Monte Carlo Evaluation of Resampling-Based Hypothesis Tests

Dennis D. Boos and Ji Zhang *

October 1998

Abstract

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.

Institute of Statistics Mimeo Series No. 2512

Department of Statistics, North Carolina State University

*Dennis D. Boos is Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695–8203. Ji Zhang is Director, Clinical Biostatistics, Merck Research Laboratories, Rahway, NJ 07065. Email addresses are: boos@stat.ncsu.edu, zhangj@merck.com.

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 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 Oden (1991).

Suppose that we want to analyze a test procedure which 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 data sets 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 we have in mind, however, the p-value for each data set is computationally difficult, and an estimated p-value $\widehat{p}$ 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 $\widehat{p} \leq \alpha$ 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 $E\widehat{p} = p$,

$$\text{E(power estimate)} = P(\widehat{p} \leq \alpha) = E\delta(\widehat{p} \leq \alpha) \neq \delta(\text{E}\widehat{p} \leq \alpha)$$

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 since we use a sample of $\widehat{p}$'s (p's measured with error) in the estimation procedure.

If one replaces the original test procedure $p \leq \alpha$ by $\widehat{p} \leq \alpha$, then this modified test procedure is called a Monte Carlo test (introduced first by Barnard in a discussion of Bartlett, 1963), and the above 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 the an infinite number of bootstrap samples. Our thinking is that for a particular data set, one can typically make $\widehat{p}$ 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

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

where s_p^2 is the pooled sample variance. Let us consider three methods of constructing rejection regions and p-values. Let T_0 be 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, one could compute the p-value $p = 1 - t_{m+n-2}(T_0)$, where t_{m+n-2} is a central t distribution function with $m + n - 2$ degrees of freedom, and

reject H_0 if $p \leq \alpha$.

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 $T_1^*, \ldots, T_I^*$, and compute an estimated p-value

$$\widehat{p} = \frac{1}{I} \sum_{i=1}^{I} \delta(T_i^* \geq T_0),$$

   As $I \longrightarrow \infty$, $\widehat{p}$ approximates the p-value in 1. obtained from the t distribution. The related test procedure may be defined as: reject H_0 if $\widehat{p} \leq \alpha$.

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 $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ into two samples and counting the proportion of these T that are greater than or equal to T_0 calculated from the original data. Since 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, $T_1^*, \ldots, T_I^*$, and obtain $\widehat{p}$ as in 2. above.

The ideas in this paper are useful for estimating the power of procedures 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 data sets are generated under the null hypothesis or under some alternative. For each of these data sets there will also be an inner resampling loop of size I (often called B in the bootstrap literature) where further data sets 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 data sets OI generated is fixed.

Let $H_\Delta(\alpha) = P(p \leq \alpha)$ be the true power function of the test procedure at an alternative Δ. For each Monte Carlo sample, the estimated p-value $\widehat{p}$ based on I resamples (using sampling with replacement) will be such that $I\widehat{p}$ 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

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

has mean

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

where $[\alpha I]$ is the greatest integer part of αI. If H_Δ is a Beta(a,b) distribution function, then (2) becomes (cf. 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 $\widehat{p}$ 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), (1/19,.57) by least squares to a straight line yields $\widehat{H}_{\Delta,I}(\alpha) = .70 - 2.50(1/I)$ and results in a bias-reduced estimate at $1/\infty = 0$ of .70. The basic approach of this paper 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. 7-8).

Since unconditionally $O\widehat{H}_{\Delta,O,I}(\alpha)$ has a binomial $(O, H^*_\Delta(\alpha))$ distribution, the mean

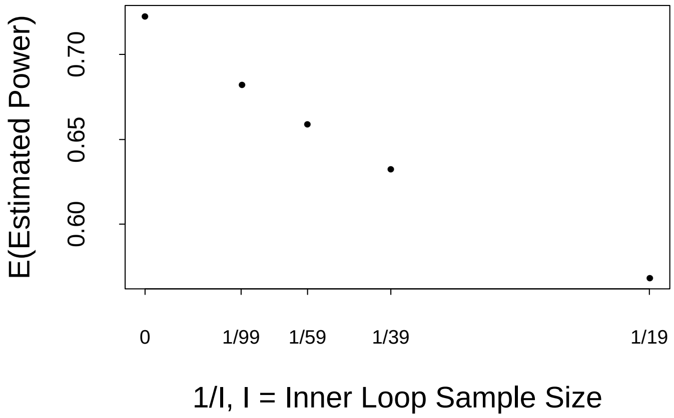

Figure 1: E(Estimated Power) at $\alpha = .05$: $H^*_{\Delta,I}(.05)$ for I $= \infty$, 99, 59, 39, 19 versus $1/I$ when $H_\Delta(t)$ is a Beta(1,25) Distribution.

squared error of $\widehat{H}_{\Delta,OI}(\alpha)$ is (cf. Oden, 1991, eq. 3)

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

Oden (1991) considered minimizing the maximum of this mean squared error 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.

Our basic extrapolation method is introduced in Section 2. Section 4 gives a Monte Carlo analysis of the method in the context of estimating the power of the two-sample parametric t test mentioned above. In Section 5 we illustrate how to use the method in practice, this time for the permutation t test. Section 6 is a brief summary.

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 data sets under the alternative, computing $\widehat{p}_{I,1}, \ldots, \widehat{p}_{I,O}$ by resampling I times for each of the data sets, and then defining

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

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 $I_1 > I_2 > \ldots > I_k$, 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 $I = \infty$, i.e., 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 the Introduction is an example. (Although the t statistic has a nuisance parameter σ estimated by s_p, the generation of the normal data sets 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 data sets under the null hypothesis and compute the test statistic for each data set, 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 $\widehat{p}_I = \{\#\text{ of } T_i^* \geq T_0\}/I$ has a discrete uniform distribution due to exchangeability of the sample:

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

The resulting test defined by the rejection region $\widehat{p} \leq \alpha$ 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 $\widehat{p}$ 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 having 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 $\widehat{p}$ by $(\{\#\text{ of } T_i^* \geq T_0\}+1)/(I+1)$, but results using the rule $\widehat{p} \leq \alpha$ 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 only results in 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) gives 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 Hald (1968) who gives an expansion for (2):

$$P_\Delta(\widehat{p}_I \leq \alpha) = H^*_{\Delta,I}(\alpha) = H_\Delta(\alpha) + \frac{B_1(\alpha)}{I} + \frac{B_2(\alpha)}{I^2} + O(I^{-3}), \qquad (4)$$

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) above would suggest, but the variablility of the estimated intercept for the linear regression is typically much smaller than that for the quadratic regression. (The variance of the least squares estimate of the intercept in the linear case must be no larger than the variance for the quadratic.) In general our recommendation is to use the linear estimate unless O values are very large.

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

$$\widehat{\text{pow}}_\infty = 1.01137(\widehat{\text{pow}}_{59}) + 0.61294(\widehat{\text{pow}}_{39}) - 0.62430(\widehat{\text{pow}}_{19}) \tag{5}$$

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

$$\begin{aligned} \widehat{\text{pow}}_\infty &= 0.46688(\widehat{\text{pow}}_{99}) + 0.41631(\widehat{\text{pow}}_{79}) + 0.33145(\widehat{\text{pow}}_{59}) \\ &+ 0.15956(\widehat{\text{pow}}_{39}) - 0.37420(\widehat{\text{pow}}_{19}) \end{aligned} \tag{6}$$

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

Equation (3) suggests another curve to fit to the raw power estimates $\widehat{\text{pow}}_I$. Suppose that for fixed Δ, $H_\Delta(\alpha)$ 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$(\widehat{a}, \widehat{b})$ 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 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 $1 - \Phi(-\Delta + \Phi^{-1}(1 - \alpha))$, 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 $\widehat{\mathbf{pow}}_{I_j}$, $j = 2, \ldots, k$. Consider the data that result from generating O data sets, computing T for each of those, denoted $T_{0k}, k = 1, \ldots, O$, and then for each outer loop situation, resampling I_1 times to get $\widehat{p}_{I_1 k}, k = 1, \ldots, O$. To be specific, for $I_1 = 59$ we have

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

Notice that for each set of resamples, we get $I_1 = 59$ 0's and 1's that we average to get the estimate $\widehat{p}_{I_1 k}$ and then average the $u_{59,k} = \delta(\widehat{p}_{I_1 k} \leq \alpha)$ to get $\widehat{\text{pow}}_{59}$. 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. Since 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:

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

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 $\widehat{\text{pow}}_{39}$. 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' 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

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

Thus to get $u_{39,k}$ we need only to 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.

Since $\widehat{\text{pow}}_{59}$, $\widehat{\text{pow}}_{39}$, and $\widehat{\text{pow}}_{19}$ 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 by 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 Sec. 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 $\widehat{\text{pow}}_{I_1}(1 - \widehat{\text{pow}}_{I_1})/I_1$ variance estimate for raw power estimates. Thus it hardly seems worthwhile estimating the correlations.

3. 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 $\Delta/\delta = 0.5$, 1.0, 1.5, and 2.0 with true powers 0.189, 0.451, 0.737, and 0.918, respectively.

For this study we need three simulation loops. The outside loop is of size 100 replications. Then the two inner loops are actually what we have called previously 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, $(O = 1000, I = 59)$ and $(O = 596, I = 99)$, that have about 59000 computations each.

Table 1 reports estimates of the root mean squared error ($\sqrt{\text{MSE}}$) and bias $\times$ 1000 of the various power estimates. The standard errors of the estimates are in the range .001 to .002 for $\sqrt{\text{MSE}}$ and around 2 for the bias $\times$ 1000.

The first and seventh rows (p_∞) of Table 1 give results for power estimates based on the true known t percentiles appropriate for normal data. They are labeled p_∞ 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 $\sqrt{\text{MSE}}$ could have been calculated simply by $\sqrt{\text{power(1-power)}/O}$. For a given O, p_∞ represents the best power estimates possible.

The second ($\widehat{p}_{59}$) and eighth rows ($\widehat{p}_{99}$) 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 $\sqrt{\text{MSE}}$ than $(O = 1000, I = 59)$ for all but $\Delta = 0.5$ because the bias is a large factor except at $\Delta = 0.5$.

The other estimators in Table 1 are

1. $\widehat{p}_{\text{lin}}$: the simple linear extrapolation method using (5) for the $(O = 1000, I = 59)$ case and (6) for the $(O = 596, I = 99)$ case. In either case $\widehat{p}_{\text{lin}}$ is just the intercept estimator from ordinary least squares simple linear regression.

2. $\widehat{p}_{\text{gls}}$ is similar to $\widehat{p}_{\text{lin}}$ except that estimated generalized least squares is used, based on empirical covariance estimates for ($\widehat{p}_{59}$,$\widehat{p}_{39}$,$\widehat{p}_{19}$) and ($\widehat{p}_{99}, \widehat{p}_{79}, \widehat{p}_{59}, \widehat{p}_{39}, \widehat{p}_{19}$), respectively.

3. $\widehat{p}_{\text{quad}}$ is the quadratic extrapolation estimator, i.e., the ordinary least squares intercept estimator from a simple quadratic model.

4. $\widehat{p}_{\text{beta}}$ 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$(\widehat{a}, \widehat{b})$ distribution.

From Table 1 we see that the the linear extrapolation estimators, $\widehat{p}_{\text{lin}}$ and $\widehat{p}_{\text{gls}}$, perform the best and very similarly. Their similarity is likely due to the fact (not displayed) that the estimated covariance matrix of the $\widehat{p}_I$ used as dependent variables in the regressions has nearly equal diagonal elements and nearly equal off-diagonal elements. The estimators $\widehat{p}_{\text{lin}}$ and $\widehat{p}_{\text{gls}}$ 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 since their $\sqrt{\text{MSE}}$ values are close to $\sqrt{\text{power(1-power)}/O}$, it hardly seems worth the trouble.

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 $\sqrt{\text{MSE}}$ of $\widehat{p}_{\text{quad}}$ 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 $O = 8900$ the linear and quadratic estimators would have the same mean squared error. For $I = 99$, we estimate that at $O = 3700$ they would have the same mean squared error. It is interesting that the quadratic estimator is more efficient at the $(O = 596, I = 99)$ combination than at the $(O = 1000, I = 59)$ 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 $\widehat{p}_{\text{beta}}$ performed almost as well as the linear estimators except for the $\Delta = 0.5$ alternative where the nonlinear least squares program had trouble converging.

Should one go to the extra effort of computing $\widehat{\text{pow}}_{39}$ and $\widehat{\text{pow}}_{19}$ and plugging in equation (5) to get $\widehat{p}_{\text{lin}}$? At $\Delta = 0.5$ and in general near the null hypothesis, the bias of the raw estimator $\widehat{p}_{59}$ is small and $\widehat{p}_{\text{lin}}$ has little advantage over $\widehat{p}_{59}$. However, at $\Delta = 1.5$ and $\Delta = 2.0$, the mean squared error of $\widehat{p}_{\text{lin}}$ is much smaller than the mean squared error of the raw estimator. Using the combinations of O and I that are approximately optimal for the raw estimator ($(O, I) = (1100, 219)$ for $\Delta = 1.5$ and $(O, I) = (800, 279)$ for $\Delta = 2.0$), we find that it would take about 4 times as large an OI value for the raw estimator to have mean squared errors equal to $\widehat{p}_{\text{lin}}$. And the savings could be even greater since we rarely would know the optimal values for O and I in a new situation.

Table 1. $\sqrt{\text{MSE}}$ and Bias of Power Estimates for Two-sample t-Statistic with Parametric Bootstrap Critical Values

	$\sqrt{\text{MSE}}$				bias × 1000			
$\Delta =$	0.5	1.0	1.5	2.0	0.5	1.0	1.5	2.0
				$O = 1000$, $I = 59$				
p_∞	.012	.017	.015	.010	1.4	0.3	−1.6	0.1
$\widehat{p}_{59}$	.013	.031	.041	.030	−6.0	−25.7	−38.3	−28.4
$\widehat{p}_{\text{lin}}$	.013	.020	.017	.011	−0.4	−6.4	−6.0	3.4
$\widehat{p}_{\text{gls}}$	.013	.020	.018	.010	−1.2	−8.4	−7.0	3.9
$\widehat{p}_{\text{quad}}$	.018	.025	.025	.019	0.7	−0.9	−1.8	1.7
$\widehat{p}_{\text{beta}}$	*	.022	.018	.012	*	2.4	2.1	6.2
				$O = 596$, $I = 99$				
p_∞	.016	.019	.018	.010	3.8	0.2	−0.6	0.8
$\widehat{p}_{99}$	.018	.025	.030	.019	−2.2	−15.8	−23.9	−15.3
$\widehat{p}_{\text{lin}}$	.018	.021	.019	.013	1.1	−4.8	−5.1	4.4
$\widehat{p}_{\text{quad}}$	.021	.023	.022	.015	1.9	−0.9	−2.6	2.2
$\widehat{p}_{\text{beta}}$	*	.023	.020	.013	*	1.1	0.0	6.2

* Convergence problems with the nonlinear least squares routine. Results based on 100 replications. Standard errors for $\sqrt{\text{MSE}}$ $\approx .001 - .002$. Standard errors for bias × 1000 ≈ 2.

4. 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 0.189, 0.451, 0.737, and 0.918, respectively, for standardized mean difference $\Delta/\sigma = 0.5$, 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 the Introduction that makes no use of the fact that the data are normally distributed.

First we ran all four alternatives with $O = 1000$ and $I = 59$ and obtained the linear estimators based on (5). Using Splus these four runs took about 2.5 hours each on a Sparcstation 4-110MHz. The estimated standard errors using the least squares formula $(X^TX)^{-1}X^T[\widehat{\Sigma}/1000]X(X^TX)^{-1}$ for the intercept were very close to those from the binomial calculation $\sqrt{\text{power}(1-\text{power})/O}$. Here $\widehat{\Sigma}$ is the sample covariance matrix from the 1000 by 3 matrix with kth row $(u_{59,k}, u_{39,k}, u_{19,k})$.

Since 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 mean squared error to the quadratic estimators. These runs took about 18 hours each. 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 0.189, 0.451, 0.737, and 0.918. 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 (+).

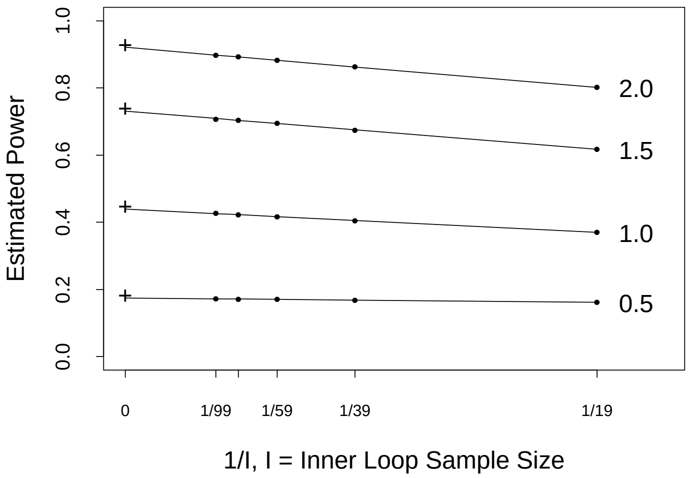

Figure 2: Estimated power of permutation t test at $\alpha = .05$ by linear extrapolation. Data are normally distributed with standardized mean difference $\Delta/\sigma = 0.5, 1.0, 1.5, 2.0$, $n_1 = 8, n_2 = 4$.

5. Choice of (O, I) with No Extrapolation

The thesis of this paper 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 simplest to use and program. Thus we want to give a little guidance on (O, I) combinations that have relatively low mean squared error.

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 $I = 2\sqrt{O}$ to $I = 4\sqrt{O}$. This seems like a good place to start, and we will 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 mean squared error under the null hypothesis may not be relevant for estimation of 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 Beta(a, b) distribution functions. For those situations we found the $I/\sqrt{O}$ ratios with lowest mean squared error 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 mean squared error is $(O, I) = (296, 199)$ and $I/\sqrt{O} = 11.6$. For Beta(.1,1.0) the optimal combination is $(O, I) = (746, 79)$ with $I/\sqrt{O} = 2.9$. This example shows the wide variety 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) distribution that hopefully covers 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 mean squared error of (1) for $I = 19$ to 499 by 20 and for $OI = 59,000$. We repeated the process for $OI = 590,000$. Then

1. We found the (O, I) combination with lowest mean squared error 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 $I = 1999$ 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 mean squared error over Beta(a, b) distributions for each $I/\sqrt{O}$ value. Then we computed the $I/\sqrt{O}$ value with the lowest average mean squared error. For $\alpha = .05$ and $\alpha = .10$, the $I/\sqrt{O}$ ratios with lowest average mean squared errors 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 mean squared error over all α values.

6. 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 $O < 4000$ and the quadratic estimator when $O \geq 4000$ and $I \geq 99$. Splus code is available from the first author upon request.

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, etc., so that OI computations are possible in the allotted time and so that linear or quadratic extrapolation can be used. The resulting $\sqrt{\text{mse}}$ should be only slightly above $\sqrt{\text{power}(1-\text{power})/O}$. If you prefer 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 mean squared error was often around .3; this implies that one should then expect $\sqrt{\text{mse}}$ to be around $1.2\sqrt{\text{power}(1-\text{power})/O}$.

Acknowledgement

We would like to 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.

REFERENCES

Barnard, G. A. (1963), "Discussion on The Spectral Analysis of Point Processes (by M. S. Bartlett)," *Journal of the Royal Statistical Society, Series 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*, Society for Industrial and Applied Mathematics, Philadelphia.

Hald, A. (1968), "The Mixed Binomial Distribution and the Posterior Distribution pf p for a Continuous Prior Distribution," *Journal of the Royal Statistical Society, Series 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, Series B*, 51, 459-467.

Hope, A. C. A. (1968), "A Simplified Monte Carlo Test Procedure," *Journal of the Royal Statistical Society, Series 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.