Empirically relevant critical values for hypothesis tests: A bootstrap approach

Joel L. Horowitz*, N.E. Savin

Department of Economics, University of Iowa, 108 Pappajohn Business Bldg., Iowa City IA 52242, USA

Abstract

Tests of statistical hypotheses can be based on either of two critical values: the Type I critical value or the size-corrected critical value. The former usually depends on unknown population parameters and cannot be evaluated exactly in applications, but it can often be estimated very accurately by using the bootstrap. The latter does not depend on unknown population parameters but is likely to yield a test with low power. The critical values used in most Monte Carlo studies of the powers of tests are neither Type I nor size-corrected. They are irrelevant to empirical research. © 2000 Elsevier Science S.A. All rights reserved.

JEL classification: C12; C15

Keywords: Hypothesis test; Critical value; Size; Type I error; Bootstrap

1. Introduction

A key objective in classical testing of statistical hypotheses is achieving good power while controlling the probability that the test makes a Type I error, that is, the probability that the test rejects a correct null hypothesis, \( H_0 \). If \( H_0 \) is

*Corresponding author. Tel.: + 1-319-335-0844; fax: + 1-319-335-1956.
E-mail address: joel-horowitz@uiowa.edu (J.L. Horowitz)
simple, then controlling the probability that the test makes a Type I error is not
difficult. A simple $H_0$ completely specifies the data generation process (DGP).
Therefore, the finite-sample distribution of a test statistic under a simple $H_0$ can
be calculated, as can the probability that the test makes a Type I error for any
critical value. Similarly, it is possible to calculate the critical value correspond-
ing to any probability of a Type I error.

Most hypotheses in econometrics are composite, meaning that $H_0$ does not
completely specify the DGP. It specifies only that the DGP belongs to a given
set. As a consequence, the sampling distribution of the test statistic under $H_0$ is
unknown except in special cases because it depends on where the true DGP is in
the set specified by $H_0$. One way of solving this problem is through the concept
of the size of the test. The size is the supremum of the test’s rejection probability
over all DGPs contained in $H_0$. The $z$-level size-corrected critical value is the
critical value that makes the size equal to $z$. In principle, the exact size-corrected
critical value can be calculated if it exists, but this is rarely done in applications,
possibly because doing so typically entails very difficult computations.

A second approach to dealing with a composite $H_0$ is to base the test on
an estimator of the Type I critical value. This is the critical value that would
be obtained if the exact finite-sample distribution of the test statistic under the
true DGP were known. In general, the true Type I critical value is unknown
because the exact finite-sample distribution of the test statistic depends on
population parameters that are not specified by $H_0$. However, an approxima-
tion to the Type I critical value often can be obtained by using the (first-order)
asymptotic distribution of the test statistic under $H_0$ to approximate its finite-
sample distribution. This approximation is useful because most test statistics in
econometrics are asymptotically pivotal: their asymptotic distributions do not
depend on unknown population parameters when the hypothesis being tested is
true. Thus, an approximate Type I critical value can be obtained from asympto-
totic distribution theory without knowledge of where the true DGP is in the set
specified by $H_0$.

The main disadvantage of this approach is that first-order asymptotic approxima-
tions can be very inaccurate with samples of the sizes encountered in
applications. As a result, the true and nominal probabilities that a test rejects
a correct $H_0$ can be very different when the critical value is obtained from the
asymptotic distribution of the test statistic. We argue in Section 4, however, that
the bootstrap can often be used to obtain a good estimator of the true Type
I critical value. When the bootstrap does this, the resulting test based on an
estimated Type I critical value has greater power than a test based on an exact
size-corrected critical value (if one exists). We argue later in this paper that in
most testing situations, the best way to achieve high power while maintaining
control over the probability of a Type I error is to base the test on a good
estimate of the Type I critical value. Such an estimate is often provided by the
bootstrap.
In econometrics, Monte Carlo studies of the finite-sample powers of tests usually take a different approach to obtaining critical values. Most studies report powers based on critical values that are called size-corrected but, as is explained later in this paper, are really critical values for essentially arbitrary simple null hypotheses. They have no empirical analogs and cannot be used in applications. In other words, the size-corrected critical values usually obtained in Monte Carlo studies of power are both misnamed and irrelevant to empirical research. We propose an alternative method for obtaining critical values in power studies. This method can be implemented in applications by using the bootstrap. Thus it provides a way to obtain empirically relevant critical values in Monte Carlo studies of power and in applications.

Possibly because of the empirical irrelevance of most size-corrected critical values reported in the econometrics literature, applied research relies heavily on asymptotic distribution theory to obtain critical values for tests. Reliance on asymptotic distribution theory can blur the distinction between size-corrected and Type I critical values. We emphasize the distinction here so as to avoid possible confusion later in this paper. The asymptotic null distribution of an asymptotically pivotal test statistic does not depend on unknown population parameters, so it may appear that the asymptotic critical value approximates the size-corrected critical value. This appearance is misleading except in special cases. Except in special cases, the asymptotic distribution of a test statistic is not a good approximation to its finite sample distribution uniformly over the null-hypothesis set (Beran, 1988). Regardless of how large the sample is, there may be points in the null hypothesis for which the asymptotic distribution is very far from the exact finite-sample distribution. Such points are typically crucial for determining the size-corrected critical value. When they are, the critical value obtained from the asymptotic distribution need not be close to the size-corrected critical value even if the sample is very large. Except in special cases, the asymptotic critical value approximates the Type I critical value at the parameter point corresponding to the true DGP. It does not approximate the size-corrected critical value. This distinction is discussed further in Section 3.

The remainder of this paper is organized as follows. The Type I critical value is defined in Section 2. The size-corrected critical value is reviewed in Section 3, and the bootstrap critical value is developed in Section 4. These critical values are all based on sampling under the assumption that the null hypothesis is true. In Section 5, we discuss the properties of the bootstrap critical value when the null hypothesis is false. This section also describes our method for obtaining critical values for use in Monte Carlo studies of power. Section 6 concludes the paper.

The statistical tests that we consider in this paper consist of a test statistic and a critical value. The paper is concerned with the choice of critical value *given the test statistic*. The paper does not address the important problem of the choice of a test statistic. In addition, we restrict attention to ‘regular’ hypotheses whose
test statistics are asymptotically normally or chi-square distributed under the null hypothesis. Thus, for example, we do not consider hypotheses that are defined through one-sided inequality constraints such as $H_0: \beta \geq c$ for some population parameter $\beta$ and constant $c$. We do consider hypotheses of the form $H_0: \beta = c$.

2. Type I critical values when the null is true

Let the data be a random sample of size $n$ from a probability distribution whose cumulative distribution function (CDF) is $F$. Denote the data by $\{X_i; i = 1, \ldots, n\}$. $F$ is assumed to belong to a family of CDFs that is indexed by the finite or infinite-dimensional parameter $\theta$ whose population value $\theta^*$. We write $F(x, \theta^*)$ for $P(X \leq x)$ and $F(\cdot, \theta)$ for a general member of the parametric family. The unknown parameter $\theta$ is restricted to a parameter set $\Theta$. The null hypothesis $H_0$ restricts $\theta$ to a subset $\Theta_0$ of $\Theta$. If $H_0$ is composite, then $\Theta_0$ contains two or more points.

Let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic for testing $H_0$. Let $G_n(\tau, F(\cdot, \theta)) = P(T_n \leq \tau|\theta)$ be the exact finite sample CDF of $T_n$ when the CDF of the sampled distribution is $F(\cdot, \theta)$. Consider a symmetrical, two-tailed test of $H_0$. $H_0$ is rejected by such a test if $|T_n|$ exceeds a suitable critical value and accepted otherwise. The exact, $\alpha$-level Type I critical value of $|T_n|$, $z_{n\alpha}$, is defined as the solution to the equation $G_n(z_{n\alpha}^2, F(\cdot, \theta^*)) - G_n(-z_{n\alpha}^2, F(\cdot, \theta^*)) = 1 - \alpha$ for $\theta^* \in \Theta_0$. A test based on this critical value rejects $H_0$ if $|T_n| > z_{n\alpha}$. Such a test makes a Type I error with probability $\alpha$. However, $z_{n\alpha}$ can be calculated in applications only in special cases. If $H_0$ is simple so that $\Theta_0$ contains only one point, then $\theta^*$ is specified by $H_0$, and $z_{n\alpha}$ can be calculated or estimated with arbitrary accuracy by Monte Carlo simulation. If, as usually happens in econometrics, $H_0$ is composite, then $\theta^*$ is unknown and $z_{n\alpha}$ cannot be evaluated unless $G_n(\tau, F(\cdot, \theta))$ does not depend on the location of $\theta \in \Theta_0$ when $H_0$ is true. In this special case, $T_n$ is said to be pivotal. The Student $t$ statistic for testing a hypothesis about the mean of a normal population or a slope coefficient in a normal linear regression model is pivotal. However, pivotal test statistics are not available in most econometric applications unless strong distributional assumptions are made. When $T_n$ is not pivotal, its Type I critical value $z_{n\alpha}$ can be very different at different points in $\Theta_0$. This is illustrated by Example 1 at the end of this section.

When $H_0$ is composite and $T_n$ is not pivotal, it is necessary to replace the true Type I critical value with an approximation or estimator. First-order asymptotic distribution theory provides one approximation. Most test statistics in econometrics are asymptotically pivotal. Indeed, the asymptotic distributions of most commonly used test statistics are standard normal or chi-square under $H_0$, regardless of the details of the DGP. If $n$ is sufficiently large and $T_n$ is asymptotically pivotal, then $G_n(\cdot, F(\cdot, \theta^*))$ can be approximated accurately by the
asymptotic distribution of $T_n$. The asymptotic distribution does not depend on
the location of $\theta^* \in \Theta_0$ when $T_n$ is asymptotically pivotal and $H_0$ is true, so
approximate critical values for $T_n$ can be obtained from the asymptotic distribu-
tion without having to know $\theta^*$.

Critical values obtained from asymptotic distribution theory are widely used
in applications. However, Monte Carlo experiments have shown that first-order
asymptotic theory often gives a poor approximation to the distributions of test
statistics with the sample sizes available in applications. As a result, the true and
nominal probabilities that a test makes a Type I error can be very different when
an asymptotic critical value is used. Under conditions explained later, the
bootstrap provides an approximation that is more accurate than the approxi-
mation of first-order asymptotic theory.

We conclude this section with an example that illustrates the extent to which
the Type I critical value of a test can vary among DGPs in the null-hypothesis
parameter set $\Theta_0$.

**Example 1.** Sensitivity of the Type I critical value to the location of $\theta^*$ in $\Theta_0$.

This example uses Monte Carlo simulation to compute exact Type I critical
information-matrix (IM) test of a binary probit model. The IM test is a specifi-
cation test for parametric models that are estimated by maximum likelihood. It
tests the null hypothesis that the Hessian and outer-product forms of the
information matrix are equal. Rejection implies that the model is misspecified.

In this example, the model being tested is

$$P(Y = 1 | X = x) = \Phi(\beta' x),$$

(2.1)

where $Y$ is a dependent variable whose only possible values are 0 and 1,
$X = (X_1, X_2, X_3)'$ is a vector of explanatory variables, $\beta = (\beta_1, \beta_2, \beta_3)'$ is a vec-
tor of parameters, and $\Phi$ is the standard normal CDF. As in Horowitz (1994),
$X$ consists of constant, $X_1$, and two covariates, $X_2, X_3$, whose values are
sampled independently from the N(0, 1) distribution and are the same in
all Monte Carlo replications. Each Monte Carlo replication consists of generat-
ing an estimation data set of size $n = 100$ by random sampling from the model
(2.1) with a given value of $\beta$, estimating $\beta$ by maximum likelihood, and
computing the IM test statistic using the full vector of indicators. In other
words, all distinct elements of the Hessian and outer product information
matrices are tested simultaneously. The 0.05-level Type I critical value is the 0.95
quantile of the empirical distribution of the IM statistic obtained in 1000 such
replications.

Type I critical values were calculated for the grid of $\beta$ values obtained by
allowing each component of $\beta$ to range over the interval $[0, 2]$ in units of 0.2.
The 0.05 Type I critical values varied from 11.95 to 83.56, depending on the
point in the grid. For example, fixing $\beta_2$ and $\beta_3$ at 0 and letting $\beta_1$ vary yields the following critical values:

<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
<th>1.0</th>
<th>1.2</th>
<th>1.4</th>
<th>1.6</th>
<th>1.8</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crit. Val.</td>
<td>11.95</td>
<td>12.45</td>
<td>14.30</td>
<td>17.08</td>
<td>20.96</td>
<td>26.35</td>
<td>34.92</td>
<td>47.98</td>
<td>67.02</td>
<td>77.60</td>
<td>83.56</td>
</tr>
</tbody>
</table>

The asymptotic critical value is 11.07. Thus, first-order asymptotic theory provides a very poor approximation to the Type I critical value at most points in $\Theta_0$. The main reason is that Studentization of the test statistic entails estimation of the covariance matrix of the covariance matrix of a parameter estimator (Chesher and Spady, 1991). Estimators of such higher-order moments of a distribution are often very imprecise.

3. Size-corrected critical values

When $T_n$ is not pivotal, the size of a test based on $T_n$ can be calculated and controlled if $\Theta_0$ is a sufficiently small set. Suppose that $H_0$ is rejected if $|T_n| > c_{nz}$ for some $c_{nz}$ and accepted otherwise. The size of this test is $\alpha$ and the $\alpha$-level, size-corrected critical value is $c_{nz}$ if

$$\sup_{\theta \in \Theta_0} P(|T_n| > c_{nz}) = \alpha.$$ 

The size can be interpreted as the supremum of the power function when $H_0$ is true.

The $\alpha$-level Type I critical value cannot exceed the $\alpha$-level size-corrected critical value. That is, $z_{nz} \leq c_{nz}$. This follows from the definition of the size of a test. Moreover, the probability of a Type I error is less than or equal to $\alpha$ when the test uses the $\alpha$-level size-corrected critical value, whereas it is exactly $\alpha$ with the Type I critical value:

$$P(|T_n| > c_{nz}) \leq P(|T_n| > z_{nz}) = \alpha$$

for $\theta^* \in \Theta_0$. If $T_n$ is pivotal, the Type I and size-corrected critical values are equal. Otherwise, they are unequal except in special cases. The result that $z_{nz} \leq c_{nz}$ is illustrated in Example 1. If $\Theta_0$ contains the grid used in that example, the 0.05-level size-corrected critical value must be at least 83.56. This value exceeds all of the other Type I critical values in the grid. A further important implication of $z_{nz} \leq c_{nz}$ is that the power of a test based on $c_{nz}$ cannot exceed and may be much smaller than the power of a test based on $z_{nz}$. This point is discussed further in Section 5.

The advantage of the size of a test in comparison to the probability of a Type I error is that the size and the size-corrected critical value do not depend on

Therefore, a size-corrected test can be implemented without knowledge of $\theta$. The $\alpha$-level size-corrected critical value is the solution to a nonlinear optimization problem and usually cannot be calculated analytically. In principle, it can be estimated with arbitrary accuracy by numerical methods. The feasibility of doing this in practice depends on the dimension of $\Theta_0$. Numerical estimation of the size-corrected critical value is feasible if the dimension is low and almost certainly infeasible if it is infinite.

In addition to the difficulty of computing size-corrected critical values, size-corrected tests (that is, tests based on size-corrected critical values) have two fundamental problems. First, the size-corrected critical value may be infinite, in which case a test based on the size concept has no power. Second, even if the size-corrected critical value is finite, the power of the test may be the same as its size. Dufour (1997) gives several examples of the first problem. Bahadur and Savage (1956) demonstrate the second for the case of testing a hypothesis about a population mean. Both problems are consequences of having a null-hypothesis set $\Theta_0$ that is too large. Dufour (1997) has shown that this happens in a wide variety of settings that are important in applied econometrics. Thus, even if computation of a size-corrected critical value is not an issue, size-corrected tests are available only when $\Theta_0$ is a sufficiently small set. Usually, this is accomplished by restricting $F(\cdot, \theta)$ to a suitably small, finite-dimensional family of functions.

Size-corrected critical values are rarely computed in practice, possibly because of the foregoing computational and conceptual problems. However, many papers in econometrics claim to report size-corrected critical values, usually as the results of Monte Carlo experiments designed to compare the powers of competing tests. In most of these papers, what is really computed is the exact Type I critical value at a specific and essentially arbitrary $\theta$ in $\Theta_0$. The empirical relevance of such a critical value is questionable at best. Because the chosen $\theta$ value is arbitrary, the critical value has no empirical analog and cannot be implemented in an application. Except, perhaps, in special cases, the chosen $\theta$ does not correspond to the true DGP in an application if $H_0$ is true.

---

and is not relevant to testing $H_0$ under whatever DGP has generated the data if $H_0$ is false.

We close this section by re-emphasizing that critical values obtained from the asymptotic distribution of a test statistic do not approximate size-corrected critical values in general, regardless of how large $n$ is. For many important tests, the size-corrected critical value is infinite for all finite values of $n$ (Dufour, 1997; De Jong et al., 1992). The asymptotic critical values of these tests are finite, however. Thus, there is no sense in which the asymptotic critical value approximates the size-corrected critical value. Instead, when $n$ is sufficiently large, the asymptotic critical value approximates the Type I critical value at the point $\theta$ corresponding to the true DGP. This point has also been made by Beran (1988).

4. The bootstrap critical value when the null is true

The bootstrap provides a way to obtain approximations to the Type I critical value of a test and the probability of a Type I error that are more accurate than the approximations of first-order asymptotic distribution theory. The bootstrap does this by using the information in the sample to estimate $\theta$ and, thereby, $G_n[\cdot, F(\cdot, \theta^*)]$. The estimator of $G_n[\cdot, F(\cdot, \theta^*)]$ is $G_n(\cdot, F_n)$ where $F_n(x) = F(x, \theta_n)$ and $\theta_n$ is an $n^{1/2}$-consistent estimator of $\theta \in \Theta_0$. The idea is that $\theta_n$ has a high probability of being close to $\theta^*$ when $H_0$ is true. Therefore, $F_n$ is close to $F(\cdot, \theta^*)$. The bootstrap estimator of the $\alpha$-level Type I critical value for $|T_n|$, $z_{n\alpha}^*$, solves $G_n(z_{n\alpha}^*, F_n) - G_n(-z_{n\alpha}^*, F_n) = 1 - \alpha$. Thus, the bootstrap estimator of the Type I critical value is, in fact, the exact Type I critical value under sampling from the distribution whose CDF is $F(\cdot, \theta_n)$.

Usually, $G_n(\cdot, F_n)$ and $z_{n\alpha}^*$ cannot be evaluated analytically. They can, however, be estimated with arbitrary accuracy by carrying out a Monte Carlo experiment in which random samples are drawn from $F_n$. Although the bootstrap is usually implemented by Monte Carlo simulation, its essential characteristic is the use of $F_n$ to approximate $F$ in $G_n[\cdot, F(\cdot, \theta^*)]$, not the method that is used to evaluate $G_n(\cdot, F_n)$. From this perspective, the bootstrap is an analog estimator in the sense of Manski (1988); it simply replaces the unknown $F$ with a sample analog.

The bootstrap usually provides a good approximation to $G_n[\cdot, F(\cdot, \theta^*)]$ and $z_{n\alpha}$ if $n$ is sufficiently large. This is because under mild regularity conditions, $\sup_x |F_n(x) - F(x, \theta^*)|$ and $\sup_x |G_n(\tau, F_n) - G_n[\tau, F(\cdot, \theta^*)]|$ converge to zero in probability or almost surely. Of course, first-order asymptotic distribution theory also provides a good approximation if $n$ is sufficiently large. It turns out, however, that if $T_n$ is asymptotically pivotal and certain technical conditions are satisfied, the bootstrap approximations are more accurate than those of first-order asymptotic theory. See Hall (1992) for the details.
In particular, the bootstrap is more accurate than first-order asymptotic theory for estimating distribution of a ‘smooth’ asymptotically pivotal statistic. It can be shown that
\[ P(|T_n| > z_{n_2}^*) = \alpha + O(n^{-2}) \]
when \( H_0 \) is true and regularity conditions are satisfied. See, for example, Hall (1992). Thus, with the bootstrap critical value and under regularity conditions, the difference between the true and nominal probabilities that a symmetrical test makes a Type I error is \( O(n^{-2}) \) if the test statistic is asymptotically pivotal. In contrast, when a critical value based on first-order asymptotic theory is used, the difference is \( O(n^{-1}) \). This ability to improve upon first-order asymptotic approximations makes the bootstrap an attractive method for estimating Type I critical values. Horowitz (1997) presents implementation details and the results of Monte Carlo experiments showing that the use of bootstrap critical values can dramatically reduce the difference between the true and nominal probability that a test makes a Type I error.\(^2\)

5. Critical values when the null hypothesis is false

The discussion of Type I critical values up to this point has assumed that \( H_0 \) is true so that there is a \( \theta^* \in \Theta_0 \) corresponding to the true DGP. The exact, \( \alpha \)-level Type I critical value of a symmetrical test based on the statistic \( T_n \) is the \( 1 - \alpha \) quantile of the distribution of \( |T_n| \) that is induced by the DGP corresponding to \( \theta = \theta^* \in \Theta_0 \). When \( H_0 \) is false, there is no \( \theta^* \in \Theta_0 \) that corresponds to the true DGP, so it is not immediately clear what \( \theta \) value should be used to obtain the exact Type I critical value. In this section, we propose using a specific \( \theta \) called the pseudo-true value. The bootstrap estimates the exact Type I critical value at the pseudo-true value of \( \theta \). Therefore, when \( H_0 \) is false, the bootstrap provides an empirical analog of a test based on the exact Type I critical value evaluated at the pseudo-true value of \( \theta \).

The problem discussed in this section does not arise with size-corrected critical values or asymptotic critical values for asymptotically pivotal test statistics. This is because size-corrected critical values and asymptotic critical values

\(^2\)Regularity conditions under which the bootstrap provides \( O(n^{-2}) \) accuracy are satisfied in many but not all settings of interest in econometrics. The conditions are satisfied, for example, by ‘regular’ maximum-likelihood and method-of-moments estimators. When the regularity conditions are not satisfied, the accuracy provided by the bootstrap may be worse than \( O(n^{-2}) \) or the bootstrap may be inconsistent. Satisfaction of the conditions for achieving \( O(n^{-2}) \) accuracy does not guarantee good numerical performance of the bootstrap in an application, but the results of many Monte Carlo experiments suggest that the bootstrap performs very well in many settings that are important in applied econometrics.
values for asymptotically pivotal statistics do not depend on $\theta$. Thus, the problem discussed in this section arises only in connection with higher-order approximations to the Type I critical value such as that provided by the bootstrap.

Section 5.1 defines the pseudo-true parameter value and describes the behavior of the bootstrap when $H_0$ is false. Section 5.2 explains how critical values based on the pseudo-true parameter value can be obtained in Monte Carlo studies of the powers of tests. Section 5.3 provides a numerical illustration of the relative powers of tests based on size-corrected critical values and Type I critical values evaluated at the pseudo-true $\theta$.

5.1. The Type I critical value when $H_0$ is false

When $H_0$ is false the true parameter value $\theta^*$ is in the complement of $\Theta_0$ in $\Theta$. There is no $\theta \in \Theta_0$ that corresponds to the true DGP. What value of $\theta$ should then be used to define the Type I critical value? If $H_0$ is simple, $\Theta_0$ consists of a single point, and this point can be used to define the Type I critical value whether $H_0$ is true or false. If $H_0$ is composite, however, there are many points in $\Theta$, and it is not clear which of them, if any, should be used to define the Type I critical value. Our solution to this problem is guided by two considerations. First, in an application it is not known whether $H_0$ is true or false. Second, it must be possible to estimate the chosen $\theta$ value consistently, regardless of whether $H_0$ is true or false.

As has already been explained, we recommend using the bootstrap to estimate the Type I critical value. Therefore, it is useful to consider how the bootstrap operates when $H_0$ is false. Let the test statistic be $T_n$ and let its CDF under the true DGP be $G_n[\cdot, F(\cdot, \theta_0)]$. As has already been explained, the bootstrap estimator of the Type I critical value is obtained from the distribution whose CDF is $G_n[\cdot, F(\cdot, \theta_n)]$, where $\theta_n$ is an estimator of $\theta$. Under regularity conditions (see, e.g. White, 1982; Amemiya, 1985) $\theta_n$ converges in probability or almost surely as $n \to \infty$ to a nonstochastic limit $\theta^0$ and $n^{1/2}(\theta_n - \theta^0) = O_p(1)$. If $H_0$ is true, then $\theta^0 = \theta^*$. Otherwise, $\theta^0 \neq \theta^*$ in general and will be called the pseudo-true value of $\theta$. Regardless of whether $H_0$ is true, the bootstrap computes the distribution of $T_n$ at a point $\theta_n$ that is a consistent estimator of $\theta^0$. Therefore, it is natural to let $\theta^0$ be the parameter point used to calculate the Type I critical value when $H_0$ is false. Doing this insures that $\theta^0$ and the Type I critical value are consistently estimable (e.g. by the bootstrap) regardless of whether $H_0$ is true. Indeed, it can be shown that when $H_0$ is false, the bootstrap provides a higher-order approximation to the Type I critical value based on $\theta^0$ (Horowitz, 1994).

The Type I critical value based on $\theta^0$ is related in a straightforward way to the critical value obtained from first-order asymptotic approximations. If $T_n$ is asymptotically pivotal with asymptotic CDF $G_n(\cdot)$ and $G_n$ satisfies mild smoothness conditions, then as $n \to \infty$, $G_n[\tau, F(\cdot, \theta^0)] \to G_\tau(\tau)$ almost surely or
in probability uniformly over $\tau$. Therefore, the critical value based on $\theta^0$ converges almost surely or in probability to the asymptotic critical value as $n \to \infty$.

The pseudo-true value may or may not be in $\Theta_0$. It is in $\Theta_0$ if the restrictions of $H_0$ are imposed on the estimator $\theta_n$ (only). Otherwise, $\theta^0$ is not in $\Theta_0$. It is of little importance whether $\theta^0$ is in $\Theta_0$ when $H_0$ is false. The important considerations are twofold: (a) the $\theta$ value used to obtain the Type I critical value coincides with $\theta^*$ when $H_0$ is true and has an empirical analog that can be implemented in applications regardless of whether $H_0$ is true, and (b) the resulting test has good power in comparison to alternatives when $H_0$ is false. We have already explained that $\theta^0$ satisfies the first of these conditions. Moreover, convergence of the Type I critical value based on $\theta^0$ to the asymptotic critical value insures that a test based on this Type I critical value inherits any asymptotic optimality properties of a test based on the asymptotic critical value. Section 5.3 provides numerical illustrations of the finite sample power of tests based on Type I critical values with $\theta = \theta^0$.

5.2. Obtaining Type I critical values in Monte Carlo studies of power

Because the Type I critical value based on $\theta^0$ has an empirical analog that can be implemented in applications, it is more attractive for use in Monte Carlo studies of power than the so-called but misnamed ‘size-corrected’ critical values that are often used in such studies. As was explained in Section 3, the critical values obtained in most Monte Carlo studies of power have no empirical analog and, therefore, are irrelevant to applications. The bootstrap method for obtaining Type I critical values that has been proposed in this paper can be implemented in Monte Carlo studies by incorporating bootstrap estimation of the critical value into the Monte Carlo procedure. That is, at each Monte Carlo replication, the bootstrap can be used to estimate the critical value conditional on the data set generated at that replication. This critical value is then used to decide whether $H_0$ is accepted or rejected in the replication. The bootstrap method can also be implemented by using numerical techniques to compute $\theta^0$ for the DGP under consideration and then using the methods of conventional Monte Carlo studies to estimate the exact, finite-sample critical value at $\theta^0$. Horowitz (1994) provides illustrations of both of these ways of obtaining critical values in a Monte Carlo study of power.

5.3. A numerical illustration of power

This section presents a numerical example that illustrates:

(1) The relation between the power of an exact test based on the Type I critical value at $\theta = \theta^0$ and the same test based on the bootstrap critical value, and
the large loss of power that can result from using the size-corrected critical value instead of the Type I critical value based at $\theta = \theta^0$.

**Example 2.** The powers of tests based on Type I and size-corrected critical values.

This example uses Monte Carlo simulation to compute the power of the Chesher (1983) and Lancaster (1984) version of the IM test of a binary probit model. Power is computed using the Type I critical value at the pseudo-true parameter value, the bootstrap critical value, and a version of the size-corrected critical value.

The model tested in this example is (2.1). Data are generated by simulation from two different models. These are

\begin{align*}
P(Y = 1 | X = x) &= \Phi(\beta' x e^{-0.75 \beta' x}) \quad (5.1) \\
P(Y = 1 | X = x) &= \Phi(\beta' x + 1.5 x_2 x_3). \quad (5.2)
\end{align*}

In each model, $X$ is as described in Example 1, and $\beta = (0.5, 1, 1)'$. In (5.2), $x_2$ and $x_3$ are the second and third components of $x$. The sample size is $n = 100$.

Model (5.1) is a heteroskedastic binary probit model. It is equivalent to the model $Y = I(\beta' X + U > 0)$, where $I$ is the indicator function and $U$ is an unobserved random variable whose distribution is $N(0, e^{\beta' X})$. When data are generated from (5.1), model (2.1) is misspecified because it does not take account of heteroskedasticity. When data are generated from (5.2), model (2.1) is misspecified because it does not take account of the interaction term $X_2 X_3$.

The powers of the Chesher–Lancaster IM test based on Type I critical values at $\theta^0$ and on bootstrap critical values are reported in Horowitz (1994), who also gives further details of the simulation procedure. The results are:

<table>
<thead>
<tr>
<th>Model</th>
<th>Type I critical value</th>
<th>Bootstrap critical value</th>
<th>Size-corrected critical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5.1)</td>
<td>0.715</td>
<td>0.667</td>
<td>0.000</td>
</tr>
<tr>
<td>(5.2)</td>
<td>0.925</td>
<td>0.875</td>
<td>0.000</td>
</tr>
</tbody>
</table>

As expected, the powers based on Type I and bootstrap critical values are close.

Now consider the power based on the size-corrected critical value. The results of Example 1 show that this critical value is at least 83.56 if the parameter set for $\beta$ is at least as large as $[0, 2]^3$. The table above shows that with this critical value, the power of the test is zero against both alternatives. This result can easily be understood. The Type I critical values at $\theta^0$ are 21.7 and 23.4 with models (5.1) and (5.2), respectively. The low power of the size-corrected test results from its use of a critical value that is much larger than the Type I critical value.
6. Conclusions

This paper has been concerned with the choice of critical values for tests of statistical hypotheses. Two types of critical values have been defined, the Type I critical value and the size-corrected critical value. These critical values are conceptually distinct and their numerical values are unequal except in special cases that usually entail strong distributional assumptions. The Type I critical value usually depends on details of the data generation process that are unknown in applications, so it cannot be evaluated exactly in applications. However, it can often be estimated very accurately by using the bootstrap. It can also be estimated by using first-order asymptotic distribution theory, but the resulting approximation can be highly inaccurate in samples of practical size. The size-corrected critical value does not depend on unknown population parameters, so it does not require making asymptotic approximations that may be inaccurate. It can, however, be difficult to compute, possibly intractably so. More importantly, the size-corrected critical value is not necessarily finite. Even when it is finite, the power of a size-corrected test is likely to be lower than the power of the same test with the Type I critical value. Often, the power of a size-corrected test is equal to or less than its size. We conclude, therefore, that tests based on Type I critical values are preferable to size-corrected tests. Such tests can often be implemented empirically with the bootstrap or, when the sample size is sufficiently large, by use of first-order asymptotic approximations.

This paper has also pointed out that the methods used to obtain the so-called size-corrected critical values in most Monte Carlo studies of the finite-sample powers of tests cannot be implemented in applications. The critical values obtained in these studies are irrelevant to empirical research. Because the methods proposed in this paper can be implemented in applications, it seems to us that they should replace existing methods for obtaining critical values in Monte Carlo studies of the powers of tests.

Acknowledgements

We thank Art Goldberger, Beth Ingram, Tom Rothenberg, and Paul Weller for comments on an earlier draft of this paper. The research of Joel L. Horowitz was supported in part by NSF grant SBR-9617925.

References

Hansen, B.E., 1996. Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64, 413–430.