

ADJUSTED POWER ESTIMATES IN MONTE CARLO EXPERIMENTS

Ji Zhang

Biostatistics and Research Data Systems
Merck Research Laboratories
Rahway, NJ 07065-0914

and

Dennis D. Boos

Department of Statistics, North Carolina State University
Raleigh, NC 27695-8203

Key Words and Phrases: Critical value; empirical level; McNemar's test; true level.

ABSTRACT

Critical values and powers of competing tests are often evaluated through Monte Carlo simulations. Since the true levels of those tests are often very different, comparisons of the power estimates are not valid. We suggest that Monte Carlo estimates of critical values be used to create adjusted power estimates which are then comparable. The main contribution is to analyze the variability of the adjusted estimates and to point out implications for the planning of Monte Carlo power studies.

1 INTRODUCTION

Monte Carlo experiments are often the simplest way to estimate and compare the power functions of complex test statistics. Unfortunately, at the end of such studies one may find results such as the following:

TABLE I

	$\delta = 0$	$\delta = 1.5$	$\delta = 3.0$	$\delta = 4.5$
Test 1:	.08	.24	.37	.74
Test 2:	.03	.20	.30	.62

where $\delta = 0$ corresponds to the null hypothesis, .05 is the nominal level of the tests, and the entries such as .08 are the proportion of test rejections in N Monte Carlo replications. If N is large enough so that the difference $.08 - .03$ is not the result of sampling variability, then these two power functions are not comparable in the usual sense because Test 1 has a “head start” over Test 2.

Thus it would seem important to adjust the power curves if comparing them is of interest. The simple solution is to just estimate the true critical values by Monte Carlo simulation and use those estimates throughout the study. Many experimenters would automatically take this approach.

The purpose of this paper is to analyze such power estimates when the critical values themselves have been estimated by Monte Carlo methods. We use the term “adjusted power estimates” because test statistics often come with standard critical values (usually from asymptotic arguments) such as normal or chi-squared quantiles. We then think of the use of estimated critical values as an adjustment to make the power curves comparable. Table II in Section 2 shows how we prefer to display the final results.

The paper is organized as follows. Section 2 introduces the notation and definitions. Section 3 gives an asymptotic analysis of the mean and variance of the adjusted power estimate, and Section 4 comments on the effect the estimation of critical values has on McNemar's test for equality of power curves. Results in both Sections 3 and 4 have implications for the planning of Monte Carlo power studies.

2 THE ADJUSTED POWER ESTIMATE

Consider a test statistic T for which a critical value C_α^* of nominal level α has been proposed. A typical Monte Carlo experiment might draw N_0 independent samples at the null hypothesis and N_1, \dots, N_k samples at points in the alternative hypothesis. For simplicity we shall often assume $k=1$. The estimated true level of the test is then

$$\hat{p}_0^* = \frac{1}{N_0} \sum_{i=1}^{N_0} I(T_{0i} \geq C_\alpha^*), \quad (1)$$

where T_{01}, \dots, T_{0N_0} are the test statistics for the N_0 Monte Carlo samples under the null hypothesis, and $I(A)$ is the indicator function of a set A such that $I(A) = 1$ if A is true and 0 otherwise. Thus the sum above just counts the number of test rejections using the critical value C_α^* . Similarly, the estimated power at an alternative is

$$\hat{p}_1^* = \frac{1}{N_1} \sum_{i=1}^{N_1} I(T_{1i} \geq C_\alpha^*), \quad (2)$$

where T_{11}, \dots, T_{1N_1} are the test statistics for the N_1 Monte Carlo samples under the alternative hypothesis.

A typical result for \hat{p}_0^* and \hat{p}_1^* are the values .08 and .24 in row 1 of Table I. If N_0 is larger than about 200, then a simple binomial test of H_0 : true level =

.05 would suggest that the true level is bigger than the nominal level $\alpha = .05$. This makes the value .24 not comparable to power results for other tests with true level = .05 or for a test such as in the second row of Table I.

At this point the experimenters appear forced to rerun the experiment unless they have saved the individual outcomes from each Monte Carlo replication (which is always a good practice if the cost of running the experiment is fairly high). We shall assume that they have saved those values: T_{01}, \dots, T_{0N_0} under the null and T_{11}, \dots, T_{1N_1} under the alternative. In this case they can estimate the critical value for T from the $(1 - \alpha)$ quantile of the null observed values, i.e., $\hat{C}_\alpha =$ the $(1 - \alpha)N_0$ th value of T_{01}, \dots, T_{0N_0} when these are placed in order from smallest to largest.

The adjusted power estimate is found by replacing C_α^* by \hat{C}_α in (2) resulting in

$$\hat{p}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} I(T_{1i} \geq \hat{C}_\alpha). \quad (3)$$

Of course the adjusted power at the null hypothesis is just α because of the way \hat{C}_α is chosen. If we carried out these calculations for the two statistics in Table I, we might replace Table I by

TABLE II

	$\delta = 0$	$\delta = 1.5$	$\delta = 3.0$	$\delta = 4.5$
Test 1:	.08 (.05)	.24 (.19)	.37 (.30)	.74 (.62)
Test 2:	.03 (.05)	.20 (.22)	.30 (.33)	.62 (.68)

The adjusted powers in parentheses are then comparable. On the other hand we have left the original power estimates in Table II because they show what kind of power one would obtain by using the standard critical value C_α^* even though such a test does not have level α .

In the next section we will analyze the adjusted power estimate \hat{p}_1 .

3 PROPERTIES OF THE ADJUSTED POWER ESTIMATE

Let $F_0(x) = P(T \leq x)$ and $F_1(x) = P(T \leq x)$ be the distribution functions of the test statistic T under the null and alternative hypotheses, respectively. The true level α critical value C_α satisfies $F_0(C_\alpha) = 1 - \alpha$, and $1 - F_1(C_\alpha)$ is the power which \hat{p}_1 tries to estimate.

We assume that F_0 and F_1 are twice differentiable at C_α and that their densities satisfy $0 < f_0(C_\alpha) < \infty$ and $0 < f_1(C_\alpha) < \infty$. Then using results from the Bahadur representation theory found in Serfling (1980, Sec. 2.5), we have that

$$\begin{aligned} \hat{p}_1 - [1 - F_1(C_\alpha)] &= -f_1(C_\alpha) \frac{1}{N_0} \sum_{i=1}^{N_0} \left[\frac{(1 - \alpha) - I(T_{0i} \leq C_\alpha)}{f_0(C_\alpha)} \right] \\ &\quad + \frac{1}{N_1} \sum_{i=1}^{N_1} I(T_{1i} \geq C_\alpha) + R_{N_0} + R_{N_1}, \end{aligned} \quad (4)$$

where $\sqrt{N_0}R_{N_0} \xrightarrow{p} 0$ and $\sqrt{N_1}R_{N_1} \xrightarrow{p} 0$ as $N_0 \rightarrow \infty$, $N_1 \rightarrow \infty$. Then by the CLT applied to (4) we have that \hat{p}_1 is asymptotically normal with mean $1 - F_1(C_\alpha)$ and variance

$$\text{AVAR} = \frac{f_1^2(C_\alpha) \alpha(1 - \alpha)}{f_0^2(C_\alpha) N_0} + \frac{F_1(C_\alpha)[1 - F_1(C_\alpha)]}{N_1}. \quad (5)$$

Using Taylor expansions we can give further insight by directly approximating the mean and variance of \hat{p}_1 :

$$\begin{aligned} E(\hat{p}_1) &= E E(\hat{p}_1 | T_{01}, \dots, T_{0N_0}) \\ &= E[1 - F_1(\hat{C}_\alpha)] \\ &= [1 - F_1(C_\alpha)] + O(N_0^{-1}) \end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{p}_1) &= \text{Var} E(\hat{p}_1 | T_{01}, \dots, T_{0N_0}) + E \text{Var}(\hat{p}_1 | T_{01}, \dots, T_{0N_0}) \\
&= \text{Var}[1 - F_1(\hat{C}_\alpha)] + \frac{E F_1(\hat{C}_\alpha)[1 - F_1(\hat{C}_\alpha)]}{N_1} \\
&= \frac{f_1^2(C_\alpha) \alpha(1 - \alpha)}{f_0^2(C_\alpha) N_0} + O(N_0^{-2}) + \frac{F_1(C_\alpha)[1 - F_1(C_\alpha)]}{N_1} + O(N_1^{-2}) \\
&= \text{AVAR} + O(N_0^{-2}) + O(N_1^{-2}).
\end{aligned}$$

These latter results actually require further assumptions on f_0 and f_1 but show more clearly the error of approximation. In particular we see that the square of the bias is of lower order than the terms of AVAR, and thus we can concentrate attention on AVAR.

Note first that AVAR has two components, the first depends on the squared ratio of the densities at C_α divided by N_0 , and the second is the usual binomial variance for when C_α is known instead of estimated.

To see the relative importance of the two terms of AVAR, let $\text{AVAR} = A/N_0 + B/N_1$. Figure 1 plots the ratio A/B versus power for a one-sided test of a normal mean with known variance for three values of α , $\alpha = .10$, $.05$, and $.01$. The results are in ascending order with the maximum of the ratio = 1.86 at $\alpha = .10$, = 2.22 at $\alpha = .05$, and = 8.87 at $\alpha = .01$.

The calculations in Figure 1 are simple: at $\alpha = .05$ $f_1(C_\alpha) = \phi(1.645 - \delta)$, $f_0(C_\alpha) = \phi(1.645)$, and $\text{power}(\delta) = 1 - F_1(C_\alpha) = 1 - \Phi(1.645 - \delta)$, where ϕ and Φ are the standard normal density and distribution functions, respectively, and δ is the alternative. Similar computations were also carried out for chi-squared type tests. Surprisingly, results very similar to Figure 1 hold true for any test which has a non-central chi-squared power function. Thus for such cases it would appear that the variance of \hat{p}_1 is about three times the variance of \hat{p}_1^* when α is near $.05$ and $N_0 = N_1$ and about nine times as much when α is near $.01$.

If this analysis is made after the experiment is completed, then not much

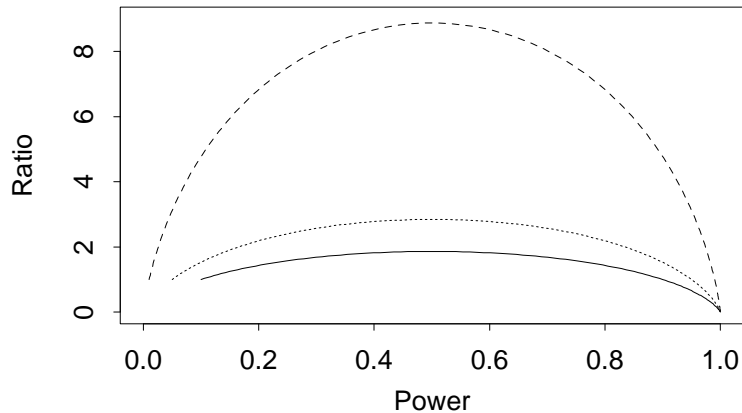


Figure 1: Ratio of terms in AVAR for one-sided mean test with normal data and known variance. In ascending order the curves are for $\alpha = .10, .05,$ and $.01$.

else can be said. However, if one knows before the experiment that adjusted power will be needed (or equivalently that Monte Carlo critical values \hat{C}_α will be required), then we might try to optimize the choice of sample sizes N_0, \dots, N_k , at least if the value of α is specified (and using the variance ratios suggested by Figure 1). Rather than give formal procedures, we merely suggest that N_0 be chosen considerably larger than N_1, \dots, N_k . For example, if $k > 1$ a rough rule of thumb might be $N_0 = 10N_1, N_1 = N_2 = \dots = N_k$.

4 TESTS FOR EQUALITY OF TWO POWER FUNCTIONS

Sometimes it is of interest to provide a formal statistical test of whether two power functions are equal. Consider first the results in Table I. Suppose that one wanted to compare the estimates .24 and .20 at $\delta = 1.5$ (ignoring for the moment that the true levels are not equal). If the individual test results are available for each of the N_1 Monte Carlo replications, then one would form

the two by two table

		Test 2		Total
		Reject	Accept	
Test 1	Reject	a	b	$a + b$
	Accept	c	d	$c + d$
Total		$a + c$	$b + d$	N_1

and use McNemar's test to test for equal power (see, e.g., Agresti, 1990, p. 350). The approximate normal statistic typically used is

$$Z = \frac{b - c}{\sqrt{b + c}}. \tag{6}$$

An exact test may be obtained by noting that b conditioned on $b + c$ is binomially distributed with $b + c$ trials and $p = 1/2$. McNemar's test is required here because the rejection results for Test 1 and Test 2 are paired since the statistics are both computed on the same Monte Carlo samples.

Now suppose that we want to use McNemar's test to compare the adjusted power estimates. For notation we let $\hat{a}, \hat{b}, \hat{c}, \hat{d}$ be the counts in the 2 x 2 table when Monte Carlo estimated critical values are used, and $\hat{Z} = (\hat{b} - \hat{c})/\sqrt{\hat{b} + \hat{c}}$.

Using expansions similar to (4) we can show that $\hat{Z} \xrightarrow{d} N(0, 1 + \lambda V)$ as $N_0 \rightarrow \infty$ and $N_1 \rightarrow \infty$ with $N_1/N_0 \rightarrow \lambda$, $0 \leq \lambda < \infty$, and V is a positive constant depending on α and on the joint density of the test statistics under both null and alternative hypotheses.

This result tells us that we need to choose N_0 large relative to N_1 in order to justify the use of $N(0, 1)$ critical values with \hat{Z} . Of course the magnitude of V is also important but hard to characterize in general because of the way that it depends on the joint densities of the test statistics.

BIBLIOGRAPHY

Agresti, A. (1990). *Categorical Data Analysis*, New York: John Wiley.

Serfling, R. J. (1980). *Approximation Theorems of Statistics*, New York: John Wiley.