H_{\Delta,I}^*(\alpha) = \frac{1}{B(a,b)} \sum_{k=0}^{[\alpha I]} \binom{I}{k} B(k+a, I-k+b), \tag{3}$$

where $B(a,b)$ is the beta function with parameters a and b. Equation (3) is just the distribution function of a beta-binomial random variable evaluated at $[\alpha I]$.

To illustrate the damping effect of using $\hat{p}$ rather than p, Figure 1 plots (2) when $H_\Delta(t)$ is a beta(1, 25) distribution. In a real situation, $H_\Delta(t)$ would not be known, but the points in Figure 1 corresponding to finite I values could be estimated. The true power for this example is the point at $1/I = 0, .72$, but using $I = 59$ in a Monte Carlo experiment would cause the estimated power to be centered at .66 instead. However, the regular pattern of the points suggests that a curve could be fit and extrapolated back to remove the bias. For example, fitting the points (1/59, .66), (1/39, .63), and (1/19, .57) by least squares to a straight line yields $\hat{H}_{\Delta,I}(\alpha) = .70 - 2.50(1/I)$ and results in a bias-reduced estimate at $1/\infty = 0$ of .70. The basic approach of this article is then to estimate $H_{\Delta,I}^*(\alpha)$ for several values of I, fit a curve (usually just a straight line), and extrapolate back to 0, which corresponds to $I = \infty$. This is similar in principle to the SIMEX method of Cook and Stefanski (1994). It could also be called a generalized jackknife procedure (see Efron 1982, p. 7).

Because unconditionally $O\hat{H}_{\Delta,O,I}(\alpha)$ has a binomial $(O, H_{\Delta,I}^*(\alpha))$ distribution, the mean squared error (MSE) of $\hat{H}_{\Delta,O,I}(\alpha)$ is (Oden 1991, eq. 3)

$$\mathrm{MSE}_\Delta(\alpha) = [H_\Delta(\alpha) - H_{\Delta,I}^*(\alpha)]^2 + \frac{H_{\Delta,I}^*(\alpha)(1 - H_{\Delta,I}^*(\alpha))}{O}.$$

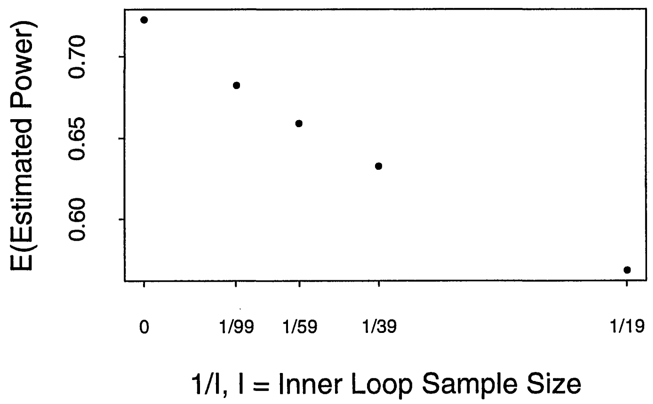

Figure 1. E(Estimated Power) at $\alpha = .05$: $H_{\Delta,I}^*(.05)$ for $I = \infty$, 99, 59, 39, 19 Versus 1/I When $H_\Delta(t)$ is a Beta(1, 25) Distribution.

Oden (1991) considered minimizing the maximum of this MSE over α, and obtained $I = 2\sqrt{O}$ under the null hypothesis and suggested the range $I = 2\sqrt{O}$ to $I = 4\sqrt{O}$ for general usage. In Section 5 we give empirical results suggesting that $I = 8\sqrt{O}$ would appear to be a better rule of thumb when using no correction for bias.

We introduce our basic extrapolation method in Section 2, and provide a theorem on the asymptotic order of the MSE in Section 3. In Section 4 we give a Monte Carlo analysis of the method in the context of estimating the power of the two-sample parametric t test mentioned earlier. In Section 5 we illustrate how to use the method in practice, this time for the permutation t test. We give recommendations somewhat different from Oden (1991) for the standard approach without extrapolation in Section 6, and a brief summary in Section 7.

2. THE EXTRAPOLATION METHOD

To estimate the power of an α-level test based on a statistic T under some alternative, we envision generating O datasets under the alternative, computing $\hat{p}_{I,1}, \ldots, \hat{p}_{I,O}$ by resampling I times for each of the datasets, and then defining

$$\widehat{\text{pow}}_I = \frac{1}{O} \sum_{j=1}^{O} \delta(\hat{p}_{I,j} \le \alpha).$$

For example, we recommend $I = 59$ or $I = 99$ when using $\alpha = .05$.

The basic extrapolation method is to use the same simulated data to get power estimates for a number of values of I, say $\widehat{\text{pow}}_{I_1}, \ldots, \widehat{\text{pow}}_{I_k}$, where $I_1 > I_2 > \cdots > I_k$, fit a curve to the pairs $(\widehat{\text{pow}}_{I_1}, 1/I_1), \ldots, (\widehat{\text{pow}}_{I_k}, 1/I_k)$, and extrapolate back to estimate the power at $I = \infty$; that is, at $1/I = 0$.

Although the method is straightforward, there are a number of questions to answer:

1. What are suitable values of $I_1, \ldots, I_k$?
2. Why should one fit a curve to $1/I$ values?
3. What kind of curves should be fit?
4. How can we obtain $\widehat{\text{pow}}_{I_j}, j = 2, \ldots, k$, if the simulation is carried out only for $I = I_1$?

Suitable Choice of ***I***. For simplicity, consider a parametric bootstrap situation with no nuisance parameters needed by the data generation process and a continuous statistic T for which large values are evidence that the alternative hypothesis is true. The two-sample t situation given in Section 1 is an example. (Although the t statistic has a nuisance parameter σ estimated by s_p, the generation of the normal datasets does not require that σ be estimated.) For a given sample, we compute the statistic T and call that value T_0. If we generate I independent datasets under the null hypothesis and compute the test statistic for each dataset, then the resulting $T_0, T_1^*, \ldots, T_I^*$ is an iid sample from the distribution of the test statistic. The estimated p value $\hat{p}_I = \{\#$ of $T_i^* \ge T_0\}/I$ has a discrete uniform distribution due to exchangeability of the sample:

$$P(\hat{p}_I = 0) = P(\hat{p}_I = 1/I) = \cdots = P(\hat{p}_I = 1) = 1/(I+1).$$

The resulting test defined by the rejection region $\hat{p} \le \alpha$ has exact level α if $(I+1)\alpha$ is an integer. For example, if $\alpha = .05$ and $I = 99$, then we reject the null hypothesis if $\hat{p}$ is either 0, 1/99, 2/99, 3/99, or $4/99 = .0404$. The associated probabilities add up to exactly α: $5 \times 1/(I+1) = 5/100 = .05$. If one uses the more natural value of, say, $I = 100$, then the resulting test is liberal, with level $6/101 = .059$. Similar results are true for permutation tests [although discreteness can make the tests conservative when $(I+1)\alpha$ is an integer] and approximately true for the nonparametric bootstrap (see Hall and Titterington 1989). Some people prefer to define $\hat{p}$ by $(\{\#$ of $T_i^* \ge T_0\} + 1)/(I+1)$, but results using the rule $\hat{p} \le \alpha$ are similar.

If $\alpha = .05$ and $I_1 = 99$, then the only smaller values so that $(I+1)\alpha$ is an integer are $I_2 = 79, I_3 = 59, I_4 = 39$, and $I_5 = 19$. These would be one natural set to use in the extrapolation process. The smallest choice of I_1 would be $I_1 = 39$, but this results in only two pairs, $(\widehat{\text{pow}}_{39}, 1/39)$ and $\widehat{\text{pow}}_{19}, 1/19)$, available for the extrapolation process. Of course, if $\alpha = .01$, then $(I_1, I_2) = (199, 99)$ would be the smallest pair.

Suitable Curves. Hope (1968) and Jockel (1986) gave conditions under which Monte Carlo tests have power functions that are monotone in I and thus in $1/I$. Moreover, plots like Figure 1 are very suggestive of a regression of $\widehat{\text{pow}}_I$ on $1/I$. But a nice technical justification is based on work of Hald (1968), who gave an expansion for (2):

$$\begin{aligned} P_\Delta(\hat{p}_I \le \alpha) &= H^*_{\Delta,I}(\alpha) \\ &= H_\Delta(\alpha) + \frac{B_1(\alpha)}{I} + \frac{B_2(\alpha)}{I^2} + O(I^{-3}), \quad (4) \end{aligned}$$

where B_1 and B_2 are functions depending on H_Δ. In general, of course, we do not know B_1 and B_2, but we can estimate them by regressing estimated power on $1/I$ and $1/I^2$. Actually, our goal is to estimate $H_\Delta(\alpha)$. Thus we can fit either a linear regression or a quadratic regression of $\widehat{\text{pow}}_I$ on $1/I$ and use the estimated intercept as our adjusted power estimate. In the simulation study of Section 4, we find that the least squares quadratic regression gives a less biased power estimate, as (4) would suggest, but the variability of the estimated intercept for the linear regression is typically much smaller than that for the quadratic regression. In general, our recommendation is to use the linear estimate unless O values are very large.

To be very specific, if $(I_1, I_2, I_3) = (59, 39, 19)$, then the adjusted power estimate given by the ordinary least squares estimate of the intercept is

$$\begin{aligned} \widehat{\text{pow}}_{\text{lin}} = 1.01137(\widehat{\text{pow}}_{59}) &+ .61294(\widehat{\text{pow}}_{39}) \\ &- .62430(\widehat{\text{pow}}_{19}). \quad (5) \end{aligned}$$

If $(I_1, I_2, I_3, I_4, I_5) = (99, 79, 59, 39, 19)$, then the estimate is

$$\begin{aligned} \widehat{\text{pow}}_{\text{lin}} = .46688(\widehat{\text{pow}}_{99}) &+ .41631(\widehat{\text{pow}}_{79}) + .33145(\widehat{\text{pow}}_{59}) \\ &+ .15956(\widehat{\text{pow}}_{39}) - .37420(\widehat{\text{pow}}_{19}). \quad (6) \end{aligned}$$

Here $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ is just $\bar{y} - \hat{b}\bar{x}$, where the y's are $\widehat{\mathrm{pow}}_I$, the x's are $1/I$, and $\hat{b}$ is the simple linear slope estimate $\hat{b} = \sum(x_i - \bar{x})y_i / \sum(x_i - \bar{x})^2$.

Equation (3) suggests another curve to fit to the raw power estimates $\widehat{\mathrm{pow}}_I$. Suppose that for fixed $\Delta, H_\Delta(\alpha)$ can be well approximated by a beta(a, b) distribution function. Then we can just use nonlinear least squares to fit (3) as a function of (a, b). The adjusted power estimate is then given by the probability that a beta$(\hat{a}, \hat{b})$ random variable is less than or equal to α. For the simulations in Section 4, we find that this method produces good results, but not quite as good as those from the linear extrapolant.

To show that a beta(a, b) distribution function assumption for $H_\Delta(\alpha)$ makes sense, consider X that is distributed as normal(μ, σ^2), where σ^2 is known. The "Z" test that rejects $H_0 : \mu = \mu_0$ for large X has power $H_\Delta(\alpha) = 1 - \Phi(-\Delta + \Phi^{-1}(1-\alpha))$, where $\Delta = (\mu_0 - \mu)/\sigma$ for alternative μ. Using nonlinear least squares, we found that the following fits were excellent for the "Z" test with $\alpha = .05$: $(\Delta = .1, a = .94, b = 1.06), (\Delta = .5, a = .75, b = 1.33), (\Delta = 1.0, a = .56, b = 1.80), (\Delta = 2.0, a = .30, b = 3.61)$, and $(\Delta = 3.0, a = .13, b = 8.19)$.

Estimation of $\widehat{\mathrm{pow}}_{I_j}, \boldsymbol{j = 2, \ldots, k}$. Consider the data that result from generating O datasets; computing T for each of those, denoted by $T_{0k}, k = 1, \ldots, O$; and then for each outer loop situation, resampling I_1 times to get $\hat{p}_{I_1 k}, k = 1, \ldots, O$. To be specific, for $I_1 = 59$, we have

$$
\begin{array}{llll}
1: & T_1^*, \ldots, T_{59}^* \rightarrow & \delta(T_1^* \geq T_{01}) & \\
 & & \delta(T_2^* \geq T_{01}) & \\
 & & \cdot & \rightarrow \hat{p}_{I_1 1} \rightarrow u_{59,1} = \delta(\hat{p}_{I_1 1} \leq \alpha) \\
 & & \cdot & \\
 & & \delta(T_{59}^* \geq T_{01}) & \\
 & & & \\
2: & T_1^*, \ldots, T_{59}^* \rightarrow & \delta(T_1^* \geq T_{02}) & \\
 & & \delta(T_2^* \geq T_{02}) & \\
 & & \cdot & \rightarrow \hat{p}_{I_1 2} \rightarrow u_{59,2} = \delta(\hat{p}_{I_1 2} \leq \alpha) \\
 & & \cdot & \\
 & & \delta(T_{59}^* \geq T_{02}) & \\
\cdot & & \cdot & \\
\cdot & & \cdot & \\
O: & T_1^*, \ldots, T_{59}^* \rightarrow & & \rightarrow \hat{p}_{I_1 O} \rightarrow u_{59,O} = \delta(\hat{p}_{I_1 O} \leq \alpha).
\end{array}
$$

Notice that for each set of resamples, we get $I_1 = 59$ 0's and 1's, which we average to get the estimate $\hat{p}_{I_1 k}$ and then average the $u_{59,k} = \delta(\hat{p}_{I_1 k} \leq \alpha)$ to get $\widehat{\mathrm{pow}}_{59}$. To get a similar estimate for $I_2 = 39$, we just need to average over $I_2 = 39$ instead of 59 of the 0's and 1's within each resampling outcome. Because there are $\binom{59}{39}$ different subsets, the natural approach is take an average over those subsets after converting each subset to an estimated p value:

$$
u_{39,k} = \frac{1}{\binom{59}{39}} \sum \cdots \sum \delta\left(\frac{1}{39} \sum_{i=1}^{39} \delta(T_{j_i}^* \geq T_{0k}) \leq \alpha\right). \tag{7}
$$

The notation in (7) is not perfect, but $u_{39,k}$ is a U statistic with a kernel of degree $I_2 = 39$, and the outer sum is over all distinct subsets. We average the $u_{39,k}$ values to get $\widehat{\mathrm{pow}}_{39}$. The calculation in (7) looks formidable until one thinks in terms of hypergeometric probabilities. Consider an urn with $N = 59$ total 0's and 1's, of which K are 1's and $N - K$ are 0's. If we randomly sample $n = 39$ 0's and 1's from this urn and let S be the sum, then

$$
u_{39,k} = E\delta\left(\frac{S}{39} \leq \alpha\right) = P(S \leq 39\alpha).
$$

Thus to get $u_{39,k}$, we need only get the probability that a hypergeometric $(N = 59, K, n = 39)$ random variable is less than or equal to 39α. In the Monte Carlo simulation, one can tabulate these hypergeometric probabilities for quick retrieval.

Because $\widehat{\mathrm{pow}}_{59}, \widehat{\mathrm{pow}}_{39}$, and $\widehat{\mathrm{pow}}_{19}$ are all based on the same sets of 0's and 1's, these estimates are highly correlated. One can estimate the correlations from the $O \times 3$ matrix with rows $(u_{59,k}, u_{39,k}, u_{19,k})$ and use these to get estimated generalized least squares (EGLS) estimates of the intercept and to get proper variance estimates for the intercept estimate. In the simulation study of Section 4, however, we found that EGLS is not better than ordinary least squares and that the variance of the adjusted power estimates are only slightly larger than the simple $\widehat{\mathrm{pow}}_{I_1}(1 - \widehat{\mathrm{pow}}_{I_1})/O$ variance estimate for raw power estimates. Thus it hardly seems worthwhile to estimate the correlations.

3. THEORETICAL CALCULATIONS

Our approach is to estimate the intercept using simple linear or quadratic regression motivated by (4) and using the (y, x) pairs $(\widehat{\mathrm{pow}}_{I_1}, 1/I_1), \ldots, (\widehat{\mathrm{pow}}_{I_k}, 1/I_k)$, where $I_1 > I_2 > \cdots > I_k$. As explained in the previous section, $\widehat{\mathrm{pow}}_{I_1}$ is the usual power estimate based on I_1 inner loop resamples and on O outer loop datasets. Thus $\mathrm{var}(\widehat{\mathrm{pow}}_{I_1}) = H^*_{\Delta,I_1}(\alpha)(1 - H^*_{\Delta,I_1}(\alpha))/O$. The other estimates $\widehat{\mathrm{pow}}_{I_j}, j = 2, \ldots, k$, have variances bounded by $H^*_{\Delta,I_j}(\alpha)(1 - H^*_{\Delta,I_j}(\alpha))/O$, due to the fact that they are U statistics. We envision finite k and $I_2, \ldots I_k$ close to I_1, so that each converges to ∞ at the same rate as $I_1 \rightarrow \infty$. In particular, for $\alpha = .05$, we are thinking of the set $(I_1, I_2 = I_1 - 20, \ldots, I_k = I_1 - 20(k-1))$.

In this setup, we now give a theorem on the asymptotic MSE of our adjusted power estimate $\widehat{\mathrm{pow}}_{\mathrm{lin}}$, which is just the estimate of the intercept, as both I_1 and $O \rightarrow \infty$.

Theorem 1. Suppose that (4) holds and that

$$
H^*_{\Delta,I}(\alpha) = H_\Delta(\alpha) + \frac{B_1(\alpha)}{I} + \frac{C_I(\alpha)}{I^2},
$$

where $|C_I| \leq C_0$ for all integer values of $(I+1)\alpha$ and C_0 is a positive constant. If I_k is the same order as I_1, as I_1 and $O \rightarrow \infty$, then the MSE of $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ for estimating the true power $H_\Delta(\alpha)$ satisfies

$$
E(\widehat{\mathrm{pow}}_{\mathrm{lin}} - H_\Delta(\alpha))^2 = \mathbf{O}(I_1^{-4}) + \mathbf{O}(O^{-1}).
$$

Proof. Direct calculation shows that the bias of $\widehat{\mathrm{pow}}_{\mathrm{lin}} = \bar{y} - \hat{b}\bar{x}$ is $\mathbf{O}(I_1^{-2})$, using the fact that $\bar{x} = \mathbf{O}(I_1^{-1})$ and $E(\hat{b} - B_1(\alpha)) = \mathbf{O}(I_1^{-1})$. The variance calculation is similarly direct, due to the linear nature of $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ and the bounds on $\mathrm{var}(\widehat{\mathrm{pow}}_{I_j}), j = 1, \ldots, k$.

4. SIMULATION RESULTS FOR THE TWO-SAMPLE t STATISTIC

In this section we report on a small simulation study of Monte Carlo power estimates for the two-sample t statistic using the parametric bootstrap for $n_1 = 8$ and $n_2 = 4$ with normal data and equal standard deviation σ. The reason to consider such a simple situation is that we know the true power exactly, and computing time for each replication is small. In fact, the alternatives used are standardized mean difference $\Delta/\sigma = .5, 1.0, 1.5$, and 2.0 with true powers .189, .451, .737, and .918.

For this study, we need three simulation loops. The outside loop is of size 100 replications; the two inner loops are actually what we previously have called the outer and inner loops. Thus this study is actually an average of 100 separate Monte Carlo estimates of power. We look at two (O, I) combinations, $(O = 1{,}000, I = 59)$ and $(O = 596, I = 99)$, that each have about 59,000 computations.

Table 1 reports estimates of the root mean squared error (RMSE) and bias $\times 1{,}000$ of the various power estimates. The standard errors of the estimates are in the range of .001 to .002 for RMSE and around 2 for the bias $\times 1{,}000$.

The first and seventh rows ($\widehat{\mathrm{pow}}_{\infty}$) of Table 1 give results for power estimates based on the true known t percentiles appropriate for normal data. They are labeled $\widehat{\mathrm{pow}}_{\infty}$ to reflect the fact that resampling with I approaching ∞ would give this result. These of course are unbiased (the nonzero bias results in Table 1 just reflect Monte Carlo variation), and here the RMSE could have been calculated simply by $\sqrt{\mathrm{power}(1 - \mathrm{power})/O}$. For a given $O, \widehat{\mathrm{pow}}_{\infty}$ represents the best power estimates possible.

The second ($\widehat{\mathrm{pow}}_{59}$) and eighth rows ($\widehat{\mathrm{pow}}_{99}$) are the raw power estimates based on $I = 59$ and $I = 99$. For these raw estimates, the $(O = 596, I = 99)$ situation is more efficient in terms of RMSE than $(O = 1{,}000, I = 59)$ for all but $\Delta = .5$, because the bias is a large factor except at $\Delta = .5$.

Table 1. RMSE and Bias of Power Estimates for Two-Sample t Statistic with Parametric Bootstrap Critical Values

	RMSE				*bias × 1,000*			
$\Delta =$	*.5*	*1.0*	*1.5*	*2.0*	*.5*	*1.0*	*1.5*	*2.0*
O = 1,000, *I* = 59								
$\widehat{\mathrm{pow}}_{\infty}$	.012	.017	.015	.010	1.4	.3	−1.6	.1
$\widehat{\mathrm{pow}}_{59}$	.013	.031	.041	.030	−6.0	−25.7	−38.3	−28.4
$\widehat{\mathrm{pow}}_{\mathrm{lin}}$	.013	.020	.017	.011	−.4	−6.4	−6.0	3.4
$\widehat{\mathrm{pow}}_{\mathrm{gls}}$	.013	.020	.018	.010	−1.2	−8.4	−7.0	3.9
$\widehat{\mathrm{pow}}_{\mathrm{quad}}$	.018	.025	.025	.019	.7	−.9	−1.8	1.7
$\widehat{\mathrm{pow}}_{\mathrm{beta}}$	*	.022	.018	.012	*	2.4	2.1	6.2
O = 596, *I* = 99								
$\widehat{\mathrm{pow}}_{\infty}$	.016	.019	.018	.010	3.8	.2	−.6	.8
$\widehat{\mathrm{pow}}_{99}$	.018	.025	.030	.019	−2.2	−15.8	−23.9	−15.3
$\widehat{\mathrm{pow}}_{\mathrm{lin}}$	.018	.021	.019	.013	1.1	−4.8	−5.1	4.4
$\widehat{\mathrm{pow}}_{\mathrm{quad}}$	.021	.023	.022	.015	1.9	−.9	−2.6	2.2
$\widehat{\mathrm{pow}}_{\mathrm{beta}}$	*	.023	.020	.013	*	1.1	0	6.2

* Convergence problems with the nonlinear least squares routine.

NOTE: Results based on 100 replications. Standard errors for RMSE $\approx$ .001 — .002. Standard errors for bias $\times$ 1,000 $\approx$ 2.

The other estimators in Table 1 are bias-adjusted estimators using $(I_1, \ldots, I_k) = (59, 39, 19)$ in the $(O = 1{,}000, I = 59)$ case and (99, 79, 59, 39, 19) in the $(O = 596, I = 99)$ case:

1. $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ is the simple linear extrapolation method using (5) for the $(O = 1{,}000, I = 59)$ case and (6) for the $(O = 596, I = 99)$ case. In either case, $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ is just the intercept estimator from ordinary least squares simple linear regression.
2. $\widehat{\mathrm{pow}}_{\mathrm{gls}}$ is similar to $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ except that EGLS is used, based on empirical covariance estimates for ($\widehat{\mathrm{pow}}_{59}$, $\widehat{\mathrm{pow}}_{39}$, $\widehat{\mathrm{pow}}_{19}$) and ($\widehat{\mathrm{pow}}_{99}$, $\widehat{\mathrm{pow}}_{79}$, $\widehat{\mathrm{pow}}_{59}$, $\widehat{\mathrm{pow}}_{39}$, $\widehat{\mathrm{pow}}_{19}$).
3. $\widehat{\mathrm{pow}}_{\mathrm{quad}}$ is the quadratic extrapolation estimator; that is, the ordinary least squares intercept estimator from a simple quadratic model.
4. $\widehat{\mathrm{pow}}_{\mathrm{beta}}$ is the estimator based on fitting a beta(a, b) distribution to (3) by nonlinear least squares. The estimator is then $P(B \leq \alpha)$, where B has a beta$(\hat{a}, \hat{b})$ distribution.

From Table 1 we see that the linear extrapolation estimators, $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ and $\widehat{\mathrm{pow}}_{\mathrm{gls}}$, perform the best and very similarly. Their similarity is likely due to the fact (not displayed) that the estimated covariance matrix of the $\widehat{\mathrm{pow}}_{I}$ used as dependent variables in the regressions has nearly equal diagonal elements and nearly equal off-diagonal elements. The estimators $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ and $\widehat{\mathrm{pow}}_{\mathrm{gls}}$ would be identical if the estimated covariance matrix had equal diagonal elements and equal off-diagonal elements. Estimating the covariance matrix is also useful for estimating the variance of either estimator, but because their RMSE values are close to $\sqrt{\mathrm{power}(1 - \mathrm{power})/O}$, this latter formula will often be adequate.

The quadratic extrapolator is less biased than the linear estimators, but the additional variance due to fitting the $1/I^2$ term is too costly and keeps the RMSE of $\widehat{\mathrm{pow}}_{\mathrm{quad}}$ relatively high. Of course, at large enough O values, the smaller bias will make the quadratic estimator preferable to the linear ones. For $I = 59$, we estimate that at $O = 8{,}900$, the linear and quadratic estimators would have the same MSE. For $I = 99$, we estimate that at $O = 3{,}700$ they would have the same MSE. It is interesting that the quadratic estimator is more efficient at the $(O = 596, I = 99)$ combination than at the $(O = 1{,}000, I = 59)$ combination, even though their biases are similar. Apparently it is much more stable to fit the quadratic with five points rather than with three, even with dependent observations.

The fitted beta distribution estimator, $\widehat{\mathrm{pow}}_{\mathrm{beta}}$, performed almost as well as the linear estimators except for the $\Delta = .5$ alternative, where the nonlinear least squares program had trouble converging.

Having computed the raw estimator $\widehat{\mathrm{pow}}_{59}$, should one go to the extra effort of computing $\widehat{\mathrm{pow}}_{39}$ and $\widehat{\mathrm{pow}}_{19}$ (using the hypergeometric formulas) and plugging in (5) to get $\widehat{\mathrm{pow}}_{\mathrm{lin}}$? At $\Delta = .5$ and in general near the null hypothesis, the bias of the raw estimator $\widehat{\mathrm{pow}}_{59}$ is small and $\widehat{\mathrm{pow}}_{\mathrm{lin}}$ has little advantage over $\widehat{\mathrm{pow}}_{59}$. However, at $\Delta = 1.5$ and

$\Delta = 2.0$, the MSE of $\widehat{\text{pow}}_{\text{lin}}$ is much smaller than the MSE of the raw estimator. Using the combinations of O and I that are approximately optimal for the raw estimator $[(O, I) = (1{,}100, 219)$ for $\Delta = 1.5$ and $(O, I) = (800, 279)$ for $\Delta = 2.0]$, we find that it would take about four times as large an OI value for the raw estimator to have MSEs equal to $\widehat{\text{pow}}_{\text{lin}}$. And the savings could be even greater, because we rarely would know the optimal values for O and I in a new situation.

5. EXAMPLE: ESTIMATING THE POWER OF THE PERMUTATION *t*

To give a specific illustration, we consider the testing situation used in Section 3, but here we study the permutation t test instead of the normal theory t test. Recall that the latter has power .189, .451, .737, and .918 for standardized mean difference $\Delta/\sigma = .5, 1.0, 1.5$, and 2.0, in normal samples of size $n_1 = 8$ and $n_2 = 4$. Here we are interested in how much power we lose by using the permutation approach described in Section 1 that makes no use of the fact that the data are normally distributed.

First, we ran all four alternatives with $O = 1{,}000$ and $I = 59$ and obtained the linear estimators based on (5). Using S-PLUS, these four runs took about 2.5 hours each on a Sparcstation 4-110 MHz. The estimated standard errors using the least squares formula $(X^TX)^{-1}X^T[\hat{\Sigma}/1{,}000]X(X^TX)^{-1}$ for the intercept were very close to those from the binomial calculation $\sqrt{\text{power}(1 - \text{power})/O}$. Here $\hat{\Sigma}$ is the sample covariance matrix from the $1{,}000 \times 3$ matrix with kth row $(u_{59,k}, u_{39,k}, u_{19,k})$.

Because the standard errors ranged from .010 to .015, we decided to lower these by rerunning the four alternatives at $O = 4000$ and $I = 99$. Also, this combination of O and I is approximately where we might expect the linear estimators to have similar MSE to the quadratic estimators. These runs each took about 18 hours. The linear estimators using (6), with standard errors in parentheses, were .175 (.006), .439 (.008), .731 (.007), .921 (.005), and the quadratic estimators were .176 (.007), .444 (.009), .730 (.009), and .918 (.006). The linear and quadratic estimators are very close to one another and to the normal theory test powers of .189, .451, .737, and .918. It seems somewhat amazing that the power of the permutation t test is so close to that of the normal theory test for these small sample sizes (asymptotically they are equivalent to first order).

Figure 2 displays the raw estimates $\widehat{\text{pow}}_{99}, \widehat{\text{pow}}_{79}, \widehat{\text{pow}}_{59}, \widehat{\text{pow}}_{39}$, and $\widehat{\text{pow}}_{19}$ for all four alternatives and the linear estimators (+).

6. CHOICE OF (*O*, *I*) WITH NO EXTRAPOLATION

The thesis of this article is that extrapolation is a simple way to improve power estimates that are biased due to using small values of I in the resampling step. We realize, however, that the extrapolation requires extra computing to get the $\widehat{\text{pow}}_{I_j}$ needed to use (5) or (6). Certainly, the raw estimate $\widehat{\text{pow}}_{I_1}$ is the simplest to use and program. Thus we want to give a little guidance on (O, I) combinations that have relatively low MSE.

Oden (1991) used a minimax approach at the null hypothesis and arrived at the rule that for fixed OI values, one should choose I in the range of $I = 2\sqrt{O}$ to $I = 4\sqrt{O}$. This seems like a good place to start, and we express our results in terms of $I/\sqrt{O}$ ratios. We believe, however, that improved recommendations are possible because the range of interest for α is usually at most (.01, .10) rather than (0, 1.0) (over which Oden's maximum is taken), and minimizing the MSE under the null hypothesis may not be relevant for estimating the power at alternatives.

First, we consider the simple Z test mentioned in Section 2. There we noted that we could approximate $P(p \leq \alpha)$ by $\text{beta}(a, b)$ distribution functions. For those situations, we found the $I/\sqrt{O}$ ratios with lowest MSE for estimating the power at a range of alternative hypotheses by searching over a grid of $I/\sqrt{O}$ values. These optimal ratios are almost exactly linear in the true power, yielding the equation

$$\text{optimal}(I/\sqrt{O}) = .16 + 5.9(\text{power}).$$

Thus for power = .30, the optimal ratio $I/\sqrt{O}$ is around 2, and for power = .65, the optimal $I/\sqrt{O}$ value is around 4. These tend to confirm Oden's recommendations.

A second example is from the simulation in Section 3 for the t test. Using the values estimated there, we fit a similar optimal curve and obtained

$$\text{optimal}(I/\sqrt{O}) = .19 + 8.9(\text{power}).$$

Here we see that Oden's recommended ratios are a bit too low when power is high.

How representative are these two examples? Consider $P(p \leq \alpha)$ given by beta(2.2, 39) and beta(.1, 1.0) distribution functions. At $\alpha = .05$, the true power is .544 for beta(2.2, 39) and .549 for beta(.1, 1.0). The bias for estimating the power with small I, however, is much greater for the beta(2.2, 39) case than for the beta(.1, 1.0) case. Thus for beta(2.2, 39) at $OI = 59{,}000$, the combination (O, I) with lowest MSE is $(O, I) = (296, 199)$ and $I/\sqrt{O} = 11.6$. For beta(.1, 1.0), the optimal combination is $(O, I) = (746, 79)$

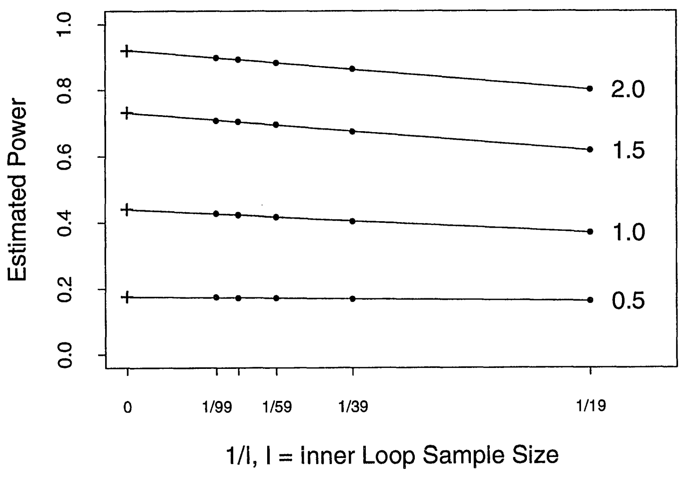

Figure 2. Estimated Power of the Permutation t Test at α = .05 by Linear Extrapolation. Data are normally distributed with standardized mean difference Δ/σ = .5, 1.0, 1.5, 2.0, n_1 = 8, n_2 = 4.

with $I/\sqrt{O} = 2.9$. This example shows the wide variety that one can obtain using two different beta distributions producing almost the same power.

Thus we decided to look at a wide variety of beta(a, b) distributions that hopefully cover the spectrum of functions $P(p \leq \alpha)$ one might find in practice. For each beta(a, b) distribution on the grid $a = .1$ to 4 by .1 with $b = 1$ to 40 by 1, we computed the MSE of (1) for $I = 19$ to 499 by 20 and for $OI = 59{,}000$. We repeated the process for $OI = 590{,}000$, and obtained the following:

1. We found the (O, I) combination with lowest MSE for each of the beta(a, b) distributions. The median $I/\sqrt{O}$ values (separately over $OI = 59{,}000$ and $OI = 590{,}000$) at $\alpha = .05$ and $\alpha = .10$ were around 7. For $\alpha = .01$, we used a grid for I from $I = 99$ to $I = 1{,}999$ by 100 and obtained median values for $I/\sqrt{O}$ of 12 to 14. Bias seems to be relatively more important than variance for $\alpha = .01$.
2. We also found the average MSE over beta(a, b) distributions for each $I/\sqrt{O}$ value. Then we computed the $I/\sqrt{O}$ value with the lowest average MSE. For $\alpha = .05$ and $\alpha = .10$, the $I/\sqrt{O}$ ratios with lowest average MSEs were between 8 and 9. For $\alpha = .01$, the best $I/\sqrt{O}$ values were 14 for $OI = 590{,}000$ and 21 for $OI = 59{,}000$.

Thus, looking at a wide spectrum of power functions suggests $I/\sqrt{O}$ ratios around 8 for $\alpha = .05$ and $\alpha = .10$ and even higher values for $\alpha = .01$. For example, if $OI = 59{,}000$, then $O = 371, I = 139$ gives $I/\sqrt{O} = 8.25$. The high end of Oden's recommendations would be $O = 595, I = 99$ with $I/\sqrt{O} = 4.06$. Our recommendations are thus higher than those given by Oden (1991) and perhaps reflect most strongly the fact that we focus on a few specific α values rather than on minimizing maximum MSE over all α values.

7. SUMMARY

Computing time is still a major restriction when estimating the power of resampling-based hypothesis tests. The good news is that at least four-fold reductions in computing time are available at the cost of a little extra programming to estimate and remove the bias due to the use of small I. Hald's equation (4) provides the motivation for either linear or quadratic extrapolation, and Figures 1 and 2 illustrate the dependence of the bias on $1/I$. At $\alpha = .05$ or $\alpha = .10$, we suggest the simple linear estimator when $O < 4{,}000$ and the quadratic estimator when $O \geq 4{,}000$ and $I \geq 99$.

How should one choose O and I? In general, we suggest starting with the binomial variance formula and choose O so that $\sqrt{\text{power}(1 - \text{power})/O}$ is suitably small. Then choose I to be 59, 99, 199, and so on, so that OI computations are possible in the allotted time and linear or quadratic extrapolation can be used. The resulting RMSE should be only slightly above $\sqrt{\text{power}(1 - \text{power})/O}$.

If one prefers not to use extrapolation, then we suggest choosing I close to $8\sqrt{O}$ such that $(I+1)\alpha$ is an integer. In our study of beta distributions, the ratio of squared bias to MSE was often around .3; this implies that one should then expect RMSE to be around $1.2\sqrt{\text{power}(1 - \text{power})/O}$.

[Received November 1998. Revised September 1999.]

REFERENCES

Barnard, G. A. (1963), Discussion on "The Spectral Analysis of Point Processes," by M. S. Bartlett, *Journal of the Royal Statistical Society*, Ser. B, 25, 294.

Cook, J. R., and Stefanski, L. A. (1994), "Simulation-Extrapolation Estimation in Parametric Measurement Error Models," *Journal of the American Statistical Association*, 89, 1314–1328.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

Hald, A. (1968), "The Mixed Binomial Distribution and the Posterior Distribution of p for a Continuous Prior Distribution," *Journal of the Royal Statistical Society*, Ser. B, 30, 359–367.

Hall, P., and Titterington, D. M. (1989), "The Effect of Simulation Order on Level Accuracy and Power of Monte Carlo Tests," *Journal of the Royal Statistical Society*, Ser. B, 51, 459–467.

Hope, A. C. A. (1968), "A Simplified Monte Carlo Test Procedure," *Journal of the Royal Statistical Society*, Ser. B, 30, 582–598.

Jockel, K. (1986), "Finite-Sample Properties and Asymptotic Efficiency of Monte Carlo Tests," *The Annals of Statistics*, 14, 336–347.

Lock, R. H. (1991), "A Sequential Approximation to a Permutation Test," *Communications in Statistics, Part B—Simulation and Computation*, 20, 341–363.

Oden, N. L. (1991), "Allocation of Effort in Monte Carlo Simulation for Power of Permutation Tests," *Journal of the American Statistical Association*, 86, 1074–1076.