ESTADÍSTICA ESPAÑOLA
Vol. 37, Núm. 140, 1995, págs. 497 a 509

Selección del número de replicaciones en un estudio de simulación

por
IGNACIO DÍAZ-EMPARANZA
Departamento de Econometría y Estadística
Universidad del País Vasco - Euskal Herriko Unibertsitatea
Avda. Lehendakari Aguirre, 83.
E48015 BILBAO Spain
e-mail: etpdihei@bs.ehu.es

RESUMEN

Para aproximar una distribución de probabilidades por medio de simulación es necesario determinar un número de replicaciones. La precisión con que se estima la distribución dependerá de dicho número de replicaciones. En este trabajo se obtiene una relación entre el número de replicaciones y la precisión de la estimación, de manera que si se desea obtener un valor prefijado para la precisión es posible determinar cuál será el mínimo número de replicaciones necesario para ello.

Palabras clave: Número de replicaciones, Monte-Carlo, precisión, error de simulación.

Clasificación AMS: 62E25

1. INTRODUCCIÓN

La alta capacidad de cálculo de los ordenadores en los últimos años, está permitiendo solucionar algunos problemas estadísticos que antes parecían irresolubles. En la literatura estadística y econométrica es frecuente encontrar trabajos en los que la distribución de probabilidades de determinado estadístico, que parece imposible ser desarrollada analíticamente, se obtiene de una forma empírica mediante el uso de simulaciones por ordenador. Por ejemplo, este tipo de práctica es muy usual en los artículos que tratan sobre procesos no estacionarios con raíces unitarias (Ver p. ej: Dickey y Fuller (1981), Dickey, Hasza y Fuller (1984), Hylleberg, Engle, Granger y Yoo (1990) o Beaulieu y Miron (1993). También se utilizan simulaciones por ordenador para comprobar la potencia de determinados estadísticos ante hipótesis alternativas que implican una distribución desconocida, por ejemplo: Dickey y Fuller (1979), Phillips y Perron (1988)

Cuando el analista se enfrenta a un problema concreto en el que necesita disponer de la distribución de probabilidades –aunque sea aproximada– de un determinado estadístico, una de las decisiones que ha de tomar, si decide calcularla por medio de simulación, se refiere al número de replicaciones que ha de realizar. Cualquier analista sabe que el número óptimo de replicaciones a utilizar al hallar una distribución empírica por medio de simulaciones es infinito, pero evidentemente, en la práctica es imposible trabajar con series de datos de infinitas observaciones.

Como cualquiera puede imaginarse, el error que se comete al aproximar una distribución mediante simulación es inversamente proporcional al número de replicaciones realizadas. Por eso, si se quiere minimizar el error, el número de replicaciones ha de ser lo más alto posible. Sin embargo, en la mayoría de los trabajos que utilizan estos métodos, el número de replicaciones se elige de una forma un tanto arbitraria, sin tener una idea sobre la precisión que se consigue al estimar la distribución de probabilidades. Esto puede verse, por ejemplo, en cualquiera de los artículos referenciados en el primer párrafo. De manera que sea cual fuere el problema concreto que se estudia, uno siempre puede preguntarse las mismas cuestiones:

– ¿Habría una ganancia significativa –en términos de precisión con que se aproxima la distribución– si se realizan, digamos, 1.000 replicaciones más?

– Dado que, a veces, el cálculo de una sóla replicación puede ser lento y costoso ¿Será suficiente considerar, por ejemplo, 1.000 replicaciones?

Ayudar a responder a estas dos preguntas es el objetivo fundamental del presente trabajo.

2. APROXIMACIÓN EMPÍRICA A LA DISTRIBUCIÓN TEÓRICA

Supongamos que se dispone de una muestra de tamaño N del vector de variables y que es de dimensión P. Supondremos también que la distribución de probabilidades de y –sea cual fuere– es conocida. Sea Y la matriz (NxP) que contiene en cada columna las N observaciones de cada uno de los componentes de y, y f una función –que se suele denominar *estadístico*– tal que a cada valor de Y le hace corresponder un valor X real, es decir,

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

La distribución de probabilidades de X es, en general, desconocida. A continuación se estudiará el problema de hallar su aproximación mediante el método de Monte Carlo.

La forma usual de aproximar una distribución de probabilidades mediante el método de Monte Carlo es la siguiente:

1. En primer lugar se generan mediante ordenador T muestras distintas (se suelen denominar *replicaciones*) de tamaño N para el vector y, a partir de su distribución teórica que es conocida.

2. Para cada una de las replicaciones se calcula el valor que toma el estadístico f: $X_t = f(Y_t)$, donde Y_t es el valor simulado de la matriz Y en la replicación t-ésima y X_t es el valor obtenido para el estadístico en dicha replicación, con t=1,...,T.

3. Los valores recogidos de X_1 ,..., X_T se ordenan y su distribución de frecuencias relativas se toma como aproximación de la función de densidad, que era desconocida.

A partir de la distribución de frecuencias relativas se calculan intervalos de confianza y se realizan contrastes de hipótesis como si ésta fuera la distribución teórica.

3. PRECISIÓN DE LA APROXIMACIÓN EMPÍRICA

Sea H un intervalo cualquiera definido sobre la recta real. Definiremos ahora una variable ficticia, X_H , de la siguiente forma:

$$X_{H\,t} = \begin{cases} 1 & \text{Si } X_t \in H \\ 0 & \text{Si } X_t \notin H \end{cases}$$

De manera que cada observación de X_t lleva asociada una observación –con valor 0 ó 1– de la variable X_{Ht}. La función de densidad teórica –desconocida– de X_t asigna una probabilidad p_H al intervalo H. Esto significa que

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

Producir T replicaciones del vector y, implica disponer de una muestra de T "observaciones" de la variable real X. Esta muestra lleva asociada, a su vez, una muestra de tamaño T de la variable X_H. Esta variable sigue una distribución binaria de parámetro p_H, así que la suma de las T observaciones de X_H, $Z_H = X_{H1} + \ldots + X_{HT}$, sigue una distribución binomial $b(p_H, T)$.

Es oportuno aquí hacer una adaptación al presente contexto del concepto de *estimación precisa* de Finster (1987)

Definición 1. Z_H/T es una *estimación precisa* de p_H con nivel de imprecisión A y confianza $1-\alpha$ (con $0 < \alpha < 1$), si

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

El conjunto de precisión [-A, A] es el conjunto de errores de simulación aceptables.

En lo que sigue a continuación se intentará determinar cuál es el número de replicaciones mínimo para obtener una estimación de p_H con nivel de imprecisión fijo A y confianza $1-\alpha$.

El teorema de Moivre (ver por ejemplo Fz. de Trocóniz 1993) prueba que la sucesión $b(p_H,1)$, $b(p_H,2),\ldots,b(p_H,T)$, es asintóticamente normal $N(T\,p_H, T\,p_H[1-p_H])$ de manera que si $T\,p_H > 18$ se suele tomar como válida la siguiente aproximación a la distribución de Z_H:

$$Z_H \approx N(T \cdot p_H, T \cdot p_H(1-p_H)) \qquad [2]$$

entonces, para la frecuencia binomial, Z_H/T, se tiene

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

Si $t_{\alpha/2}$ es el cuantil $\alpha/2$ correspondiente a la cola derecha de la distribución N(0,1),

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

de aquí se obtiene que un intervalo de confianza aproximada 1-α para la probabilidad p_H es

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

o expresándolo de otra forma,

$$\Pr\left[\left|\frac{Z_H}{T} - p_H\right| < t_{\frac{\alpha}{2}}\sqrt{\frac{p_H(1-p_H)}{T}}\right] \simeq 1-\alpha \qquad [3]$$

Comparando la expresión (1) con la (3) se aprecia que la parte derecha de la desigualdad juega aquí el papel del nivel de imprecisión *A*. Esto proporciona una forma de relacionar el número de replicaciones con el nivel de imprecisión:

$$A = t_{\frac{\alpha}{2}}\sqrt{\frac{p_H(1-p_H)}{T}} \qquad [4]$$

Por tanto, para obtener una estimación de p_H con un nivel de imprecisión prefijado *A*, a un nivel de confianza 1-α, el número mínimo de replicaciones que ha de producirse es:

$$T = \frac{t^2_{\frac{\alpha}{2}} p_H(1-p_H)}{A^2} \qquad [5]$$

Veamos a continuación algunos ejemplos sobre la utilización de estas dos últimas fórmulas.

Ejemplo 1. Se desea aproximar con confianza 99% la cola derecha de probabilidad 0,05 de una distribución desconocida, con nivel de imprecisión 0,005. Es decir, se desea aproximar la cola de probabilidad 0,05 mediante la frecuencia relativa, con un número de replicaciones tal que haga que con un 99% de confianza la probabilidad teórica de esa cola se encuentre en el intervalo [0,045 0,055]. ¿Cuál es el mínimo número de replicaciones necesario para ello?

$$T = \frac{t^2_{\frac{\alpha}{2}} p_H(1-p_H)}{A^2} = \frac{2{,}57^2 \cdot 0{,}05 \cdot (1-0{,}05)}{0{,}005^2} = \frac{0{,}313732}{0{,}000025} = 12.549{,}31$$

Utilizando $T \geq 12.550$ se obtendrá una estimación precisa de p_H de acuerdo con la definición 1.

Ejemplo 2. Se ha aproximado un intervalo H de probabilidad $p_H = 0{,}1$ por medio de una distribución empírica calculada con 1.000 replicaciones. Al 99% de confianza. ¿Cuál será la ganancia en precisión si se duplica el número de replicaciones?

$$A_{1.000} = 2{,}57\sqrt{\frac{0{,}1 \cdot 0{,}9}{1.000}} = 0{,}02438$$

$$A_{2.000} = 2{,}57\sqrt{\frac{0{,}1 \cdot 0{,}9}{2.000}} = 0{,}01724$$

Ganancia en precisión: $A_{1.000} - A_{2.000} = 0{,}0074$.

Aunque la aplicación de las ecuaciones (4) y (5) es francamente sencilla, en la práctica puede ser útil observar las tablas 1, 2 y 3 y la figura 1, que se han obtenido a partir de ellas.

4. MÉTODOS DE APLICACIÓN

Estas ecuaciones sugieren distintas estrategias de actuación dependiendo del enfoque que se desee dar al problema. En esta sección se establecerá la forma de enfrentarse a tres de ellos: en primer lugar el enfoque –que podríamos llamar básico– que corresponde al caso en que el interés se centra en determinar el número de replicaciones necesario para obtener una estimación de la probabilidad p_H con nivel de imprecisión A; en segundo lugar el caso en el que se desea estimar con imprecisión ε un valor crítico de una distribución, es decir, el valor de X que lleva asociada una probabilidad (1- p_X) en su función de distribución, en este caso la imprecisión se define sobre valores de X, no sobre probabilidades; en tercer lugar se estudiará la forma de seleccionar el número de replicaciones para realizar una prueba sobre la potencia de un contraste.

4.1. Enfoque básico

Si lo que se desea es establecer el número de replicaciones mínimo necesario para alcanzar una imprecisión menor o igual que A en la estimación de p_H el método a seguir puede ser el siguiente:

1. En primer lugar determinar el nivel de confianza $1-\alpha$ y el grado de imprecisión A que se quiere tolerar en la aproximación por el método de Monte Carlo del cuantil de probabilidad p_H.

2. Con los valores así determinados, aplicar la fórmula (5) para obtener el mínimo número de replicaciones con que se alcanzará la imprecisión A.

3. Utilizar en el proceso de simulación un número de replicaciones mayor o igual al obtenido en la etapa anterior.

4.2. Precisión definida sobre X

Si se desea estimar con imprecisión ε el valor de X que lleva asociado una probabilidad $(1- p_X)$ en su función de distribución teórica, se puede utilizar para ello un método en dos etapas como el siguiente:

1. Utilizar el método descrito en la sección anterior para determinar el número de replicaciones necesario para estimar la probabilidad p_X con imprecisión fija A y nivel de confianza $1-\alpha$.

2. Con un número de replicaciones igual o mayor al determinado por la ecuación (5) simular la distribución de probabilidades de la variable X. En dicha distribución, buscar la probabilidad asignada a los valores $X-\varepsilon$ y $X+\varepsilon$, que denominaremos $1-\hat{p}_{X-\varepsilon}$ y $1-\hat{p}_{X+\varepsilon}$.

3. Repetir el método del enfoque básico para determinar el número de replicaciones necesario para estimar la probabilidad p_X con imprecisión $A = \min(p_X - \hat{p}_{X+\varepsilon},\ \hat{p}_{X-\varepsilon} - p_X)$ este será, a su vez, el que determina aproximadamente una imprecisión ε en la estimación del valor de X que lleva asociado una probabilidad $(1- p_X)$.

4.3. Prueba de potencia

Si se desea hacer una prueba sobre la potencia de un contraste –basado en un estadístico de distribución conocida ó desconocida bajo la hipótesis nula– el método puede ser el siguiente:

1. Fijar el valor crítico, X_{VC}, correspondiente al nivel de significación que se desee, sobre la distribución del estadístico bajo la hipótesis nula (H_0).

2. Realizar un número de replicaciones arbitrario, por ejemplo 5.000, del estadístico bajo la hipótesis alternativa (H_a). Sobre la distribución de frecuencias así obtenida calcular la probabilidad que se asigna a X_{VC},

$$Pr(X < X_{VC} / H_a) = 1 - \hat{p}_{X_{vc}}$$

3. Utilizar el método descrito en el enfoque básico al determinar el número de replicaciones necesario para obtener una imprecisión A en la estimación de $\hat{p}_{X_{vc}}$ con confianza 1-α..

5. CONCLUSIONES

Aunque *a priori* parece imposible tener algún conocimiento sobre el error que se produce al aproximar los cuantiles de una distribución desconocida por medio de simulación, en este trabajo se ha comprobado que la teoría sobre la distribución binomial puede aportar información a este respecto. Esta teoría permite establecer una relación entre la imprecisión que se obtiene al estimar o aproximar los cuantiles de la distribución y el número de replicaciones mínimo que hay que producir para obtener esa imprecisión.

Tablas 1, 2 y 3: Número mínimo de replicaciones para obtener una aproximación de la probabilidad p_H con imprecisión A al nivel de confianza 1-α.

Tabla 1

$p_H = 0{,}1$	1-a		
A	0,9	0,95	0,99
0,025	390	553	955
0,0244362	408	579	1.000
0,02	609	864	1.493
0,0185942	704	1.000	1.727
0,017279	816	1.158	2.000
0,0156049	1.000	1.420	2.452
0,015	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
0,01	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
0,00588	7.043	10.000	17.271
0,0054641	8.156	11.580	20.000
0,005	9.741	13.830	23.885
0,0049347	10.000	14.198	24.521
0,0041578	14.086	20.000	34.542
0,004	15.220	21.609	37.320
0,0034894	20.000	28.396	49.043
0,003	27.057	38.416	66.347
0,002	60.878	86.436	149.282
0,001	243.513	345.744	597.127

Tabla 2

$p_H = 0{,}05$	1-a		
A	0,9	0,95	0,99
0,025	--(*)	--	504
0,02	--	456	788
0,0177525	408	579	1.000
0,015	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
0,01	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
0,005	5.141	7.299	12.606
0,0042717	7.043	10.000	17.271
0,004	8.033	11.405	19.697
0,0039696	8.156	11.580	20.000
0,003585	10.000	14.198	24.521
0,0030206	14.086	20.000	34.542
0,003	14.280	20.275	35.017
0,002535	20.000	28.396	49.043
0,002	32.130	45.619	78.788
0,001	128.521	182.476	315.150

(*) La aproximación (2) sólo es válida si T>360.

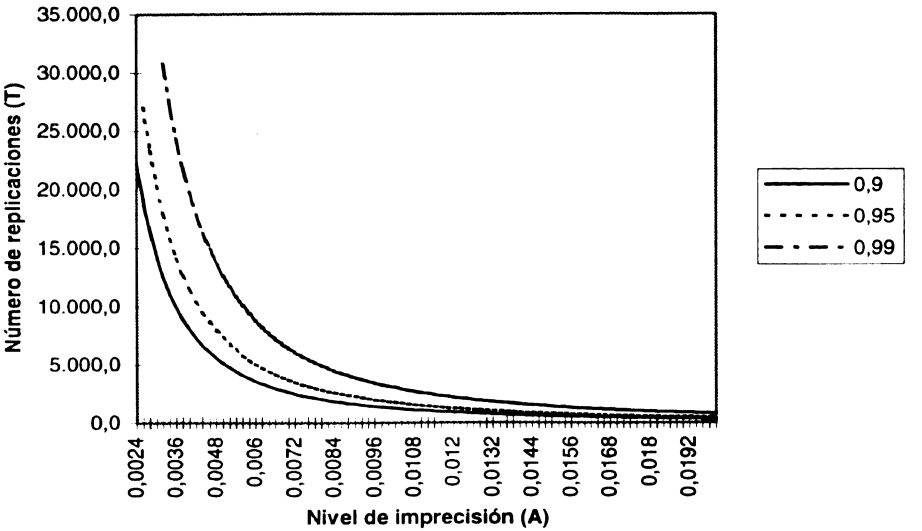
Tabla 3

$p_H = 0{,}01$	1-a		
A	0,9	0,95	0,99
0,005731	--(*)	--	2.000
0,005176	--	--	2.452
0,005	--	--	2.627
0,004679	--	--	3.000
0,004361	--	2.000	3.454
0,004	--	2.377	4.105
0,00366	2.000	2.840	4.904
0,003625	2.039	2.895	5.000
0,003561	2.113	3.000	5.181
0,003	2.976	4.226	7.298
0,002988	3.000	4.259	7.356
0,002758	3.522	5.000	8.635
0,002563	4.078	5.790	10.000
0,002315	5.000	7.099	12.261
0,002	6.697	9.508	16.421
0,00195	7.043	10.000	17.271
0,001812	8.156	11.580	20.000
0,001637	10.000	14.198	24.521
0,001379	14.086	20.000	34.542
0,001157	20.000	28.396	49.043
0,001	26.786	38.032	65.684
0,0005	107.146	152.127	262.736

(*) La aproximación (2) sólo es válida para T≥2.000

Figura 1

RELACIÓN A-T PARA CADA NIVEL DE CONFIANZA (1-α) CON $p_H = 0,05$



REFERENCIAS

BEAULIEU, J y MIRÓN, J. (1993): «Seasonal unit roots in aggregat U.S. data». *Journal of Econometrics 54*, 305-328.

DICKEY, D., HASZA, D y FULLER, W. (1984): «Testing for unit roots in seasonal time series». *Journal of American Satatistical Association 79*, 355-367

DICKEY, D., Y FULLER, W. (1979): «Distribution of the estimators for autoregressive time series with a unit root». *Econometrica 49*, 1057-1071.

DICKEY, D., Y FULLER, W. (1981): «Likelihood ratio statistics for autoregressive time series with a unit root». *Journal of the American Statistical Association 74*, 427-431.

FINSTER, M.P. (1987): «An analysis of five simulation methods for determining the number of replications in a complex Monte Carlo study». *Statistics and Probability Letters 5*, 353-360

FZ. DE TROCÓNIZ, A. (1993): «Probabilidades, Estadística, Muestreo». Ed.. Tebar Flores.

HYLLEBERG, S., ENGLE, R., GRANGER, C. y YOO, B. (1990): «Seasonal integration and cointegration». *Journal of Econometrics 44*, 215-238.

PHILLIPS, P. y PERRON, P. (1988): «Testing for a unit root in time series regression». *Biometrika 75*, 335-346.

SELECTION OF THE NUMBER OF REPLICATIONS IN A STUDY ON SIMULATION

SUMMARY

To approach a distribution of probabilities by means of simulation, it is necessary to determine a number of replications. The accuracy of the distribution's estimate depends on this number of replications. This paper achieves a relation between the number of replications and the accuracy of the estimate, so that when the aim is to obtain a pre-fixed value for the accuracy, it is possible to determine the minimum number of replications needed.

Key words:Number of replications, Monte-Carlo, precision, simulation error

Classification AMS: 62E25