Published in *Statistical Papers* 43,567-577 (2002)
The final publication is available at www.springerlink.com

Is a Small Monte Carlo Analysis a Good Analysis?

Checking the Size, Power and Consistency of a Simulation-Based Test

Ignacio Díaz-Emparanza[1] ⋆

Departamento de Econometría y Estadística e Instituto de Economía Pública. Universidad del País Vasco - Euskal Herriko Unibertsitatea.

Received: date / Revised version: date

Abstract In this paper we study the relationship between the number of replications and the accuracy of the estimated quantiles of a distribution obtained by simulation. A method for testing hypotheses on the quantiles of a theoretical distribution using the simulated distribution is proposed, as well as a method to check the hypothesis of consistency of a test.

1 Introduction

The appearance and development of computers over the last 30 years has proved to be a great advance for statistical and economic science. The calculation capabilities of computers has allowed solutions to be found to some problems that seemed impossible to solve analytically. Also the speed of processors has increased more than 50-fold in the last 10 years, so that calculations that were previously unthinkable because of their duration, can now be performed quickly, and it is also possible to manage bigger sets of data.

One of the tasks in applied statistics or econometrics which is now usually done by computer is the estimation or approximation of probability distributions by means of Monte Carlo simulations. This is a very frequent activity; so frequent, in fact, that it is hard to find articles that do not make use of it.

However, in recent years a great number of papers are being published in which the author or authors are content to carry out a Monte Carlo

⋆ Financial support from research projects PB96-1469-C05-01, UPV-038.321-G55/98 and PI9970 is gratefully acknowledged.

study with only a small number of replications —usually one thousand and sometimes less—. This is somewhat surprising. It is obvious that the quality of these approaches depends directly on the number of replications used. Then, in a world where technology is advancing so fast, is it not worthwhile taking advantage of that technology to obtain higher quality estimations?

The main goal of this work is to answer this question. In section 2 we study the relationship between the number of replications and the accuracy level in the estimation by means of simulation of the quantiles of a probability distribution; in section 3 we suggest tests on the *size* and *power* of a test; in section 4 a test for the hypothesis of consistency is proposed, section 5 shows an example of its application and section 6 gives the conclusions.

2 Accuracy of Empirical Approximation.

Let y be a $(P \times 1)$ vector and let $y_1, \ldots, y_N$ be the available sample of that vector. Let us assume that the probability distribution of y –whatever it may be— is known. Let $Y = (y_1, y_2, \ldots, y_N)'$ be the $(N \times P)$ matrix that contains in each column the N observations of each component of y and f a function such that for each value of Y there is a real value X, such that,

$$X = f(Y) \in \Re$$

The probability distribution of X is, in general, unknown. The usual way of estimating it through the Monte Carlo method is as follows:

1. Generate by computer T different samples —replications— of N observations of the vector y. That is to say, T realizations of the matrix Y, coming from its theoretical probability distribution, which is known.
2. Calculate the value of the statistic $X_t = f(Y_t)$ for each replication, where Y_t is the simulated value of the matrix Y at the t-th replication and X_t is the value obtained for the statistic in this replication, with $t = 1, \ldots, T$.
3. Order the calculated values of $X_1, \ldots, X_T$ and take their distribution of relative frequencies as an approximation of the density function, which is unknown. Starting from the distribution of relative frequencies, calculate confidence intervals and run hypotheses tests as if this was the theoretical distribution.

(Finster, 1987) defines the following concept of an *accurate estimate*.

Definition 1 *$\widehat{p}$ is an "*accurate estimate*" of p with accuracy A and confidence $1-\alpha$ (with $0 < \alpha < 1$), if*

$$\Pr\left[|\widehat{p} - p| \leq A\right] \geq 1 - \alpha \tag{1}$$

$[-A, A]$ is the set of acceptable simulation errors.

(Kleijnen, 1987, p. 38) and (Díaz-Emparanza, 1995) have studied the relationship between the number of replications used in a Monte Carlo study

and the accuracy of the approximation obtained. In the following, we will present this result.

Let H be any interval defined on $\Re$. Assume that X_H is an indicator variable defined as:

$$X_{Ht} = \begin{cases} 1 \text{ if } \ X_t \in H \\ \\ 0 \text{ if } \ X_t \notin H \end{cases} \tag{2}$$

So that each observation of X_t has an associated observation (with value 0 or 1) of the variable X_{Ht}. The (unknown) theoretical density function of X_t assigns a probability p_H to the H interval. This means that

$$\Pr\left[X_t \in H\right] = \Pr\left[X_{Ht} = 1\right] = p_H \tag{3}$$

Producing T replications of the matrix Y means having a sample of T "observations" of the real variable X. This sample also has an associated sample of size T of the variable X_H. This variable follows a binary distribution with parameter p_H, so the sum of the T observations of X_H, $Z_H = X_{H1} + \cdots + X_{HT}$, follows a binomial distribution $b(p_H, T)$. The Moivre-Laplace theorem proves that the sequence $b(p_H, 1), b(p_H, 2), \ldots, b(p_H, T), \ldots$ is asymptotically normal $N(T\,p_H, T\,p_H[1 - p_H])$. Different authors propose different conditions for considering the approximation as reasonably accurate: for example (Kleijnen, 1987, p. 38) affirms that the approximation is sufficiently precise if $T > 20$, (Fernández de Trocóniz, 1993) says that the approximation can be considered as valid if $T \cdot p_H > 18$, but the most general agreement is that the normal approximation to the binomial distribution is reasonably accurate if both

$$T \cdot p_H \geq 5 \qquad \text{and} \qquad T \cdot (1 - p_H) \geq 5 \tag{4}$$

are satisfied [see for example (Hogg & Tanis, 1988) or (Cryer & Miller, 1991)]. Under these conditions,

$$Z_H \approx N\left(T\,p_H, T\,p_H(1 - p_H)\right) \tag{5}$$

then, for the binomial frequency, Z_H/T, we have

$$\frac{Z_H}{T} \approx N\left(p_H, \frac{p_H(1 - p_H)}{T}\right) \tag{6}$$

If $\lambda_{\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile on the right tail of the distribution N(0,1),

$$\Pr\left[\left|\frac{Z_H}{T} - p_H\right| \leq \lambda_{\frac{\alpha}{2}} \sqrt{\frac{p_H(1 - p_H)}{T}}\right] \simeq 1 - \alpha \tag{7}$$

Comparing expressions (7) and (1) we can see that $\lambda_{\frac{\alpha}{2}} \sqrt{p_H(1 - p_H)/T}$ plays here the role of accuracy A in the estimation of p_H by means of

$\widehat{p}_H = Z_H/T$. This provides a way of relating the number of replications with the accuracy[1]:

$$A = \lambda_{\frac{\alpha}{2}} \sqrt{\frac{p_H(1-p_H)}{T}} \tag{8}$$

3 Tests on Size and Power

The above expression allows us to observe that for a fixed number of replications, T, the accuracy, A, is different for the different H intervals that we want to estimate, in other words, for different values of p_H. (Usually H is a quantile of the distribution and p_H its corresponding probability).

The accuracy level is maximum for the maximum of the function $p_H(1-p_H)$, that is to say, for $p_H = 0.5$ and it is minimum ($A = 0$) for $p_H = 0$ and $p_H = 1$. So, for example, for $T = 1000$, $\alpha = 0.05$ and $p_H = 0.5$ a value of $A = 0.03099$ is obtained. This is the maximum accuracy level.

It is also easy to observe that function (8) is symmetrical with respect to 0.5 so that for p_H and for $p_H^* = (1-p_H)$ the same value of A is obtained. For example, for $p_H = 0.05$ or $p_H^* = 0.95$ and the above values of T and α, we have $A = 0.01351$.

As indicated above, one of the most frequent objectives in a Monte Carlo analysis is to estimate the distribution of probabilities of some statistic to test a certain null hypothesis. There are other occasions in which one knows —or has simulated— the asymptotical distribution, and the Monte Carlo analysis is used to check whether the application of the critical values of this distribution when the number of observations is moderate will produce satisfactory results. In these applications it is vital to know one characteristic of the test under the null hypothesis: *size*, and another characteristic of the test under the alternative hypothesis: *power*. Usually a test with a small *size* —i.e. a small probability of type I error— and a big *power* —i.e. a small probability of type II error too— is considered desirable.

However accuracy A is measured in terms of nominal probability, so its effect when estimating a probability p has varying importance, and it should be valued in a different way if p is a big value than if p is small. Let us suppose, for example, that we want to analyze the *size* of a test that is carried out with a moderate number of observations, having chosen the critical value corresponding to a significance level or nominal *size* $\alpha^* = 0.05$ in the asymptotical distribution (α^* and α must not be mistaken. α^* is the significance level of the proposed test, which is a probability obtained from the distribution function —unknown— of X, and α is a probability obtained from the distribution function of the binomial frequency Z_H/T that, as we

[1] If conditions (4) are not satisfied, or if we wish for more precision in the determination of A, we can work directly with the quantiles of the binomial $b(p_H, T)$ distribution. Then if z_ε is the $1-\varepsilon$ quantile of such distribution, we obtain $A = Max\left\{\left(\frac{z_{\frac{\alpha}{2}}}{T} - p_H\right), -\left(\frac{z_{1-\frac{\alpha}{2}}}{T} - p_H\right)\right\}$.

have seen in (6), is approximately Gaussian.). The accuracy that will be obtained in the estimation of that *size* if the simulation is carried out with $T = 1000$ replications and $\alpha = 0.05$, in accordance with expression (8), will be approximately 0.01351. This means that the variation that we can hope for in *nominal size* may be as much as 27% ($A/0.05 = 0.01351/0.05 = 0.27$). On the other hand, let us suppose that we want to analyze the *power* of the test against some alternative hypothesis, and it is known (for example by an exploratory simulation) that the nominal *power* is 95 %. For the same values of $T$and α equation (8) gives the same value of $A = 0.01351$. But, as this is measured in terms of nominal probability, the variation that we can now hope for in nominal *power* is only 1.42 % (A/0.95=0.01351/0.95=0.0142). So, as we said, the same accuracy level does not have the same relative importance when it is obtained as an estimation of a probability $p = 0.05$ as it does when it is obtained as an estimation of $p = 0.95$.

Expression (6) also provides a method for checking hypotheses on the *size* and *power* of a test. Under the hypothesis $p_H = p_H^0$

$$\frac{Z_H}{T} \approx N\left(p_H^0, \frac{p_H^0(1-p_H^0)}{T}\right) \tag{9}$$

Thus, if we want to test the null hypothesis that the *size* of the test is less than a pre-set value p_H^0, we can propose the test:

$$\begin{array}{ll} H_0: & p_H \leq p_H^0 \\ H_a: & p_H > p_H^0. \end{array} \tag{10}$$

In such a case, under the null hypothesis, at significance level α,

$$\tau_s = \frac{\frac{Z_H}{T} - p_H^0}{\sqrt{\frac{p_H^0(1-p_H^0)}{T}}} < \lambda_\alpha. \tag{11}$$

If we find that $\tau_s > \lambda_\alpha$ it can be considered that there is enough evidence against H_0 to reject it. For example, if we want to test the hypothesis that the *size* of a test is lower than 5 % when $\alpha^* = 0.05$, considering $\alpha = 0.05$ and $T = 1000$, solving (11) for Z_H/T we obtain that this hypothesis should be rejected if Z_H/T is greater than 0.06130.

On the other hand, if we want to test the null hypothesis that the *power* of a testing procedure for a given α^* is higher than a pre-set value p_H^0, it would be necessary to perform the check:

$$\begin{array}{ll} H_0: & p_H \geq p_H^0 \\ H_a: & p_H < p_H^0. \end{array} \tag{12}$$

In this case, under the null hypothesis, at significance level α,

$$\tau_p = \frac{\frac{Z_H}{T} - p_H^0}{\sqrt{\frac{p_H^0(1-p_H^0)}{T}}} > -\lambda_\alpha. \tag{13}$$

If it is found that $\tau_p < -\lambda_\alpha$, there is also enough evidence to reject H_0. For example, if we want to check the hypothesis that the *power* of a test is higher than 0.95 also considering $\alpha = 0.05$ and $T = 1000$, solving 13 for Z_H/T the result is that the hypothesis should be rejected if $Z_H/T < 0.93870$.

4 A Test for the Hypothesis of Consistency

Expression (6) also provides a way of checking whether a testing method is consistent. Remember that a test is consistent if when the sample size, N, is increased, the probability of rejecting false hypotheses also increases. A form of checking this statistically is as follows. Generating two independent groups of replications of the statistic under the alternative hypothesis, the first one with sample size N_1 and T_1 replications, the second with sample size $N_2 > N_1$ and T_2 replications, where T_2 can be bigger than, smaller than or equal to T_1, from expression (6) we have that

$$\frac{Z_1}{T_1} \approx N\left(p_1, \frac{p_1(1-p_1)}{T_1}\right) \tag{14}$$

$$\frac{Z_2}{T_2} \approx N\left(p_2, \frac{p_2(1-p_2)}{T_2}\right) \tag{15}$$

Z_1 is the number of times that the statistic falls in the critical region in the first group of replications and Z_2 in the second. p_1 is the power of the test when the statistic is calculated with N_1 observations and p_2 when it is calculated with N_2 observations. The consistency hypothesis of the test implies that p_2 must be bigger than p_1, so $p_2 - p_1 > 0$ and, as these two groups of replications are independent,

$$\frac{Z_2}{T_2} - \frac{Z_1}{T_1} \approx N\left(p_2 - p_1, \frac{p_1(1-p_1)}{T_1} + \frac{p_2(1-p_2)}{T_2}\right) \tag{16}$$

Then, for the consistency hypothesis we could set up the test:

$$\begin{array}{ll} H_0: & p_2 - p_1 \geq 0 \\ H_a: & p_2 - p_1 < 0. \end{array} \tag{17}$$

Under this null hypothesis, it must be verified that

$$\tau_c = \frac{\frac{Z_2}{T_2} - \frac{Z_1}{T_1}}{\sqrt{\frac{p_1(1-p_1)}{T_1} + \frac{p_2(1-p_2)}{T_2}}} > -\lambda_\alpha \tag{18}$$

As in this case p_1 and p_2 are unknown, in the denominator of the statistic τ_c we can replace them by their approximate values, Z_1/T_1 and Z_2/T_2. Finding that $\tau_c < -\lambda_\alpha$ is sufficient evidence to reject the consistency hypothesis at a significance level α.

A different, and maybe more detailed, way of testing the consistency hypothesis could be based on a third group of replications of the statistic,

now with sample size $N_3 > N_2$ and T_3 replications. Under this hypothesis, $p_3 > p_2$ and $p_2 > p_1$, therefore, it should happen simultaneously that $p_2 - p_1 > 0$ and $p_3 - p_2 > 0$. If the three groups of replications have been generated in an independent way, the vector $(\frac{Z_1}{T_1} \ \frac{Z_2}{T_2} \ \frac{Z_3}{T_3})'$ will have the multinomial distribution:

$$\begin{pmatrix} Z_1/T_1 \\ Z_2/T_2 \\ Z_3/T_3 \end{pmatrix} \approx N \left[\begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}, \begin{pmatrix} \frac{p_1(1-p_1)}{T_1} & 0 & 0 \\ 0 & \frac{p_2(1-p_2)}{T_2} & 0 \\ 0 & 0 & \frac{p_3(1-p_3)}{T_3} \end{pmatrix} \right] \tag{19}$$

To test the hypothesis of consistency the joint hypothesis:

$$H_0 : \begin{bmatrix} p_2 - p_1 \geq 0 \\ p_3 - p_2 \geq 0 \end{bmatrix} \tag{20}$$

could be posited. However when outlining a type-F statistic for the overall test of these two restrictions we find the problem that, with this type of statistics, situations in which the alternative hypothesis is verified (for example $p_1 >> p_2$) and extreme situations in which the null hypothesis is verified ($p_2 >> p_1$) are overlapped in the rejection region. Therefore, in this situation it is more appropiate to outline the two restrictions sequentially: first testing

$$\begin{aligned} H_{01} &: p_2 - p_1 \geq 0 \\ H_{a1} &: p_2 - p_1 < 0 \end{aligned}$$

by means of the τ_c statistic, then, if H_{01} is not rejected, checking

$$\begin{aligned} H_{02} &: p_3 - p_2 \geq 0 \\ H_{a2} &: p_3 - p_2 < 0. \end{aligned} \tag{21}$$

Here, it is necessary to keep in mind that the distribution of $\frac{Z_3}{T_3} - \frac{Z_2}{T_2}$ is not independent of that of $\frac{Z_2}{T_2} - \frac{Z_1}{T_1}$. For that reason, to build a t-Student type statistic it is necessary to take as the basis the conditional distribution

$$\begin{aligned} &\left(\frac{Z_3}{T_3} - \frac{Z_2}{T_2} / \frac{Z_2}{T_2} - \frac{Z_1}{T_1} \right) \approx \\ &N \left(p_3 - p_2 + \frac{\sigma_{12}}{\sigma_2^2} \left(\frac{Z_2}{T_2} - \frac{Z_1}{T_1} - (p_2 - p_1) \right), \ \sigma_1^2 - \frac{(\sigma_{12})^2}{\sigma_2^2} \right) \end{aligned} \tag{22}$$

where

$$\begin{aligned} \sigma_1^2 &= \frac{p_1(1-p_1)}{T_1} + \frac{p_2(1-p_2)}{T_2} \\ \sigma_{12} &= -\frac{p_2(1-p_2)}{T_2} \\ \sigma_2^2 &= \frac{p_2(1-p_2)}{T_2} + \frac{p_3(1-p_3)}{T_3} \end{aligned} \tag{23}$$

then, under the null hypothesis of consistency

$$\tau_c^{(2)} = \frac{\frac{Z_3}{T_3} - \frac{Z_2}{T_2} - \left[-\frac{\sigma_{12}}{\sigma_2^2}\left(\frac{Z_2}{T_2} - \frac{Z_1}{T_1}\right)\right]}{\sqrt{\sigma_1^2 - \frac{(\sigma_{12}^2)}{\sigma_2^2}}} > -\lambda_\alpha \tag{24}$$

So if a value $\tau_c^{(2)} < -\lambda_\alpha$ is found, the hypothesis of consistency will be rejected.

We must keep in mind that if the significance level α is used in each individual test, the significance level of the entire procedure is $\varepsilon = 2\alpha - \alpha^2$, so that if we want to use, for example, a significance level of $\varepsilon = 0.05$ in the overall procedure, it would be necessary to use approximately $\alpha = 0.025$ in each individual test.

In (22), the expression for the mean of the conditional distribution shows us the need for the sequential procedure. It is not enough to carry out the second individual test as, in it, very negative values of $p_2 - p_1$ can be compensated by very positive ones of $p_3 - p_2$ and hence the statistic does not appear in the rejection region when the test is, however, inconsistent. Therefore, more power will be obtained in the sequential test of the consistency hypothesis.

5 An example

In the literature on nonstationary (integrated) time series there are many articles in which, as analytical probability distributions of certain test statistics cannot be calculated, Monte Carlo simulations are used to approximate it. There are countless examples. Some of the main ones are: (Dickey & Fuller, 1981), (Hasza & Fuller, 1982), (Stock & Watson, 1988), (Johansen, 1988), (Hylleberg, Engle, Granger & Yoo, 1990), (Kwiatkowski, Phillips, Schmidt & Shin, 1992), ... In many of these articles, the author/s get conclusions about the respective tests based only on the sample powers and sizes of the tests, without taking into account that these are only approximations of the powers and sizes that are obtained from the nominal distributions. In the following example we will show how the results of the above sections for testing hypotheses about the size and power of a test can be applied. We will base this on the simulation results taken from the aforesaid article by Stock and Watson.

In that article, the authors showed that cointegrated multiple time series share at least one common trend. Two tests were developed for the number of stochastic trends (i.e., for the order of cointegration) in a multiple time series with and without drift. Both tests involved the roots of the ordinary least squares coefficient matrix obtained by regressing the series onto its first lag. $q_f(k, m)$ was an statistic for testing k versus m common trends and $q_c(k, k-1)$ was an statistic for testing k versus $k-1$ common trends. The

Table 1 Monte Carlo Experiment Results: Rejection Probabilities. [Stock & Watson (1988)].

		Data-generating process			
		(25), with $\phi = .4$		(26), with $\theta = .4$	
		A,	B,	C,	D,
ρ	Level	$q_f^\mu(2,1)$	$q_c^\mu(2,1)$	$q_f^\mu(2,1)$	$q_c^\mu(2,1)$
1.00	5%	.03	.03	.03	.07
	10%	.07	.06	.06	.13
.95	5%	.11	.10	.08	.22
	10%	.21	.18	.19	.35
.90	5%	.40	.34	.30	.60
	10%	.59	.50	.51	.74
.80	5%	.92	.82	.86	.99
	10%	.97	.90	.95	.99

The results were computed using 2000 Monte Carlo draws with a sample size of $N = 200$

counterparts of $q_f(k,m)$ and $q_c(k,k-1)$ when there might be a nonzero intercept or a noncero intercept and drift were respectively $q_f^\mu(k,m)$, $q_f^\tau(k,m)$, $q_c^\mu(k,k-1)$ and $q_c^\tau(k,k-1)$. The quantiles of their asymptotical distributions were obtained using 30,000 Monte Carlo replications with sample size $N = 1000$.

In section 7 of the afore-mentioned article, they report the results of a "small Monte Carlo experiment" to investigate the *size* and *power* of their tests. They study the statistics $q_f^\mu(2,1)$ and $q_c^\mu(2,1)$ in particular. The experiments were performed using $T = 2000$ replications with a sample size of $N = 200$. The results are in table 1.

Two different models are considered. In the first, Y_t is generated by a VAR(2):

$$(1 - \phi L)(1 - \Phi L)Y_t = \varepsilon_t \tag{25}$$

and in the second by a VARMA(1,1):

$$(1 - \Phi L)Y_t = (1 + \theta L)\varepsilon_t \tag{26}$$

Where in both cases $E\varepsilon_t\varepsilon_t' = G$ and

$$\Phi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho & 0 \\ 0 & 0 & 0.5 \end{bmatrix}, \qquad G = \begin{bmatrix} 1 & 0.5 & -0.25 \\ 0.5 & 1 & 0.5 \\ -0.25 & 0.5 & 1 \end{bmatrix}.$$

Both ϕ and θ are scalars that are less than 1 in absolute value. Under the null hypothesis $\rho = 1$, so there are two common trends; under the alternative $|\rho| < 1$, and there is only one common trend.

Under these conditions, using the statistic τ_s defined in (11) the null hypothesis that the *size* of the test, p_H^0, is smaller than 5 % should be rejected if the estimation Z_H/T is bigger than $p_H^0 + \lambda_\alpha \sqrt{\frac{p_H^0(1-p_H^0)}{T}} = 0.058$. When the data generating process is (25) with $\phi = 0.4$ and $\rho = 1$, at significance level $\alpha^* = 0.05$, using the statistic $q_f^\mu(2,1)$, table 1 shows an estimated rejection probability of 0.03, so the hypothesis that the *size* is inferior to 5 % cannot be rejected. But when the data generating process is (26) with $\theta = 0.4$ and $\rho = 1$, at significance level $\alpha^* = 0.05$, using the statistic $q_c^\mu(2,1)$, the table shows an estimated rejection probability of 0.07, so the null hypothesis would be rejected in this case.

As for *power*, using the statistic τ_p defined in (13) the null hypothesis that the *power* of the test (p_H^0) is bigger than —for example— 95 % should be rejected if the estimate Z_H/T is smaller than $p_H^0 - \lambda_\alpha \sqrt{\frac{p_H^0(1-p_H^0)}{T}} = 0.942$. When the data generating process is (25) with $\phi = 0.4$ and $\rho = 0.8$, at significance level $\alpha^* = 0.05$, using the statistic $q_f^\mu(2,1)$, table 1 shows an estimated rejection probability of 0.92, so at significance level $\alpha = 0.05$ the hypothesis that the *power* is bigger than 95 % is rejected. When the data generating process is (26) with $\theta = 0.4$ and $\rho = 0.8$, at significance level $\alpha^* = 0.05$, using the statistic $q_c^\mu(2,1)$, the table reports an estimated rejection probability of 0.99, so in this case this hypothesis would not be rejected.

Stock & Watson do not perform any check in their article on the consistency of their tests. Let us suppose that a new group of 2000 replications of the $q_f^\mu(2,1)$ statistic is generated, now with sample size $N = 300$, using data generating process (25) with $\phi = 0.4$ and $\rho = 0.8$, obtaining an estimated rejection probability of 0.91. The test appears to be inconsistent given that when the sample size is increased, the rejection probability of a false hypothesis decreases. However, both rejection probabilities have been calculated on the basis of simulations and are therefore only aproximations of the theoretical rejection probabilities. The theory developed in section 4 allows us to statistically test the hypothesis of consistency. According to expression (18) if the test is consistent, in this new simulation we should find that $Z_2/T > Z_1/T - \lambda_\alpha \sqrt{\frac{p_1(1-p_1)}{T} + \frac{p_2(1-p_2)}{T}}$; replacing p_1 by $\frac{Z_1}{T} = 0.92$ and p_2 by $\frac{Z_2}{T} = 0.91$ we obtain that under the consistency hypothesis $\frac{Z_2}{T} > 0.9056$, therefore the value $\frac{Z_2}{T} = 0.91$ does not imply the rejection of this hypothesis at the 5% significance level.

6 Conclusions

When a probability distribution is approximated by means of simulation it is obvious that the bigger the number of replications is, the better the approach will be. The selection of the number of replications implies the determination of an accuracy level in the estimation of the different quantiles that is given by expression (8).

When many replications are used, we obtain very reliable, highly accurate estimates. In these cases the characteristics of the estimated distribution will be very similar to those of the theoretical distribution and, consequently, the conclusions we get starting from the empirical distribution will be almost identical to those that would be obtained starting from the theoretical distribution.

However, when using a *small* number of replications the empirical distribution will be further away from the theoretical one and more care will be needed when inferring results on the latter starting from the characteristics of the former. A *small* number of replications produces a low level of precision, and this must be taken into acount when we want to reach conclusions about the characteristics of the unknown distribution. Expression (8) allows us to perform tests on the different quantiles of the distribution. In particular, tests on the *size*, *power* and the hypothesis of *consistency* of a test have been presented here.

References

Cryer, J. B. & Miller, R. B. (1991), *Statistics for Business: Data Analysis and Modelling*, PWS-KENT publishing Company. Boston.

Díaz-Emparanza, I. (1995), 'Selección del número de replicaciones en un estudio de simulación', *Estadística Española* **37**(140), 497–509.

Dickey, D. & Fuller, W. (1981), 'Likelihood ratio statistics for autoregressive time series with a unit root', *Econometrica* **49**, 1057–1071.

Fernández de Trocóniz, A. (1993), *Probabilidades. Estadística. Muestreo.*, Tebar Flores.

Finster, M. P. (1987), 'An analysis of five simulation methods for determining the number of replications in a complex monte carlo study', *Statistics & Probability Letters* **5**, 353–360.

Hasza, D. & Fuller, W. (1982), 'Testing for nonstationary parameter specifications in seasonal time series models', *The Annals of Statistics* **10**, 1209–1216.

Hogg, R. W. & Tanis, E. A. (1988), *Probability and Statistical Inference*, Macmillan Publishing Company. New York.

Hylleberg, S., Engle, R. F., Granger, C. W. J. & Yoo, B. S. (1990), 'Seasonal integration and cointegration', *Journal of Econometrics* **44**, 215–38.

Johansen, S. (1988), 'Statistical analysis of cointegration vectors', *Journal of Economic Dynamics and Control* **12**, 231–254.

Kleijnen, J. P. C. (1987), *Statistical tools for simulation practitioners*, Marcel Dekker, Inc., New York.

Kwiatkowski, D., Phillips, P., Schmidt, P. & Shin, Y. (1992), 'Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?', *Journal of Econometrics* **54**, 159–178.

Stock, J. & Watson, M. (1988), 'Testing for common trends', *Journal of the American Statistical Association* **83**, 1097–1107.