

Racine, Jeff; MacKinnon, James G.

**Working Paper**

## Simulation-based Tests that Can Use Any Number of Simulations

Queen's Economics Department Working Paper, No. 1027

**Provided in Cooperation with:**

Queen's University, Department of Economics (QED)

*Suggested Citation:* Racine, Jeff; MacKinnon, James G. (2004) : Simulation-based Tests that Can Use Any Number of Simulations, Queen's Economics Department Working Paper, No. 1027, Queen's University, Department of Economics, Kingston (Ontario)

This Version is available at:

<https://hdl.handle.net/10419/189311>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



Queen's Economics Department Working Paper No. 1027

# Simulation-based Tests that can Use Any Number of Simulations

Jeff Racine  
McMaster University

James MacKinnon  
Queen's University

Department of Economics  
Queen's University  
94 University Avenue  
Kingston, Ontario, Canada  
K7L 3N6

10-2004

# SIMULATION-BASED TESTS THAT CAN USE ANY NUMBER OF SIMULATIONS

JEFF RACINE AND JAMES G. MACKINNON

ABSTRACT. Conventional procedures for Monte Carlo and bootstrap tests require that  $B$ , the number of simulations, satisfy a specific relationship with the level of the test. Otherwise, a test that would instead be exact will either overreject or underreject for finite  $B$ . We present expressions for the rejection frequencies associated with existing procedures and propose a new procedure that yields exact Monte Carlo tests for any positive value of  $B$ . This procedure, which can also be used for bootstrap tests, is likely to be most useful when simulation is expensive.

## 1. INTRODUCTION

The use of simulation methods for inference in finite samples is well established. In the case of Monte Carlo tests, where the simulated test statistics follow the same distribution as the actual one under the null hypothesis, these methods yield exact inferences; see Dwass [5], Jöckel [10], and Hall and Titterington [9]. In the case of bootstrap tests based on asymptotically pivotal test statistics, they generally yield inferences that improve more rapidly with the sample size than tests based on asymptotic theory; see Beran [1], Hall [8], and Mammen [11], among many others.

Standard procedures for Monte Carlo and bootstrap tests require that  $B$  be chosen so that  $\alpha(B+1)$  is an integer, where  $\alpha$  is the level of the test. When computation is expensive, this condition may be burdensome, especially if  $\alpha$  is small. For  $\alpha = .01$ , it is impossible to perform a Monte Carlo test with  $B < 99$  using standard procedures.

In this paper, we propose a new procedure for performing Monte Carlo tests when  $B$ , the number of simulations, can take on any positive value. Formally, our analysis deals only

---

*Date:* October 26, 2004.

*Key words and phrases.* Resampling, Monte Carlo test, percentiles.

We would like to thank Peter Hall for helpful suggestions and guidance that led to a greatly improved version of this paper. We are also grateful to Anthony Davison for comments. Racine's research was supported by National Science Foundation award #BCS-0320284. MacKinnon's research was supported by a grant from the Social Sciences and Humanities Research Council of Canada.

with the case in which the distribution of the test statistic  $\tau$ , of which the realized value is  $\hat{\tau}$ , is identical under the null hypothesis to the distribution of the simulated statistics  $\tau_j^*, j = 1, \dots, B$ . However, this procedure can also be applied to bootstrap tests based on asymptotically pivotal test statistics. In the latter case, it will generally not yield an exact test, but, as in Hall [7], where the condition that  $\alpha(B + 1)$  is an integer is satisfied, the error due to bootstrapping as  $n \rightarrow \infty$  will be of the same order for any finite  $B$  as it is for  $B = \infty$ .

Our procedure is not needed when computation is cheap. However, there are many situations in which it is not. One such situation is when simulation-based methods are required to estimate the model; see van Dijk and Monfort [12] and Gouriéroux and Monfort [6] for introductions to simulation-based methods in econometrics. Another situation is when simulation is required to compute standard errors, or, more generally, covariance matrices. Even when standard errors can be calculated without simulation, they may not be reliable. In such cases, the double bootstrap proposed by Beran [1] can be used to obtain more accurate inferences. But this requires a substantial number of second-level simulations for each of  $B$  first-level simulations and can therefore be very expensive.

Monte Carlo tests with finite  $B$  involve some loss of power relative to (infeasible) tests that use  $B = \infty$ ; see Jöckel [10] and Hall and Titterton [9]. Of course, our procedure is not immune to this problem. But, as we show in Section 4, our procedure does not introduce any additional power loss, and power seems to increase monotonically with  $B$ .

## 2. CONVENTIONAL EDF $P$ VALUES

One can perform a Monte Carlo or bootstrap test either by comparing the actual test statistic  $\hat{\tau}$  with a critical value computed from the  $\tau_j^*$  or by computing a bootstrap  $P$  value using both  $\hat{\tau}$  and the  $\tau_j^*$  and then comparing it with the level of the test. Our procedure is based on the  $P$  value approach, which we now describe.

The most intuitive way to compute a bootstrap  $P$  value based on  $B$  simulated test statistics, for a test that rejects in the upper tail, is to estimate the empirical distribution function

(EDF) of the bootstrap statistics. Then the bootstrap  $P$  value is

$$(1) \quad P_B^* = 1 - \hat{F}(\hat{\tau}) = 1 - \frac{1}{B} \sum_{j=1}^B I(\tau_j^* \leq \hat{\tau}) = \frac{1}{B} \sum_{j=1}^B I(\tau_j^* > \hat{\tau}) = \frac{N}{B},$$

where  $I(\cdot)$  is the indicator function,  $\hat{F}(\hat{\tau})$  is the EDF of the bootstrap statistics, and  $N \equiv \sum_{j=1}^B I(\tau_j^* > \hat{\tau})$  is the number of simulated test statistics greater than  $\hat{\tau}$ . We reject the null hypothesis whenever  $P_B^* < \alpha$ . This procedure is equivalent to using  $\hat{F}(\hat{\tau})$  to estimate a critical value and rejecting the null whenever  $\hat{\tau}$  exceeds that critical value.

When  $B$  is chosen so that  $\alpha(B+1)$  is an integer and the test statistic is pivotal, this procedure yields an exact test. The reason why Monte Carlo tests are exact in this case is easy to see. If  $\hat{\tau}$  and the  $\tau_j^*$  all follow the same distribution, then the probability that  $\hat{\tau}$  will be among the  $\alpha(B+1)$  most extreme values by chance is exactly  $\alpha$ . But this is precisely the situation in which  $P_B^*$  is less than  $\alpha$ .

Some authors (for example, Davison and Hinkley [4]) use a modified version of the EDF approach in which  $P_B^*$  is replaced by

$$(2) \quad P'_B = \frac{N}{B+1} + \frac{1}{B+1}$$

and reject whenever  $P'_B \leq \alpha$ . This *biased EDF* approach leads to precisely the same inferences as the usual EDF approach based on (1) whenever  $\alpha(B+1)$  is an integer, but it leads to different results when this condition is not satisfied. A test based on (2) can never reject at all when  $B$  is less than the smallest value for which  $\alpha(B+1)$  is an integer.

It is well known that  $\hat{F}(\hat{\tau})$  has a rectangular (discrete uniform) distribution. Therefore, so will  $P_B^* = 1 - \hat{F}(\hat{\tau})$ . Thus, under the null,  $P_B^*$  assumes  $B+1$  equally probable potential outcomes. The expected rejection frequency for a test based upon  $P_B^*$  under the null can be shown to be

$$(3) \quad E[I(P_B^* < \alpha)] = \Pr[P_B^* < \alpha] = \frac{\text{ceiling}(\alpha B)}{B+1},$$

where  $\text{ceiling}(x)$  is a function returning the smallest integer not less than  $x$ . It can similarly be shown that

$$(4) \quad \mathbb{E}[I(P'_B \leq \alpha)] = \Pr[P'_B \leq \alpha] = \frac{\text{floor}(\alpha(B+1))}{B+1},$$

where  $\text{floor}(x)$  is a function returning the largest integer not greater than  $x$ . From these results, it can be seen that tests based on  $P_B^*$  and  $P'_B$  yield identical inferences when either  $\alpha(B+1)$  or  $\alpha B$  is an integer. Both tests are exact in the former case and underreject in the latter. Expressions (3) and (4) do not appear to be well known.

Figures 1 and 2, which are based on these results, plot rejection frequencies under the null for the standard and biased EDF approaches as a function of  $B$  when  $\alpha = .05$  and  $\alpha = .01$ , respectively. The standard EDF approach often overrejects severely. The biased EDF approach never overrejects, but it often underrejects severely. Both approaches can work very badly when  $B$  is small.

Let  $B_{\min}(\alpha)$  denote the smallest value of  $B$  such that  $\alpha(B+1)$  is an integer. From equations (3) and (4), it is easy to see that  $B_{\min}(\alpha) = 9$  when  $\alpha = .10$ , 19 when  $\alpha = .05$ , and 99 when  $\alpha = .01$ . The constraint that  $B$  not be less than  $B_{\min}(\alpha)$  is particularly troublesome when  $\alpha$  is small. When computation is expensive, it may be burdensome to have to perform at least 99 simulations to perform a test at the .01 level.

### 3. CONTINUOUS EDF $P$ VALUES

The size distortions associated with the conventional EDF approach arise because  $P_B^*$  (and also  $P'_B$ ) can take on only  $B+1$  discrete values. Our proposal is to modify  $P_B^*$  so that it is uniformly distributed on the  $[0, 1]$  interval by making use of a single draw from the uniform distribution in addition to the  $B$  simulated statistics.

The modified (continuous) EDF  $P$  value is given by

$$(5) \quad P_B^c = \frac{N}{B+1} + \frac{U}{B+1},$$

where  $U \sim U[0, 1]$ . Thus it has almost the same form as the biased EDF  $P$  value  $P'_B$ , but the quantity that is added is  $U/(B + 1)$  instead of  $1/(B + 1)$ . Because this quantity is a continuous random variable,  $P_B^c$  must be continuous. It is easy to see that the smallest value it can take is 0 and the largest value it can take is 1, and that these extremes both occur with zero probability.

In fact, under the null hypothesis,  $\Pr[P_B^c < \alpha] = \alpha$  for any finite  $B$ . In other words,  $P_B^c$  is itself distributed as  $U[0, 1]$ . It is easy to see why this is the case. The first term on the right-hand side of equation (5) is a discrete random variable that can take on  $B + 1$  possible values. Under the null hypothesis, each of these values is equally probable. The second term is a continuous random variable that follows the  $U[0, 1/(B + 1)]$  distribution. This second random variable fills in the gaps between the values that the first random variable can take. For each value of  $N$ ,  $P_B^c$  is uniformly distributed between  $N/(B + 1)$  and  $(N + 1)/(B + 1)$ . Since every value of  $N$  is equally probable,  $P_B^c$  must be  $U[0, 1]$ . For a formal proof of this proposition, see the Appendix.

Our procedure is formally quite similar to a randomization test; see [2], Section 4.5. It may thus be subject to the criticism that we are adding a source of randomness which has nothing to do with the original problem. However, there is a fundamental difference between our procedure and randomization tests. In our case, the underlying test statistic  $\tau$  is continuous, not discrete. Any Monte Carlo or bootstrap test based on finite  $B$  involves simulation error, whether or not our procedure is used. This error can always be made arbitrarily small by making  $B$  sufficiently large. That is as true for the simulation error associated with  $U$  as for the errors associated with the  $\tau_j^*$ , because the randomness in the second term on the right-hand side of equation (5) is, in principle, no different from the randomness in the first term that arises from simulating the  $\tau_j^*$ .

It is easy to see that  $P_B^c$  must yield exactly the same inferences as  $P_B^*$  and  $P'_B$  whenever  $\alpha(B + 1)$  is an integer. In this special case,  $\Pr(P_B^c \leq \alpha) = \Pr(N < \alpha(B + 1))$  because, whenever  $N < \alpha(B + 1)$ , it must be less by at least 1. But this implies that  $N + U \leq \alpha(B + 1)$ .

#### 4. POWER LOSS AND $B$

It is well-known that the randomness introduced by simulation causes Monte Carlo and bootstrap tests to lose power, and that this loss of power is proportional to  $1/B$ . For discussions of this type of power loss, see Jöckel [10], Hall and Titterton [9], and Davidson and MacKinnon [3]. Our procedure is, of course, subject to exactly the same type of power loss as standard ones based on  $P_B^*$  or  $P_B'$ . However, evidence from simulation strongly suggests that the power of our procedure is monotonically increasing in  $B$ . Moreover, as we have noted, our procedure yields exactly the same inferences as standard ones whenever  $\alpha(B+1)$  is an integer. It follows that using our procedure can never harm test power. If we have performed  $B$  simulations, where  $\alpha(B+1)$  is not an integer, it is always better to base a test on  $P_B^c$  than to throw away  $b$  simulations so that  $\alpha(B-b+1)$  is an integer and base a test on  $P_{B-b}^*$ .

There is potentially a severe loss of power whenever  $B < B_{\min}(\alpha)$ . In this case, the power of the test is bounded from above. Suppose that the null hypothesis is seriously false, so that the power of the test when  $B = \infty$  is unity. This implies that  $\tau_j^* < \hat{\tau}$  for every simulation, so that  $N = 0$ . When  $N = 0$ ,

$$(6) \quad \Pr(P_B^c < \alpha) = \Pr(U/(B+1) < \alpha) = \Pr(U < \alpha(B+1)) = \alpha(B+1).$$

Thus the test can never have power greater than  $\alpha(B+1)$ . When  $B$  and  $\alpha$  are both small, this bound may be quite low. For example, when  $B = 9$  and  $\alpha = .01$ , it is just .10. Of course, the bound in equation (6) has no force when  $B > B_{\min}(\alpha)$ .

Figures 3 and 4 show power functions for tests at the .05 and .01 levels for various values of  $B$ . The test statistic actually follows the  $N(\delta, 1)$  distribution, with  $\delta = 0$  under the null hypothesis. In both figures, the topmost curve, which is labelled  $B = \infty$ , is based on the  $N(0, 1)$  distribution, and the other curves are based on 5 million replications with the indicated values of  $B$ . In both figures, it can be seen that there can be quite a lot of power



loss when  $B < B_{\min}(\alpha)$ . Using  $B = B_{\min}(\alpha)$ , which is 19 for  $\alpha = .05$  and 99 for  $\alpha = .01$ , still leads to noticeable power loss. This is essentially halved by using  $B = 2B_{\min}(\alpha) + 1$ .

The effect of  $B$  on power loss is more clearly seen in Figures 5 and 6. The first of these shows the difference between power when  $B = \infty$  and power for a specified value of  $B$  plotted against  $B$  for tests at the .05 level for three different values of  $\delta$ . As the bound in equation (6) suggests, power loss is very great when  $\delta$  is large, and it declines very rapidly as  $B$  increases towards  $B_{\min}(\alpha)$ . There are sharp kinks at  $B = B_{\min}(\alpha)$ . Beyond the kinks, the curve is initially almost flat, but it then falls again much more slowly than before. There are also kinks at  $B = 2B_{\min}(\alpha) + 1$ ,  $B = 3B_{\min}(\alpha) + 1$ , and so on. However, power loss becomes very small, and the curves become quite flat, as  $B$  increases.

Figure 5 is similar to Figure 6, but it shows power loss for tests at the .01 level. The loss of power when  $\delta = 1$  is quite modest because there is not much power to be lost, as can be seen from Figure 4. For  $B < B_{\min}(\alpha)$ , where there are once again noticeable kinks, the loss of power is much greater when  $\delta = 2$  and very much greater when  $\delta = 3$ , and it declines more rapidly as  $B$  increases.

## 5. CONCLUSION

We have proposed a very simple procedure for performing Monte Carlo and bootstrap tests that is valid for any positive number of simulations  $B$ . This procedure is closely related to the EDF and biased EDF methods of calculating  $P$  values that are widely used, and it yields identical results whenever  $\alpha(B + 1)$  is an integer. The new procedure produces exact tests whenever a test statistic is pivotal. For test statistics that are only asymptotically pivotal, it produces tests with an error in rejection probability that is of the same order in the sample size for all  $B$ .

APPENDIX

The fundamental result of this paper is that the continuous EDF  $P$  value  $P_B^c$  defined in (5) follows the  $U[0, 1]$  distribution under the null hypothesis, and therefore  $\Pr[P_B^c < \alpha] = \alpha$  for any  $B > 0$ .

*Proof.* Recall that  $N \equiv \sum_{j=1}^B I(\tau_j^* > \hat{\tau})$ . Thus  $N \in \{0, \dots, B\}$ , each outcome occurring with probability  $1/(B+1)$  under the null. The probability function for  $N$  under the null,  $P(N)$ , and the density function for  $U$ ,  $f_u(U)$ , are independent, having joint probability density function  $f(N, U) = P(N) \times f_u(U)$ . The expected rejection frequency is therefore

$$\begin{aligned} \Pr[P_B^c < \alpha] &= \Pr \left[ \frac{N + U}{B + 1} < \alpha \right] \\ &= \sum_{N=0}^B P(N) \int_0^1 dF_u(U < \alpha(B + 1) - N) \\ &= \frac{1}{(B + 1)} \sum_{N=0}^B \int_0^1 dF_u(U < \alpha(B + 1) - N). \end{aligned}$$

$F_u$ , being the uniform CDF, assumes values

$$F_u(\alpha) = \begin{cases} 0 & \text{if } \alpha(B + 1) < N, \\ 1 & \text{if } \alpha(B + 1) - 1 > N, \text{ and} \\ \alpha(B + 1) - N & \text{otherwise.} \end{cases}$$

Recalling that  $N$  is an integer,

$$\begin{aligned} \Pr[P_B^c < \alpha] &= \sum_{N=0}^{N < \text{ceiling}(\alpha(B+1)-1)} \frac{1}{B+1} + \sum_{N=\text{ceiling}(\alpha(B+1)-1)} \frac{\alpha(B+1) - N}{B+1} \\ &\quad + \sum_{N > \text{ceiling}(\alpha(B+1)-1)}^B \frac{0}{B+1} \\ &= \frac{\text{ceiling}(\alpha(B+1)-1)}{B+1} + \frac{\alpha(B+1) - \text{ceiling}(\alpha(B+1)-1)}{B+1} \\ &= \alpha. \end{aligned}$$

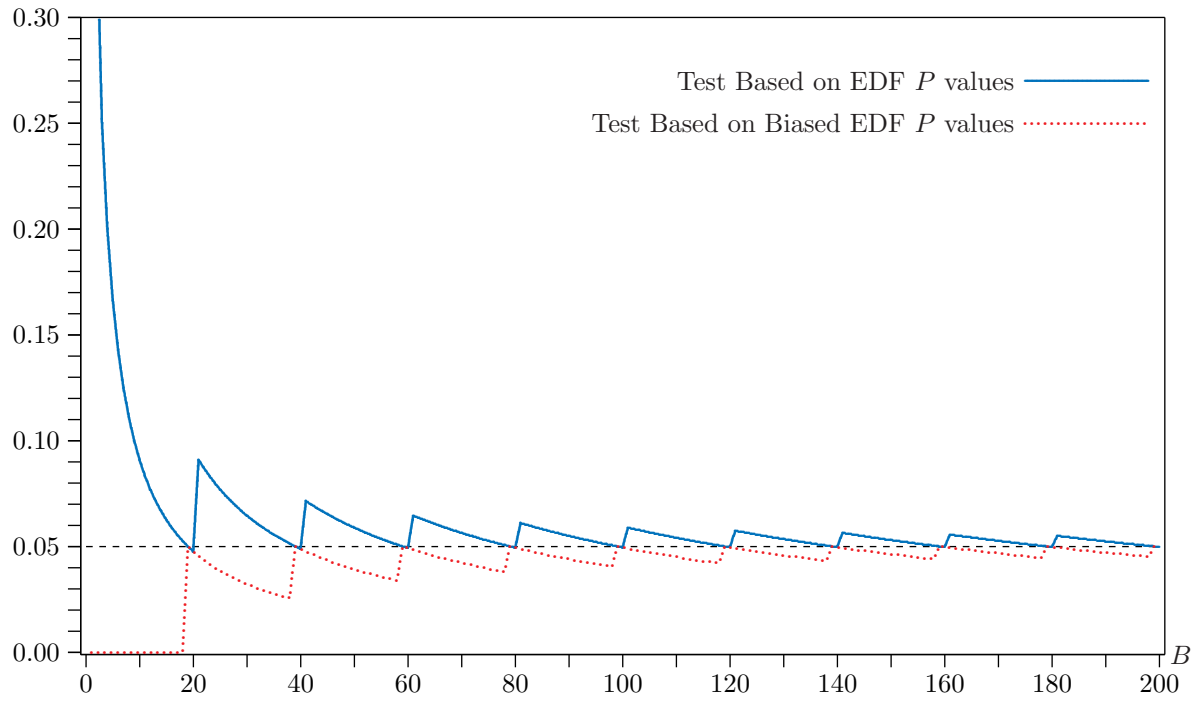
Therefore,  $\Pr[P_B^c < \alpha] = \alpha$  for any  $B > 0$ . □

## REFERENCES

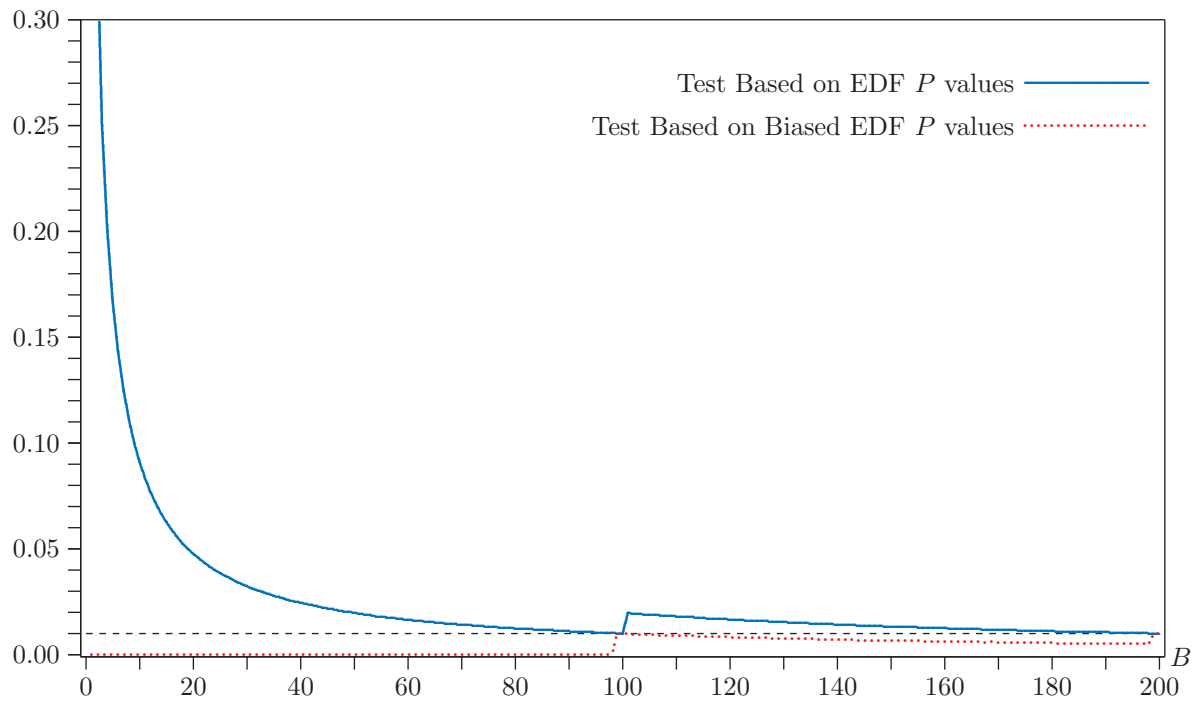
1. R. Beran, *Prepivoting test statistics: A bootstrap view of asymptotic refinements*, Journal of the American Statistical Association **83** (1988), 687–697.
2. D. R. Cox and D. V. Hinkley, *Theoretical statistics*, Chapman and Hall, London, 1974.
3. R. Davidson and J. G. MacKinnon, *Bootstrap tests: How many bootstraps?*, Econometric Reviews **19** (2000), 55–68.
4. A. C. Davison and D. V. Hinkley, *Bootstrap methods and their application*, Cambridge University Press, Cambridge, 1997.
5. M. Dwass, *Modified randomization tests for nonparametric hypotheses*, Mathematical Statistics **28** (1957), 181–187.
6. C. Gouriéroux and A. Monfort, *Simulation-based econometric methods*, Oxford University Press, Oxford, 1997.
7. P. Hall, *On the number of bootstrap simulations required to construct a confidence interval*, The Annals of Statistics **14** (1986), 1453–1462.
8. ———, *The bootstrap and edgeworth expansion*, Springer Series in Statistics, Springer-Verlag, New York, 1992.
9. P. Hall and D.M. Titterton, *The effect of simulation order on level accuracy and power of Monte Carlo tests*, Journal of the Royal Statistical Society **B 51** (1989), 459–467.
10. K.-H. Jöckel, *Finite sample properties and asymptotic efficiency of Monte Carlo tests*, Annals of Statistics **14** (1986), 336–347.
11. E. Mammen, *When does bootstrap work? Asymptotic results and simulations*, Springer-Verlag, New York, 1992.
12. H. van Dijk and A. Monfort, *Econometric inference using simulation techniques*, John Wiley and Sons, Chichester, 1995.

JEFF RACINE, DEPARTMENT OF ECONOMICS & CENTER FOR POLICY RESEARCH, SYRACUSE UNIVERSITY, SYRACUSE, NEW YORK, USA 13244-1020

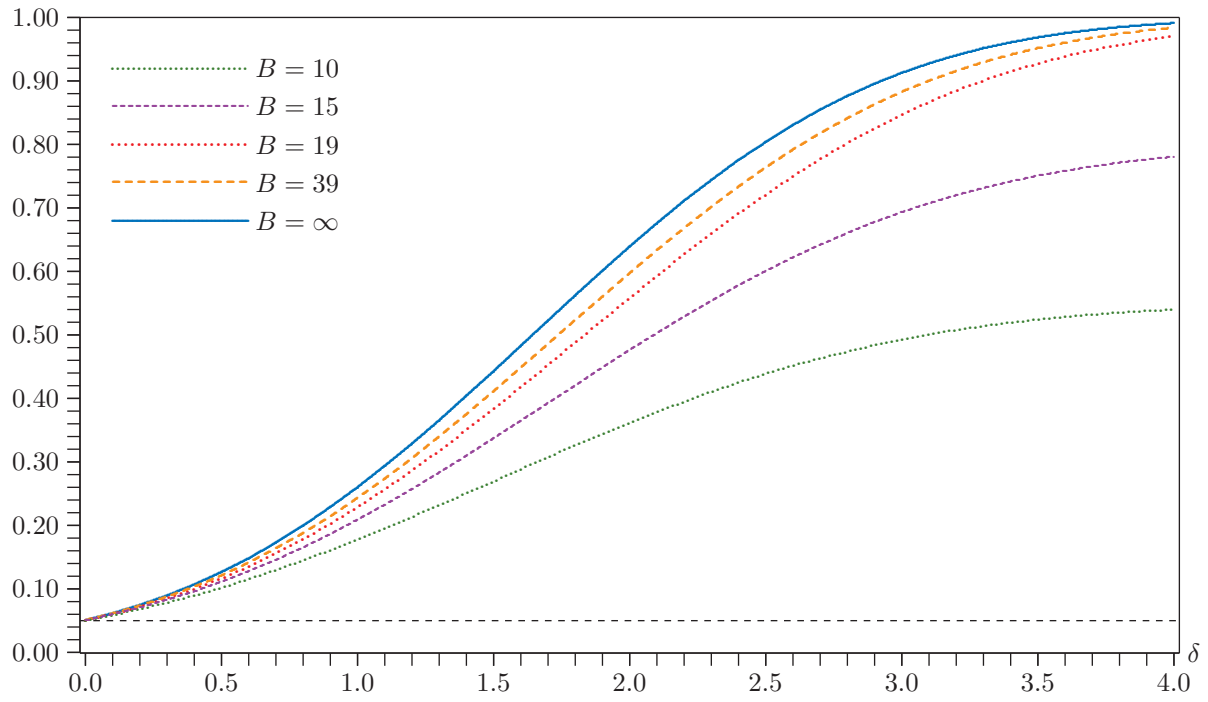
JAMES G. MACKINNON, DEPARTMENT OF ECONOMICS, QUEEN'S UNIVERSITY, KINGSTON, ONTARIO, CANADA, K7L 3N6



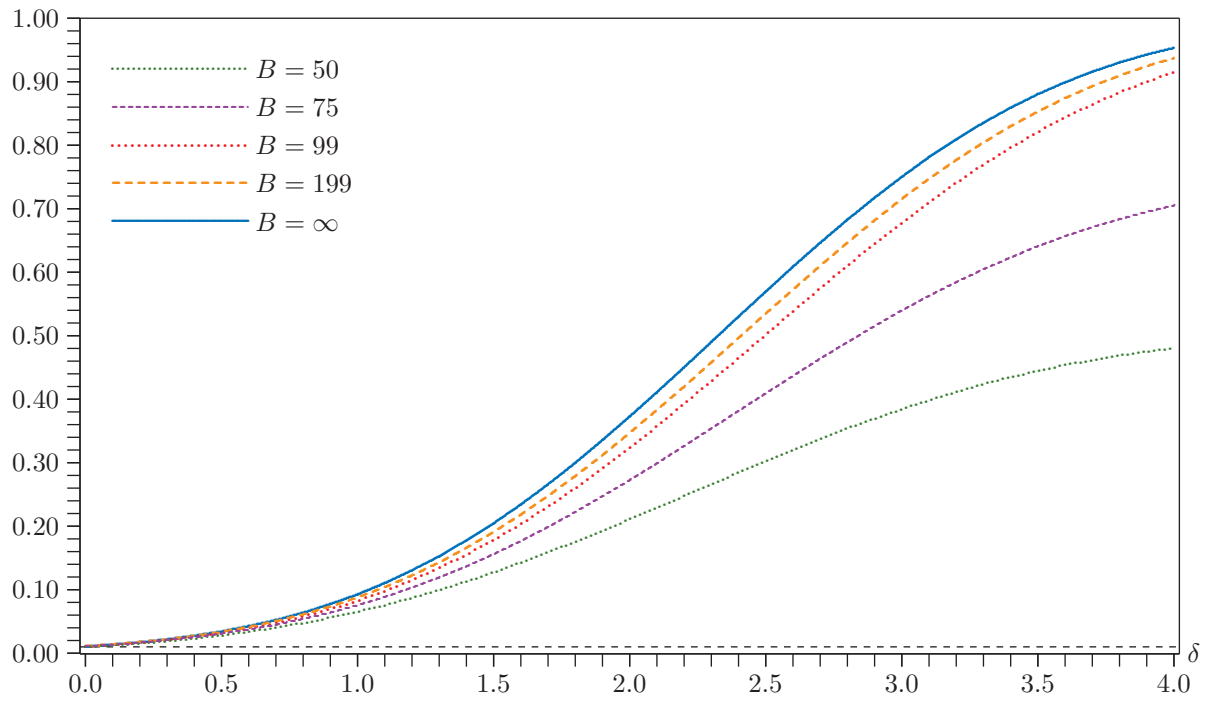
**Figure 1.** Rejection frequencies for tests at .05 level



**Figure 2.** Rejection frequencies for tests at .01 level



**Figure 3.** Power functions for tests at .05 level



**Figure 4.** Power functions for tests at .01 level

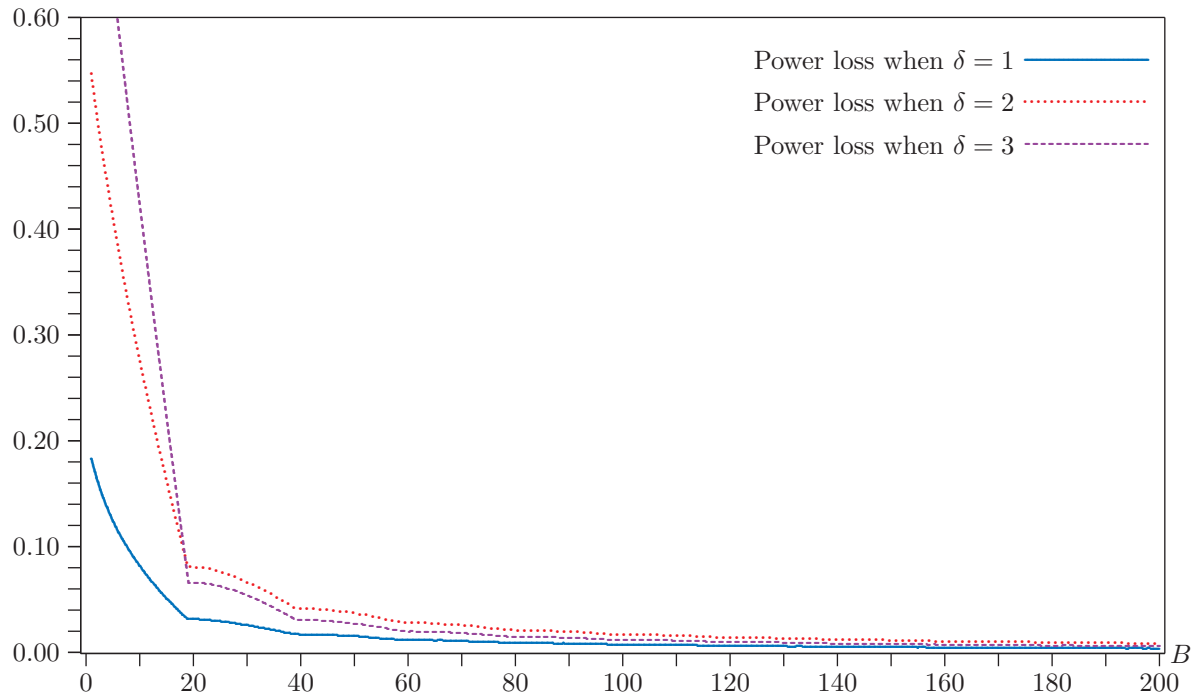


Figure 5. Power loss for tests at .05 level

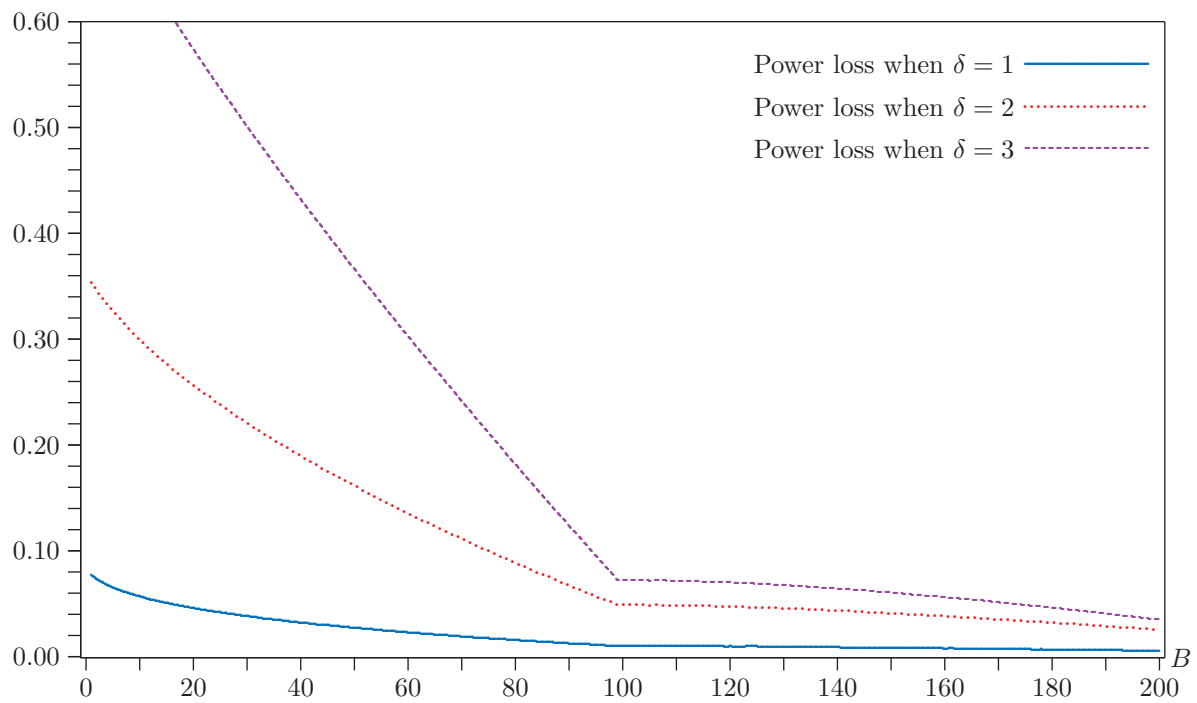


Figure 6. Power loss for tests at .01 level

