SELECTING THE NUMBER OF REPLICATIONS IN A SIMULATION STUDY

Ignacio Díaz-Emparanza

Departamento de Econometría y Estadística

Universidad del País Vasco - Euskal Herriko Unibertsitatea

Avda. Lehendakari Aguirre, 83.

E48015 BILBAO Spain

e-mail: id@alcib.bs.ehu.es

Tfno: (94) 4797432

December 16, 1996

SELECTING THE NUMBER OF REPLICATIONS IN A SIMULATION STUDY

Abstract

In order to approach a distribution of probability by means of simulation it is necessary to determine a number of replications. The accuracy with which the distribution is calculated will rely on this number of replications. In this work, a relationship between the number of replications and the accuracy of the estimate is obtained, so that if it is desired to get a preset value for the accuracy it is possible determine which it will be the minimum number of necessary replications for it.

Key words: Number of replications, Monte-Carlo, accuracy, binomial distribution.

AMS clasification: 62E25

1 Introduction

The high capacity of calculation of the computers in last years is allowing to solve some statistical problems that before seemed impossible to resolve. In the statistical and econometrical literature is frequent to find works in that the probability distribution of certain statistic, that seems impossible to be developed analitically, is obtained by means of simulations by computer. For example, this type of practice is very usual in works that deal with non stationary processes with unit roots (See e.g. Dickey & Fuller (1981), Dickey, Hasza & Fuller (1984), Hylleberg, Engle, Granger & Yoo (1990), Beaulieu & Miron (1993)). Simulations by computer are also used in order to check the power of certain statistics against alternative hypotheses that implicate an unknown distribution, for example: Dickey & Fuller (1979),Phillips & Perron (1988).

When the analyst is confronted with a concrete problem in that needs to provide of the distribution of probability —although it is approached— of a determined statistic, one of the decisions that he has to take, if he decides to calculate it by means of simulation, it refers to the number of replications that he is going to carry out. Any analyst knows that the optimum number of replications to use upon finding an empirical distribution by means of simulations is infinite, but obviously, it is impossible in practice to work with series of data of infinite observations.

As whoever could imagine, the error that is committed upon approaching a distribution by means of simulation is conversely proportional to the number of replications carried out, so, if it is wanted to minimize the error, the number of replications must be the higher possible.

But in the majority of the works that utilizes these methods, the number of replications is selected arbitrarily, without having an idea on the precision that is gotten upon estimating the distribution of probability. This could be seen, for example, in any of the articles referenced in the first paragraph. So that whatever that be the concrete problem that is studied, one could always wonder the same questions:

- Would there be a significant gain —in terms of accuracy with that the distribution is approached— if does be carried out, do we say, 1,000 replications more?

- Given that, sometimes, the calculation of just one replication could be slow and expensive, will be enough consider 1,000 replications?

Helping to respond to these two questions is the fundamental target of the present work.

2 Presentation.

Suppose that we have a sample of size N of a variable vector y that has dimension P. We will also suppose that the probability distribution of y —whatever that is— is known. Let be Y the matrix $(P \times N)$ that contains the N observations of each one of the components of y at each row, and f a function —that is usually denominated *statistic*— such that to each value of Y makes to correspond a real value X, that is to say,

$$X = f(Y) \in \Re. \tag{1}$$

The probability distribution of X is, in general, unknown. Next, the problem of finding its approximation by means of the method of Monte Carlo will be studied.

3 Empirical approximation to the theoretical distribution.

The usual way of approaching a probability distribution using the method of Monte Carlo is the following:

1. First, T different samples of size N for the vector Y are generated through computer (they are denominated *replications*), coming from its theoretical probability distribution, that is known.

2. For every replication the value that takes the f statistic: $X_t = f(Y_t)$ is calculated, where Y_t is the value simulated of the matrix Y at the t-th replication and X_t is the value obtained for the statistic at this replication, with t= 1,..., T.

3. The values calculated for $X_1, \ldots, X_T$ are ordered and their distribution of relative frequencies is taken like approximation of the density function, that is unknown.

Starting from the distribution of relative frequencies, intervals of confidence are calculated and hypothesis tests are done as if this were the theoretical distribution.

4 Accuracy of the empirical approximation.

Let be H any interval on $\mathcal{R}$. We will define now a dummy variable, X_H, of the following form:

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

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

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

Producing T replications of the vector y implicates provide a sample of T "observations" of the real variable X. This sample has associated a sample of size T of the binary variable X_H. This variable follows a binary distribution of parameter p_H, so that the sum of the T observations of X_H, $Z_H = X_{H1} + \cdots + X_{HT}$, follows a binomial distribution $b(p_H, T)$.

Is opportune here to do an adaptation to the present context of the concept of *accurate estimate* of Finster (1987).

Definition 1 *Z_H/T is an "accurate estimate" of p_H with accuracy A and confidence $1-\alpha$ $(0 < \alpha < 1)$, if*

$$\Pr\left[\left|\frac{Z_H}{T} - p_H\right| < A\right] \geq 1 - \alpha \tag{4}$$

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

In what follows, we will try to determine which is the minimum number of replications in order to get an estimate of p_H with fixed precision A and confidence $1-\alpha$.

The theorem of Moivre (see for example Fz. de Trocóniz (1993)) shows that the succession $b(p_H,1), b(p_H,2), \ldots, b(p_H,T), \ldots$ is asymptotically normal $N(T\,p_H, T\,p_H[1-p_H])$ so that if $T\,p_H > 18$ the following approach to the distribution of Z_H is taken like valid:

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

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 $t_{\frac{\alpha}{2}}$ is the $\alpha/2$ quantile of the N(0,1) distribution,

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

from here we get that an interval of approximate confidence $1-\alpha$ for the probability p_H is

$$\left[\frac{Z_H}{T} - t_{\frac{\alpha}{2}}\sqrt{\frac{p_H(1-p_H)}{T}} \;\;,\;\; \frac{Z_H}{T} + t_{\frac{\alpha}{2}}\sqrt{\frac{p_H(1-p_H)}{T}}\right] \tag{8}$$

or expressing it in another way,

$$\Pr\left[\left|\frac{Z_H}{T} - p_H\right| < t_{\frac{\alpha}{2}}\sqrt{\frac{p_H(1-p_H)}{T}}\right] \simeq 1-\alpha \tag{9}$$

Comparing (9) with (4) we can appreciate that $t_{\frac{\alpha}{2}}\sqrt{\frac{p_H(1-p_H)}{T}}$ plays in this expression the role of the accuracy A. This proportions a way of relating the number of replications with the accuracy:

$$A = t_{\frac{\alpha}{2}}\sqrt{\frac{p_H(1-p_H)}{T}} \tag{10}$$

Therefore, in order to obtain an estimate of p_H with a prefixed accuracy A at level of confidence $1-\alpha$, the minimum number of replications that we have to produce is:

$$T = \frac{t^2_{\frac{\alpha}{2}} p_H(1-p_H)}{A^2} \tag{11}$$

Next, we see some examples on the utilization of these two last formulas.

Example 1 *It is wanted to approach with confidence 99% the right tail of probability 0.05 of an unknown distribution, with accuracy 0.005. That is to say, it is desired to approach the tail of probability 0.05 by means of the relative frequency, with a number of replications such that makes that with a 99% of confidence the theoretical probability of that tail meets in the interval* $[0.045 \quad 0.055]$*, which is the minimum number of replications needed?*

$$T = \frac{t^2_{\frac{\alpha}{2}} p_H (1 - p_H)}{A^2} = \frac{2.57^2 \; 0.05 \; (1 - 0.05)}{0.005^2} = \frac{0.313732}{0.000025} = 12,549.31 \tag{12}$$

Starting from $T \geq 12,550$ *will be gotten an accurate estimate of* p_H *in accordance with definition 1.*

Example 2 *It has been approached an interval H of probability* $p_H = 0.1$ *by means of an empirical distribution calculated with 1,000 replications. At 99% of confidence, Which will it be the gain in accuracy if does the number of replications be duplicated?*

$$A_{1,000} = 2.57\sqrt{\frac{0.1 \;\; 0.9}{1,000}} = 0.02438 \tag{13}$$

$$A_{2,000} = 2.57\sqrt{\frac{0.1 \;\; 0.9}{2,000}} = 0.01724 \tag{14}$$

Gain in the level of accuracy: $A_{1,000} - A_{2,000} = 0.00714$

Although the application of equations (10) and (11) is frankly simple, it in the practice could be useful observe the table 1 and figure 1, that has been gotten from them.

5 Methods of application.

These equations suggest different strategies of performance relying on the focus that is desired to give up to each problem. In this section, it will settle down the form of being confronted with three of them: first, the focus —that we could call basic— that corresponds to the case in that the interest is centered in determining the necessary number of replications for getting an estimate of the probability p_H with accuracy A; second, the case in that we want

to estimate with accuracy ε a critical value of a distribution, that is to say, the value of X that has associated a probability $1 - p_X$ in its function of distribution, in this case the accuracy is defined on the values of X, not on probabilities; and third will be studied the form of selecting the number of replications in order to carry out a check on the power of a test.

5.1 Basic focus.

If it is wanted to establish the minimum number of replications necessary in order to reach an accuracy A in the estimate of p_H, the method to follow could be the following:

1. In first place determine the level of confidence $1-\alpha$ and the degree of accuracy A that you want to reach in the approach by the method of Monte Carlo of the quantile of probability p_H.

2. With the values so determined, apply the formula (11) in order to get the minimum number of replications that will reach the accuracy A.

3. Utilize, in the process of simulation, a number of replications greater or equal to the obtained in the previous stage.

5.2 Accuracy defined on X.

If you want to estimate with accuracy ε the value of X that has associated a probability $1 - p_X$ in its theoretical function of distribution, one could use a method in two stages like the following for it:

1. Utilize the method described in the section 5.1 in order to determine the necessary number of replications for estimate the probability p_X with fixed accuracy A and level of confidence $1-\alpha$. With a number of replications equal or greater than the determined by the equation (11) simulate the distribution of probabilities of the X variable. In this distribution, search for the probability assigned to the values $X - \varepsilon$ and $X + \varepsilon$, that we will denominate $1 - \hat{p}_{X-\varepsilon}$ and $1 - \hat{p}_{X+\varepsilon}$.

2. Repeat the method of the basic focus in order to determine the necessary number of replications for estimating the probability p_X with accuracy

$$A = min(p_X - \hat{p}_{X+\varepsilon},\ \hat{p}_{X-\varepsilon} - p_X).$$

This will determine an accuracy approximately ε in the estimate of the value of X that has an associated probability $1 - p_X$.

5.3 Power of a test.

If you want to check the power of a test —based on a statistic with a well-known or unknown distribution under the null hypothesis— the method could be the following:

1. Fix the critical value, X_{VC}, corresponding to the significance level that is wanted, on the distribution of the statistic under the null hypothesis.

2. Achieve an arbitrary number of replications, for example 5,000, of the statistic under the alternative hypothesis. On the distribution of frequencies so obtained calculate the probability,

$$\Pr(X > X_{VC}/H_a) = \hat{p}_{X_{VC}}$$

3. Utilize the method described in the basic focus upon determining the necessary number of replications for getting an accuracy A in the estimate of $\hat{p}_{X_{VC}}$ with confidence $1-\alpha$.

6 Conclusions.

Although *a priori* it seems impossible have any knowledge about the error that is produced upon approaching the quantiles of an unknown distribution by simulation, in this work have been substantiated that the theory on the binomial distribution could contribute information to this reference.

This theory allows to establish a relationship between the accuracy with which are estimated or approached the quantiles of the distribution and the minimum number of replications that one must produce in order to get that accuracy.

Table 1: Minimum number of replications in order to get an approximation of p_H of accuracy A to the level of confidence $1-\alpha$.

Table 1a

$p_H = 0,1$			
	$1-\alpha$		
A	0.90	0.95	0.99
0.0250000	390	553	955
0.0244362	408	579	1,000
0.0200000	609	864	1,493
0.0185942	704	1,000	1,727
0.0172790	816	1,158	2,000
0.0156049	1,000	1,420	2,452
0.0150000	1,082	1,537	2,654
0.0141082	1,223	1,737	3,000
0.0131481	1,409	2,000	3,454
0.0110343	2,000	2,840	4,904
0.0109282	2,039	2,895	5,000
0.0107354	2,113	3,000	5,181
0.0100000	2,435	3,457	5,971
0.0090095	3,000	4,259	7,356
0.0083156	3,522	5,000	8,635
0.0077274	4,078	5,790	10,000
0.0069787	5,000	7,099	12,261
0.0058800	7,043	10,000	17,271
0.0054641	8,156	11,580	20,000
0.0050000	9,741	13,830	23,885
0.0049347	10,000	14,198	24,521
0.0041578	14,086	20,000	34,542
0.0040000	15,220	21,609	37,320
0.0034894	20,000	28,396	49,043
0.0030000	27,057	38,416	66,347
0.0020000	60,878	86,436	149,282
0.0010000	243,513	345,744	597,127

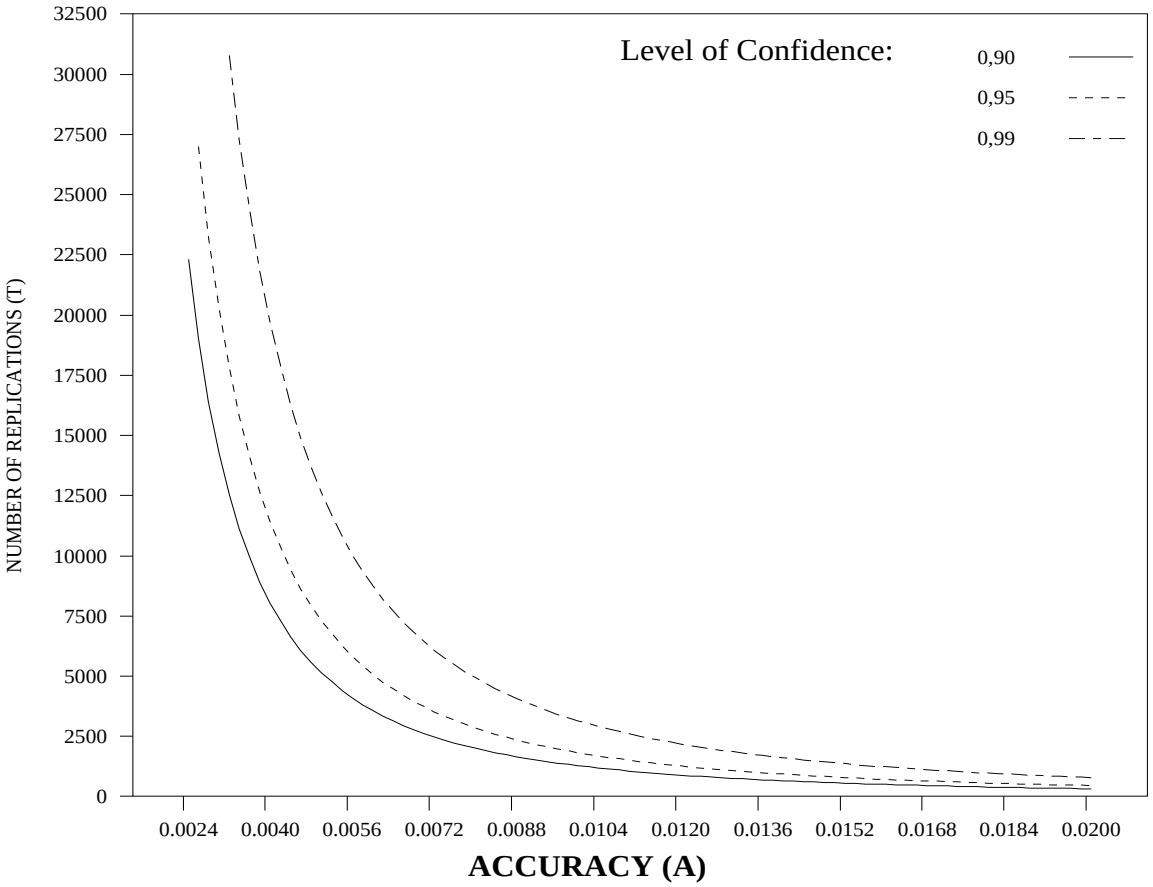
Table 1b

$p_H = 0,05$			
	$1-\alpha$		
A	0.90	0.95	0.99
0.0250000	206	292	504
0.0200000	321	456	788
0.0177525	408	579	1,000
0.0150000	571	811	1,401
0.0135084	704	1,000	1,727
0.0125529	816	1,158	2,000
0.0113367	1,000	1,420	2,452
0.0102494	1,223	1,737	3,000
0.0100000	1,285	1,825	3,152
0.0095519	1,409	2,000	3,454
0.0080163	2,000	2,840	4,904
0.0079391	2,039	2,895	5,000
0.0077991	2,113	3,000	5,181
0.0065452	3,000	4,259	7,356
0.0060411	3,522	5,000	8,635
0.0056138	4,078	5,790	10,000
0.0050699	5,000	7,099	12,261
0.0050000	5,141	7,299	12,606
0.0042717	7,043	10,000	17,271
0.0040000	8,033	11,405	19,697
0.0039696	8,156	11,580	20,000
0.0035850	10,000	14,198	24,521
0.0030206	14,086	20,000	34,542
0.0030000	14,280	20,275	35,017
0.0025350	20,000	28,396	49,043
0.0020000	32,130	45,619	78,788
0.0010000	128,521	182,476	315,150

Table 1c

$p_H = 0,01$			
	$1-\alpha$		
A	0.90	0.95	0.99
0.0100000	268	380	657
0.0090000	331	470	811
0.0081046	408	579	1,000
0.0061670	704	1,000	1,727
0.0057308	816	1,158	2,000
0.0051756	1,000	1,420	2,452
0.0050000	1,071	1,521	2,627
0.0046792	1,223	1,737	3,000
0.0043607	1,409	2,000	3,454
0.0040000	1,674	2,377	4,105
0.0036597	2,000	2,840	4,904
0.0036245	2,039	2,895	5,000
0.0035605	2,113	3,000	5,181
0.0030000	2,976	4,226	7,298
0.0029881	3,000	4,259	7,356
0.0027580	3,522	5,000	8,635
0.0025629	4,078	5,790	10,000
0.0023146	5,000	7,099	12,261
0.0020000	6,697	9,508	16,421
0.0019502	7,043	10,000	17,271
0.0018122	8,156	11,580	20,000
0.0016367	10,000	14,198	24,521
0.0013790	14,086	20,000	34,542
0.0011573	20,000	28,396	49,043
0.0010000	26,786	38,032	65,684
0.0005000	107,146	152,127	262,736

Figure 1: Relation A-T for each level of confidence with $p_H = 0.05$.



References

Beaulieu, J. & Miron, J. (1993), 'Seasonal unit roots in aggregate u.s. data', *Journal of Econometrics* **54**, 305–28.

Dickey, D. & Fuller, W. (1979), 'Distribution of the estimators for autoregressive time series with a unit root', *Journal of the American Statistical Association* **74**, 427–31.

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

Dickey, D., Hasza, D. & Fuller, W. (1984), 'Testing for unit roots in seeasonal time series', *Journal of American Statistical Association* **79**, 355–67.

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.

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

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

Phillips, P. & Perron, P. (1988), 'Testing for a unit root in time series regression', *Biometrika* **75**, 335–46.