INSTITUTE FOR ECONOMIC RESEARCH

Discussion Paper # 951

Bootstrap Testing: How Many Bootstraps?

by

Russell Davidson and James G. MacKinnon

September 1997

Queen's University
Kingston, Ontario, Canada K7L 3N6

Bootstrap Tests: How Many Bootstraps?

by

Russell Davidson

GREQAM
Centre de la Vieille Charité
2 rue de la Charité
13002 Marseille, France

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

Electronic mail: **russell@ehess.cnrs-mrs.fr**

and

James G. MacKinnon

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

Electronic mail: **jgm@qed.econ.queensu.ca**

Abstract

This note discusses how to choose the number of bootstrap samples when performing bootstrap tests. There are two important issues that arise when the number of bootstraps is finite. One is bias in the estimation of bootstrap P values or critical values, and the second is loss of power. We discuss an easy way to avoid bias and thus obtain exact tests if the underlying test statistic is pivotal. We also propose a simple pretest procedure for choosing the number of bootstrap samples so as to avoid power loss, and we illustrate its performance using sampling experiments.

This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada. We are grateful to Joel Horowitz, two referees, and numerous seminar participants for comments on earlier work.

July, 1997

1. Introduction

As a result of the remarkable increases in the speed of computers that have occurred during the past two decades, the bootstrap has become increasingly popular for performing hypothesis tests. In econometrics, the use of the bootstrap for this purpose has been advocated by Horowitz (1994), Hall and Horowitz (1996), Davidson and MacKinnon (1996), and others. There are two types of error associated with this use of the bootstrap. The first is that bootstrap P values and bootstrap critical values will generally not be correct even when the number of bootstrap samples is infinite. This type of error may occur whenever a test statistic is not pivotal, so that its distribution depends on unknown parameters or other unknown features of the data generating process. Nevertheless, inferences based on the bootstrap will generally be more accurate than inferences based on asymptotic theory, in the sense that the errors are of lower order in the sample size; see Beran (1988), Hall (1992), and Davidson and MacKinnon (1996).

The second type of error, which is the subject of this note, arises because the number of bootstrap samples, say B, is necessarily finite. Using a finite number of bootstraps actually has two consequences. The first is that, unless B is chosen correctly, estimates of P values or critical values will be biased. The second is that, whenever $B < \infty$, there will be some loss of power. This loss of power is often small, but, as we shall show, it can be quite large in some cases.

There are at least two ways to use the bootstrap for hypothesis testing. One approach uses it to compute P values, and a second uses it to compute critical values. If the objective is simply to reject or not reject a null hypothesis at some predetermined level α, these two approaches are equivalent. We emphasize the P value approach because, in most circumstances, it provides more information than the critical value approach and is a little easier to analyze.

The P value approach is very simple. One first computes a test statistic, say $\hat{\tau}$, in the usual way. Using an estimate of the model under the null hypothesis, which we denote by $\hat{\mu}$, one then draws B bootstrap samples. Each of these is used to compute a bootstrap test statistic τ_j^* in exactly the same way as the real sample was used to compute $\hat{\tau}$. For a one-tailed test with a rejection region in the upper tail, the bootstrap P value may then be estimated by

$$\hat{p}^*(\hat{\tau}) \equiv \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* \geq \hat{\tau}), \tag{1}$$

where $I(\cdot)$ is the indicator function. As $B \to \infty$, it is clear that the estimated bootstrap P value $\hat{p}^*(\hat{\tau})$ will tend to the ideal bootstrap P value $p^*(\hat{\tau})$, which is defined as

$$p^*(\hat{\tau}) \equiv \Pr_{\hat{\mu}}(\tau \geq \hat{\tau}).$$

An ideal bootstrap test rejects the null hypothesis at level α whenever $p^*(\hat{\tau}) < \alpha$. A feasible bootstrap test rejects it whenever $\hat{p}^*(\hat{\tau}) < \alpha$. With the ideal test, it is

obvious that we could obtain precisely the same results by using an infinite number of bootstrap samples to estimate a level α critical value τ_C^* and rejecting the null hypothesis whenever $\hat{\tau} > \tau_C^*$. As we discuss in the next section, this is also true for the feasible bootstrap test if B is chosen properly.

In the next section, we discuss how B should be chosen to avoid bias in the estimation of $\hat{p}$. Then, in Section 3, we discuss the power loss that can arise from bootstrapping with finite B. Finally, in Section 4, we discuss how B should be chosen, and we propose a simple way to choose it endogenously by using pretests.

2. How to Choose the Number of Bootstraps

The issues discussed in this section and the next are closely related to the literature on Monte Carlo testing. The idea of a Monte Carlo test is generally attributed to Dwass (1957) and Barnard (1963). Other early papers include Hope (1968) and Marriott (1979). Suppose that a test statistic τ is known to be pivotal but does not have a distribution that is readily computable. For a test at level α, we calculate B artificial samples, which we may still refer to as bootstrap samples, and from them we compute B test statistics, τ_j^*, $j = 1, \ldots, B$. It is essential to choose B so that $\alpha(B+1)$ is an integer. We sort all $B+1$ test statistics, $\hat{\tau}$ and the B bootstrap ones, so that the test statistic which would lead to the most decisive rejection of the null hypothesis is ranked first. Thus, if we want to reject when $\hat{\tau}$ is large, we would sort the test statistics from largest to smallest. We then see where $\hat{\tau}$ lies in the sorted list. If the rank of $\hat{\tau}$ is less than or equal to $\alpha(B+1)$, we reject the null hypothesis at level α; otherwise, we do not reject it. For example, if $\alpha = .05$ and $B = 999$, we would reject the null whenever the rank of $\hat{\tau}$ is no greater than 50.

It is easy to see why $\alpha(B+1)$ must be an integer. The rank of $\hat{\tau}$ can have $B+1$ possible values, all of them equally likely under the null hypothesis. The number of these that lead to rejection of the null is the largest integer not greater than $\alpha(B+1)$. Thus if, but only if, $\alpha(B+1)$ is an integer, the null will be rejected in exactly $\alpha(B+1)$ of the $B+1$ possible cases, that is, with probability equal to α, so that the Monte Carlo test is exact. If $\alpha(B+1)$ were not an integer, this simple argument would not go through. For example, if B were 100 and α were .05, $\hat{\tau}$ could have 101 possible ranks. If we rejected when its rank was less than or equal to $\alpha(B+1) = 5.05$, which means less than or equal to 5, the probability of rejecting would be $5/101 = .0495$. Notice that, in this case, we would obtain an estimated P value precisely equal to .05 in one sample out of 101. If that happened, the null hypothesis should not be rejected. Similarly, if B were 98, $\hat{\tau}$ could have 99 possible ranks, of which, since $\alpha(B+1) = 4.95$, only 4 would lead to rejection. Thus the probability of rejecting would be $4/99 = 0.0404$. Observe that the consequences of making B slightly too small in this example are much more severe than the consequences of making it slightly too large.

We have just seen that $\alpha(B+1)$ must be an integer if we wish to obtain an exact test. Therefore, for $\alpha = .05$, the smallest possible value of B is 19, and for

$\alpha = .01$, the smallest possible value is 99. However, unless computing costs are extraordinarily high, we almost certainly will not want to use these smallest possible values in practice. There are two reasons for this. Firstly, even though using a finite value for B does not affect the level of a Monte Carlo test, it does affect the outcome of the test. Whether or not we reject will depend on the output of the random number generator we use as well as on the data. Secondly, as we shall discuss in the next section, using a finite value for B reduces the power of the test.

It is easy to see that Monte Carlo tests must yield exactly the same outcomes as the two types of bootstrap test discussed in the introduction, provided that B is chosen correctly. The feasible bootstrap P value defined in (1) will be less than α if and only if the rank of $\hat{\tau}$ is less than or equal to $\alpha(B+1)$. For example, suppose that $B = 99$ and $\alpha = .05$. Then, if the rank of $\hat{\tau}$ is 5, $\hat{p}^* = 4/99 = .0404$, and if the rank of $\hat{\tau}$ is 6, $\hat{p}^* = 5/99 = .0505$. In the former case, we reject the null hypothesis, and in the latter case, we do not. Thus a test based on bootstrap P values will yield exactly the same results as a Monte Carlo test.

An alternative approach to bootstrap testing is to use the τ_j^* to estimate a critical value; see Horowitz (1994). When B is chosen so that $\alpha(B+1)$ is an integer, the natural estimator of the critical value is $\hat{\tau}_C^* \equiv \tau_{\alpha(B+1)}^*$, where $\tau_{\alpha(B+1)}^*$ denotes the τ_j^* with rank $\alpha(B+1)$ when the τ_j^*, without $\hat{\tau}$, are ranked from largest to smallest. It is easy to see that the rank of $\hat{\tau}$ in the ranking including $\hat{\tau}$ is less than or equal to $\alpha(B+1)$ if and only if $\hat{\tau} > \hat{\tau}_C^*$. Thus rejecting the null when $\hat{\tau} > \hat{\tau}_C^*$ is equivalent to using a Monte Carlo test, and therefore equivalent to rejecting when $\hat{p}^* < \alpha$.

The above discussion has explicitly assumed that τ is pivotal. In many situations in which bootstrap tests are used, this assumption will be false. As a consequence, even ideal bootstrap tests will not be exact. Thus the argument for choosing B so that $\alpha(B+1)$ is an integer may not be quite as strong, in practice, as the above discussion suggests. However, there does not appear to be an argument for choosing it in any other way.

3. Sampling Variability and Power Loss

Using a finite value of B inevitably results in some loss of power, for essentially the same reason that, in the context of bootstrap confidence interval estimation, it results in confidence intervals that tend to be too long (Hall, 1986). This power loss was first investigated for a rather special case by Hope (1968). Subsequently, Jöckel (1986) obtained some fundamental theoretical results for a fairly wide class of Monte Carlo tests. These results are immediately applicable to bootstrap tests.

For any test and any fixed alternative, we can define the size-power function $\eta(\alpha)$ as the probability that the test will reject the null when its true size is α. This function implicitly depends on the DGP, but for notational simplicity this dependence is not made explicit. This function is precisely what we estimate when we plot a size-power curve using simulation results; see, for example, Davidson and MacKinnon (1998). Because $\eta(0) = 0$ and $\eta(1) = 1$, and we need $\eta(\alpha) > \alpha$ for $0 < \alpha < 1$ if the

test is to have power greater than its size, in general we may expect the size-power function to be concave. For tests that follow standard noncentral distributions, such as the noncentral χ^2 and the noncentral F, the size-power function is indeed concave. However, it will not necessarily be concave for every test, regardless of what DGP generated the data.

Let the size-power function for a bootstrap test based on B bootstrap samples be denoted by $\eta^*(\alpha, m)$, where $\alpha(B+1) = m$. When the original test statistic is assumed to be pivotal, $\eta(\alpha) \equiv \eta^*(\alpha, \infty)$. Under this condition, Jöckel (1986) proves the following two results:

(i) If $\eta(\alpha)$ is concave, then so is $\eta^*(\alpha, m)$.

(ii) Assuming that $\eta(\alpha)$ is concave, $\eta^*(\alpha, m+1) > \eta^*(\alpha, m)$.

Thus increasing the number of bootstrap samples will always increase the power of the test. Just how much it will do so depends in a fairly complicated way on the shape of the size-power function $\eta(\alpha)$, and Jöckel's theoretical results on this point are not at all easy to interpret.

The assumption of pivotalness is not needed if we wish to compare the power of an ideal bootstrap test with the power of a feasible bootstrap test that corresponds to it. We simply have to interpret $\eta(\alpha)$ as the size-power function for the ideal bootstrap test. Provided this function is concave, Jöckel's two results apply. Thus we conclude that the feasible bootstrap test will be less powerful than the ideal bootstrap test whenever the size-power function for the ideal test is concave.

To see just how the choice of B affects test power, we conducted a small simulation experiment. We generated t statistics for the null hypothesis that $\gamma = 0$ in the very simple model

$$y_t = \gamma + u_t, \quad u_t \sim N(0,1), \quad t = 1, \ldots, 4. \tag{2}$$

These t statistics were then converted to P values. For the ideal bootstrap test, this was done by using the c.d.f. of the $t(3)$ distribution so as to obtain p^*. For feasible bootstrap tests corresponding to different values of B, this was done by drawing bootstrap test statistics from the $t(3)$ distribution and using equation (1), slightly modified to allow for the fact that this is a two-tailed test, to obtain $\hat{p}^*$. There were 400,000 replications.

From Jöckel's results, we know that it is only the size-power function that affects power loss. The details of (2) (in particular, the very small sample size) were chosen solely to keep the cost of the experiments down. They affect the results only to the extent that they affect the shapes of the size-power curves. Figure 1 shows these curves for four values of γ, namely, 0.5, 1.0, 2.0, and 3.0. Curves for $\gamma = 1.5$ and $\gamma = 2.5$ were also computed, but they are not shown to avoid cluttering the figure.

The values of B that were used in our experiments were 19, 24, 39, 49, 79, 99, 199, and 399. These values were chosen because they yield valid Monte Carlo tests for commonly encountered values of α. The smallest value of B that yields a valid Monte Carlo test for $\alpha = .05$ is 19, for $\alpha = .04$, 24, for $\alpha = .025$, 39, and

so on. If B were not chosen so that Monte Carlo tests are valid for certain values of α, the estimated bootstrap P value $\hat{p}^*(\hat{\tau})$ would provide a biased estimate of the true bootstrap P value $p^*(\hat{\tau})$ when the latter is equal to those values of α. As a consequence, estimates of test size and power in a simulation experiment would not vary smoothly with α. In order to eliminate this type of bias, the size-power curves for our experiments were plotted only for values of α such that $\alpha(B+1)$ was an integer. Because the distribution of the ideal bootstrap test is known for the simple case we are examining, and because B is chosen so that test size is exact for all values of α that are plotted, we can simply plot the observed power of each test against its nominal size without having to correct for any possible size distortion.

Figures 2, 3, and 4 present the results of three of the six experiments that we performed. The horizontal axis shows test size, and the vertical axis shows the difference between the power of the test based on $B = \infty$ and the power of the test based on any finite value of B. For $B \leq 79$, the points at which power loss is evaluated are shown as balls.

In Figure 2, $\gamma = 1.0$. This value of γ implies that the ideal bootstrap test is not very powerful; it has power 0.290 for a test at the .05 level. The power loss is also reasonably small, although by no means negligible for the smaller values of B. It peaks for test sizes around .10 and then declines slowly as test size increases. The fact that the power loss is relatively low is clearly a consequence of the fact that power is low. In the unreported results for $\gamma = 0.5$, where the test had very little power, the maximum power loss was substantially less than it is in Figure 2.

In Figure 3, $\gamma = 2.0$, which implies that the ideal bootstrap test is reasonably powerful; it has power 0.755 for a test at the .05 level. In this case, power loss can be quite substantial, especially for very small test sizes. For $B = 19$, the loss is nearly .15 at the .05 level, and it would clearly be even larger than this at smaller test sizes if we could perform valid tests using such a small value of B. Even for $B = 399$, power loss is greater than .01 for all test sizes less than .05.

In Figure 4, $\gamma = 3.0$, which implies that the ideal bootstrap test is very powerful; it has power 0.967 for a test at the .05 level. The power loss for very small test sizes is much larger than in Figure 3, but it declines more rapidly as test size increases. For a test at the .01 level, the power loss when $B = 99$ is remarkably large at 0.121: The test rejects only 49.9% of the time, compared with 62.0% for a test with $B = \infty$. Even when $B = 399$, the power loss for a test at the .01 level is a hefty 0.037.

The results shown in Figures 2, 3, and 4, along with unreported results for $\gamma = 0.5$, $\gamma = 1.5$, and $\gamma = 2.5$, make it clear that the power loss from bootstrapping depends in a complicated way on the shape of the size-power function. Power loss tends to be small when the size-power curve is close to the 45° line, simply because the ideal bootstrap test has little power to lose. Power loss also tends to be small when power is extremely high, presumably because the test has power to spare. On the other hand, as can be seen from Figures 3 and 4, power loss can be quite severe for tests with small sizes that are reasonably, but not extremely, powerful.

In view of these results, it would seem advisable to use a fairly large number of bootstrap samples, unless computational cost is a severe problem. The easiest approach is simply to pick B in advance. When this is done, we recommend that B be at least 399 for a test at the .05 level and at least 1499 for a test at the .01 level, in order to avoid severe loss of power. If computational cost is not a concern, considerably larger values of B should be used: 1999 and 4999 may be reasonable choices for tests at the .05 and .01 levels, respectively.

4. Choosing B by Pretesting

Although it is the easiest approach, using a fixed number of bootstrap samples is not the best one. In many cases, it will be possible to obtain conclusive results without using very many bootstrap samples. For example, if $\hat{\tau} > \tau_j^*$ for every bootstrap sample, we will not need a very large number of bootstrap samples to conclude that we should reject the null hypothesis. The probability of this event occurring by chance if the ideal bootstrap P value, $p^*(\hat{\tau})$, is equal to α is $(1-\alpha)^B$. For $B = 99$ and $\alpha = .05$, this probability is .006. Thus, if 99 bootstrap samples yield none for which $\tau_j^* > \hat{\tau}$, it is probably safe to reject the null hypothesis at the .05 level. This is one case in which the critical value approach may be attractive. If the actual values of the τ_j^* are all much lower than $\hat{\tau}$, we may be able to estimate τ_C^* and reject the hypothesis that $\tau_C^* > \hat{\tau}$ at a high level of confidence using 39 or even 19 bootstrap samples.

Similarly, if $\hat{p}^*(\hat{\tau})$ is substantially greater than α, we may be able to conclude that $p^*(\hat{\tau}) > \alpha$. According to the binomial distribution, the probability that $\tau_j^* > \hat{\tau}$ 11 or more times out of 99 if $p^*(\hat{\tau}) = .05$ is .004. Thus, if 11 or more out of 99 bootstrap samples produce test statistics more extreme than $\hat{\tau}$, it is probably safe to conclude that the null hypothesis cannot be rejected at the .05 level.

The examples just given suggest a simple pretesting procedure for choosing B endogenously. We do not claim that this procedure is optimal in any sense. However, it is easy to implement and, as we shall demonstrate shortly, it appears to work very well. The procedure involves three steps:

1. Choose $B_{\min}$, the initial number of bootstrap samples (say, 99), $B_{\max}$, the maximum number of bootstrap samples (say, 12,799) and β, the level for the pretest (say, .001). Initially, calculate τ_j^* for $B_{\min}$ bootstrap samples, and set $B = B_{\min}$ and $B' = B_{\min}$.
2. Compute $\hat{p}^*(\hat{\tau})$ based on B bootstrap samples. Depending on whether $\hat{p}^*(\hat{\tau}) < \alpha$ or $\hat{p}^*(\hat{\tau}) > \alpha$, test either the hypothesis that $p^*(\hat{\tau}) \geq \alpha$ or the hypothesis that $p^*(\hat{\tau}) \leq \alpha$ at level β. This may be done using the binomial distribution or, if αB is not too small, the normal approximation to it. If $\hat{p}^*(\hat{\tau}) < \alpha$ and the hypothesis that $p^*(\hat{\tau}) \geq \alpha$ is rejected, or if $\hat{p}^*(\hat{\tau}) > \alpha$ and the hypothesis that $p^*(\hat{\tau}) \leq \alpha$ is rejected, stop.

3. If the algorithm gets to this step, set $B = 2B'+1$. If $B > B_{\max}$, stop. Otherwise, calculate τ_j^* for a further $B'+1$ bootstrap samples and set $B' = B$. Then return to step 2.

It is easy to see how this procedure will work. When $p^*(\hat{\tau})$ is not close to α, it will usually terminate after one or two rounds with an estimate $\hat{p}^*(\hat{\tau})$ that is relatively inaccurate, but clearly different from α. When $p^*(\hat{\tau})$ is reasonably close to α, the procedure will usually terminate after several rounds with an estimate $\hat{p}^*(\hat{\tau})$ that is fairly accurate. When $p^*(\hat{\tau})$ is very close to α, it will usually terminate with $B = B_{\max}$ and a very accurate estimate $\hat{p}^*(\hat{\tau})$. Occasionally, especially in this last case, the procedure will make a mistake, in the sense that $\hat{p}^*(\hat{\tau}) < \alpha$ when $p^*(\hat{\tau}) > \alpha$, or *vice versa*. However, the magnitude of the difference between $\hat{p}^*(\hat{\tau})$ and $p^*(\hat{\tau})$ will usually be very small when this happens. The probability of such a mistake can be reduced by making β smaller or by making $B_{\max}$ larger.

In order to investigate how this procedure works in practice, we conducted several simulation experiments, with 1,000,000 replications each, based on the model (2), with different values of γ, $B_{\min}$, $B_{\max}$, and β. The same sequence of random numbers was used to generate the data for all values of these parameters. $B_{\min}$ was normally 99, and α was always .05. Because it is extremely expensive to evaluate the binomial distribution directly when B is large, we used the normal approximation to the binomial whenever $\alpha B \geq 10$.

Table 1 shows results for four different values of γ. When $\gamma = 0$, so that the null hypothesis is true, B^*, the average number of bootstrap samples, is quite small. As expected, reducing β and increasing $B_{\max}$ both cause B^* to increase. As can be seen from the column headed "Conflicts", there are very few cases in which the feasible bootstrap test yields a different result from the test with $B = \infty$. Most conflicts occur when the procedure terminates with $B^* = B_{\max}$, which implies that $p^*(\hat{\tau})$ is very near α and is estimated very accurately; the fraction of the cases with conflicts for which $B^* = B_{\max}$ is shown in the column headed "Prop. $B_{\max}$". Thus, even when the procedure yields the "wrong" answer, the investigator is not likely to be seriously misled.

The average number of bootstrap samples is higher for $\gamma = 3$ than for $\gamma = 0$, higher again for $\gamma = 1$, and higher still for $\gamma = 2$. This reflects the way in which the proportion of the $p^*(\hat{\tau})$ near .05 depends on γ. When B^* is higher, there tend to be more cases in which the feasible and ideal bootstrap tests yield different results, because the procedure terminates more frequently with $B = B_{\max}$. Some power loss is evident, but it is always very small, less than .001 in the worst case. All of the choices of β and $B_{\max}$ that were investigated appear to yield acceptable results. We very tentatively recommend $\beta = .001$ and $B_{\max} = 12,799$.

The last line in the table for each value of γ shows what happens when we choose a fixed B slightly larger than the B^* observed for the recommended values of β and $B_{\max}$. There are far more conflicts when a fixed B is used, and they are more likely to be misleading, because they are based on much less accurate estimates of $p^*(\hat{\tau})$. There is also substantially more power loss. Thus it appears that, holding

expected computer time constant, the pretesting procedure works considerably better than using a fixed value of B.

It is easy to understand why the pretesting procedure works well. From the results of Section 3, we know that, when the null hypothesis is true, B can safely be small, because we are not concerned about power at all. Similarly, when the null is false and test power is extremely high, B does not need to be large, because power loss is not a serious issue. However, when the null is false and test power is moderately high, B needs to be large in order to avoid loss of power. Of course, since one does not know whether or not the null is false when one undertakes a test, one cannot choose B in advance on the basis of these considerations. But that is more or less what the pretesting procedure does. It tends to make B small when it can safely be small and large when it needs to be large.

5. Final Remarks

An unavoidable feature of bootstrap testing is the need to choose the number of bootstrap samples, B. In Section 2, we explained why B should always be chosen so that $B+1$ times the level of the test is an integer. In Section 3, we discussed the loss of power that can occur when B is too small. The easiest way to avoid this is simply to choose B as a fairly large integer, larger for tests at the .01 level than for tests at the .05 level. However, the pretesting procedure that we discussed in Section 4 seems to work substantially better than using a fixed number of bootstrap samples.

References

Barnard, G. A. (1963). "Contribution to discussion," *Journal of the Royal Statistical Society*, Series B, 25, 294.

Beran, R. (1988). 'Prepivoting test statistics: a bootstrap view of asymptotic refinements," *Journal of the American Statistical Association*, 83, 687–697.

Davidson, R. and J. G. MacKinnon (1996). "The size distortion of bootstrap tests," GREQAM Document de Travail No. 96A15.

Davidson, R. and J. G. MacKinnon (1998). 'Graphical methods for investigating the size and power of hypothesis tests," *The Manchester School*, forthcoming.

Dwass, M. (1957). "Modified randomization tests for nonparametric hypotheses," *Annals of Mathematical Statistics*, 28, 181–187.

Hall, P. (1986). "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *Annals of Statistics*, 14, 1453–1462.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York, Springer-Verlag.

Hall, P. and J. L. Horowitz (1996). "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrica*, 64, 891–916.

Hope, A. C. A. (1968). "A simplified Monte Carlo significance test procedure," *Journal of the Royal Statistical Society*, Series B, 30, 582–598.

Horowitz, J. L. (1994). "Bootstrap-based critical values for the information matrix test," *Journal of Econometrics*, 61, 395–411.

Jöckel, K.-H. (1986). "Finite sample properties and asymptotic efficiency of Monte Carlo tests," *Annals of Statistics*, 14, 336–347.

Marriott, F. H. C. (1979). "Barnard's Monte Carlo tests: How many simulations?," *Applied Statistics*, 28, 75–77.

Table 1. Bootstrap Tests with B Chosen by Pretest

γ	$B_{\min}$	$B_{\max}$	β	B^*	Rej. B^∞	Rej. B^*	Conflicts	Prop. $B_{\max}$
0.0	99	12,799	0.01	324.4	0.04986	0.04983	0.00170	0.8019
0.0	99	6,399	0.001	318.3	0.04986	0.04980	0.00217	0.9866
0.0	99	12,799	0.001	420.1	0.04986	0.04985	0.00153	0.9824
0.0	99	12,799	0.0001	491.6	0.04986	0.04983	0.00150	0.9987
0.0	439	439		439.0	0.04986	0.04985	0.00827	1.0000
1.0	99	12,799	0.01	1056.9	0.28855	0.28858	0.00727	0.8077
1.0	99	6,399	0.001	1033.5	0.28855	0.28825	0.00919	0.9879
1.0	99	12,799	0.001	1474.3	0.28855	0.28831	0.00663	0.9804
1.0	99	12,799	0.0001	1771.0	0.28855	0.28852	0.00649	0.9985
1.0	1,499	1,499		1499.0	0.28855	0.28781	0.01910	1.0000
2.0	99	12,799	0.01	1370.6	0.75504	0.75474	0.00953	0.8098
2.0	99	6,399	0.001	1404.3	0.75504	0.75429	0.01199	0.9882
2.0	99	12,799	0.001	1977.8	0.75504	0.75462	0.00847	0.9849
2.0	99	12,799	0.0001	2409.1	0.75504	0.75475	0.00842	0.9981
2.0	1,999	1,999		1999.0	0.75504	0.75287	0.02114	1.0000
3.0	99	12,799	0.01	566.6	0.96722	0.96700	0.00290	0.8046
3.0	99	6,399	0.001	705.1	0.96722	0.96686	0.00365	0.9866
3.0	99	12,799	0.001	885.0	0.96722	0.96699	0.00261	0.9820
3.0	99	12,799	0.0001	1120.1	0.96722	0.96703	0.00262	0.9977
3.0	899	899		899.0	0.96722	0.96469	0.00994	1.0000

Notes:

B^* is the average value of B that was finally chosen.

"Rej. B^∞" is the proportion of replications for which the null hypothesis was rejected at the .05 level according to the Student's t distribution.

'Rej. B^*" is the proportion of replications for which the null hypothesis was rejected at the .05 level according to the bootstrap test, based on whatever value of B was finally used.

"Conflicts" is the proportion of replications for which the ideal bootstrap test and the feasible bootstrap test yielded different inferences.

"Prop. $B_{\max}$" is the proportion, within those replications for which a conflict occurred, where the final value of B was $B_{\max}$.

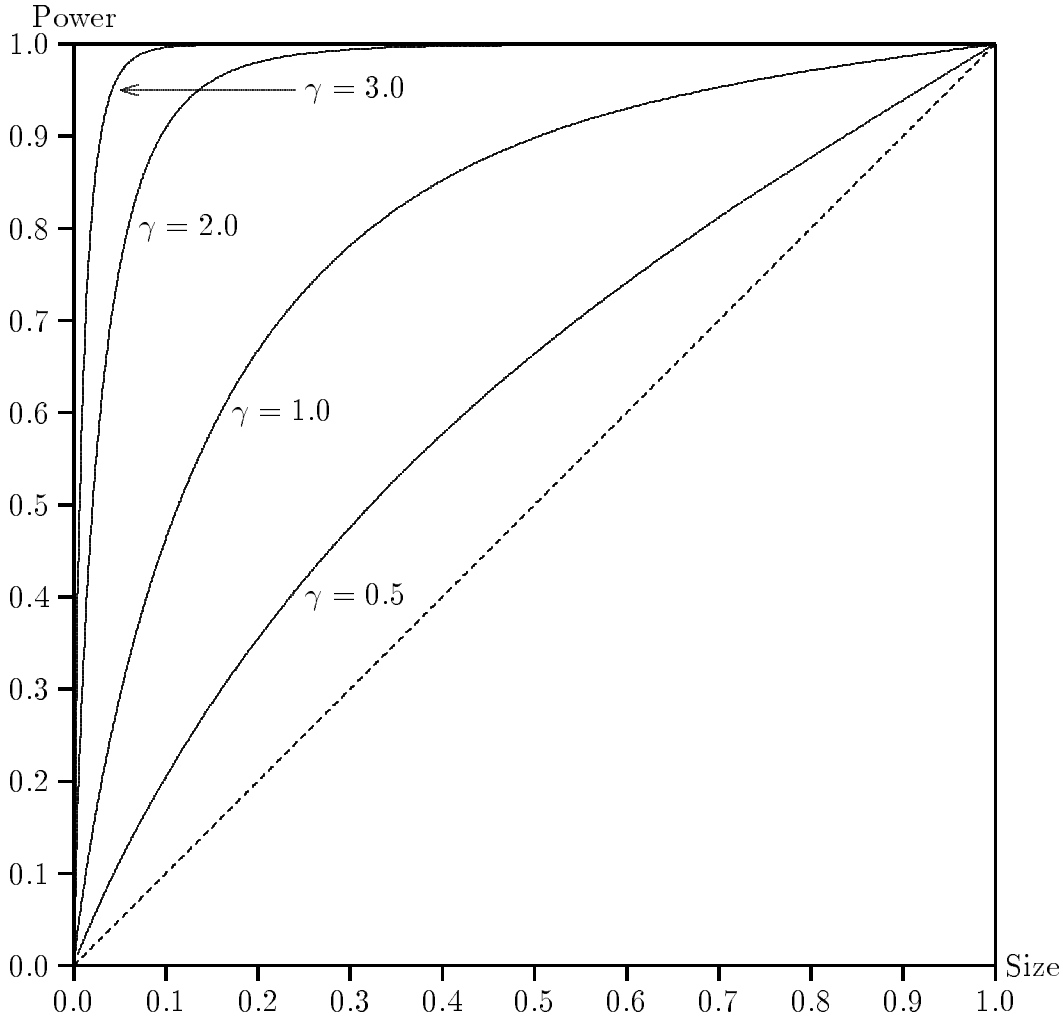

Figure 1. Size-Power Curves for $B = \infty$.

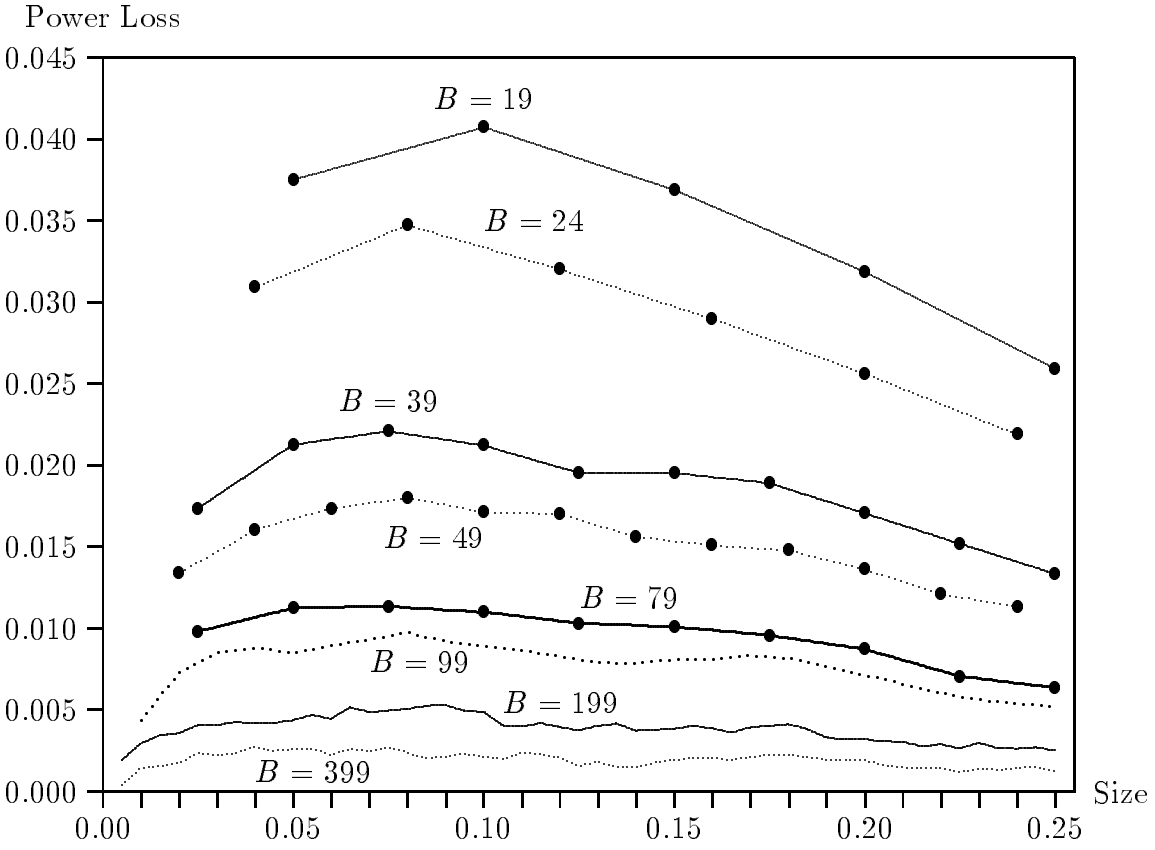

Figure 2. Power Loss from Bootstrapping: $\gamma = 1.0$.

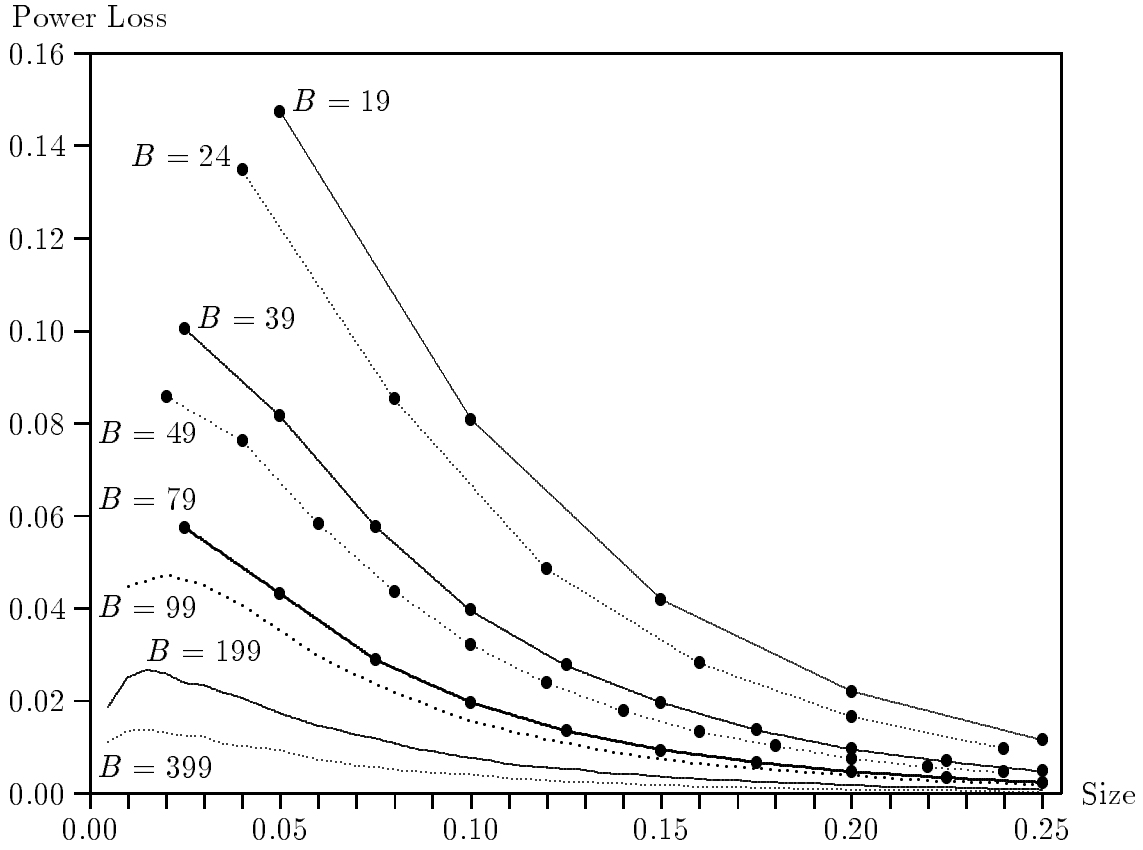

Figure 3. Power Loss from Bootstrapping: $\gamma = 2.0$.

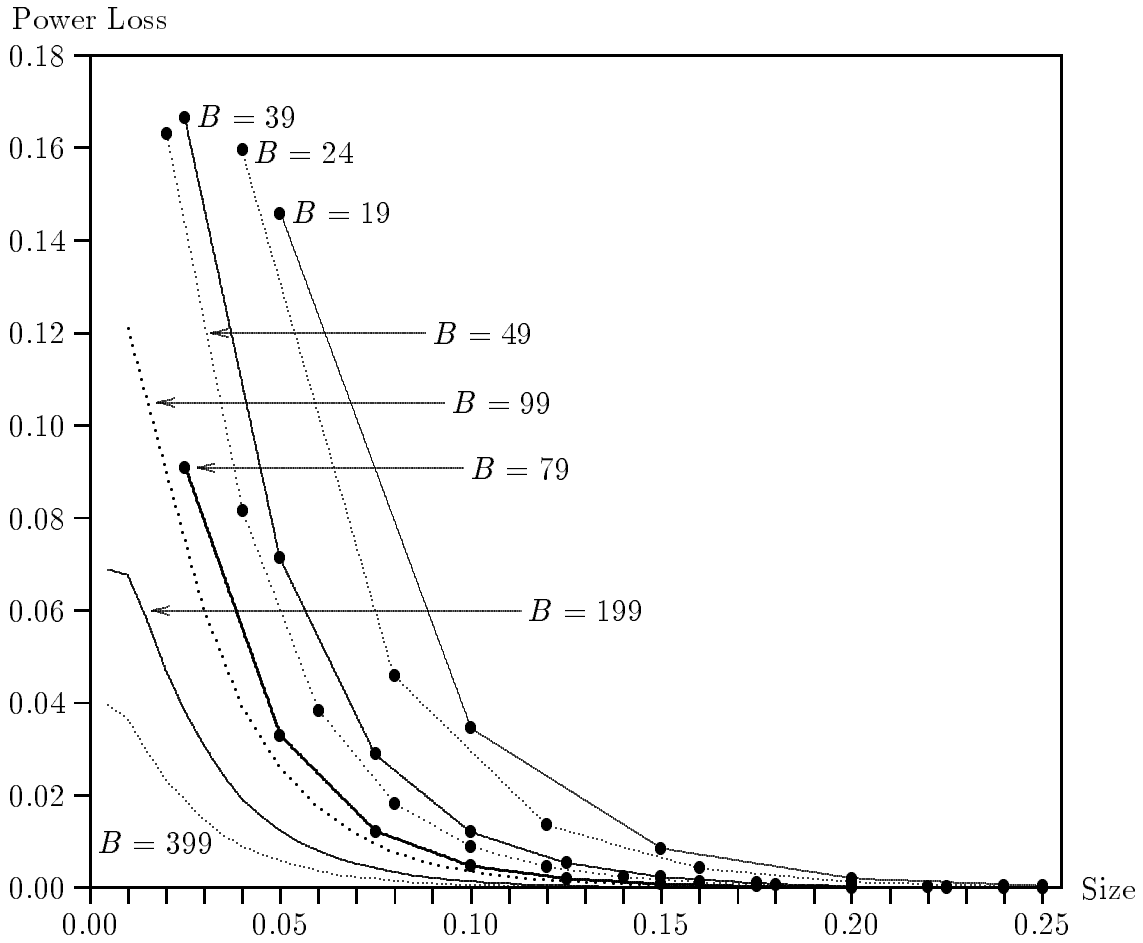

Figure 4. Power Loss from Bootstrapping: $\gamma = 3.0$.