

USO DE LA SIMULACIÓN EN HOJA DE CÁLCULO COMO HERRAMIENTA PEDAGÓGICA PARA LA INFERENCIA ESTADÍSTICA: APLICACIÓN A LAS PROPIEDADES CLÁSICAS DE LOS ESTIMADORES

Carlos Martínez de Ibarreta Zorita
Departamento de Métodos Cuantitativos
Universidad Pontificia Comillas (ICADE) de Madrid

Resumen

En esta comunicación se presenta y describe una aplicación realizada en hoja de cálculo (Excel) que, mediante el uso de métodos de simulación tipo Montecarlo, ilustra y permite experimentar con algunas de las propiedades clásicas (insesgo, eficiencia, consistencia) de diversos estimadores de parámetros poblacionales, así como las características de su distribución muestral.

El interés pedagógico de esta aplicación, ampliamente utilizada por el autor en sus cursos de docencia universitaria, radica en los siguientes aspectos: (a) ofrecer una visión más intuitiva, aplicada y complementaria de algunos de los conceptos teóricos habitualmente enseñados en los cursos de inferencia estadística, lo que facilita su comprensión y asimilación por parte de los alumnos, (b) mostrar una visión introductoria de las técnicas de simulación como herramienta de investigación y análisis, (c) permitir un aprendizaje más activo del alumno en estas materias y, como objetivo de carácter secundario y transversal, (d) posibilitar la mejora en el uso de la hoja de cálculo por el alumno como herramienta avanzada para el diseño, planteamiento y solución de problemas de carácter estadístico.

La comunicación finaliza con posibles propuestas de extensión de esta aplicación a ámbitos diferentes al presentado.

Palabras clave: simulación Montecarlo, inferencia estadística, estimación de parámetros, propiedades de estimadores, insesgo, eficiencia, consistencia, hoja de cálculo.

1.- INTRODUCCIÓN

En muchas de las licenciaturas universitarias existentes en la actualidad aparece alguna asignatura en la que se aborda el estudio de las nociones básicas de la Inferencia Estadística. Este es el caso de la licenciatura en Administración y Dirección de Empresas (ADE) que se imparte en la Universidad Pontificia Comillas (ICADE), especialidad E-2 en la que tales conceptos aparecen en la asignatura denominada “Estadística Empresarial” de tercer curso.

La experiencia docente del autor constata la dificultad inicial que suelen tener los alumnos para la comprensión de los conceptos básicos. Las explicaciones teóricas complementadas con ejercicios tradicionales a veces no bastan para que los alumnos interioricen dichos conceptos y no se limiten a la memorización de definiciones y fórmulas, sin conseguir un aprendizaje efectivo. Por ejemplo, es frecuente que al preguntar a un alumno “¿Qué significa para Vd. que un estimador sea insesgado?”, su respuesta, si es un alumno aplicado, sea “*aquel cuya esperanza coincide con el parámetro al que estima*”, pero interrogado por lo que entiende de esa definición, no sea capaz de ir más allá.

En este punto, se plantea el uso pedagógico de ejercicios de simulación realizados en hoja de cálculo para intentar cubrir esas lagunas de aprendizaje y comprensión, tendiendo puentes entre la teoría y la práctica y como método complementario a la docencia tradicional.

El desarrollo de las sesiones de simulación en hoja de cálculo, en lugar de en otra aplicación estadística más específica, responde a diversas razones, entre las que pueden destacarse las siguientes: (a) ser un programa del que disponen la mayoría de ordenadores, (b) su facilidad de aprendizaje y uso, al menos en un nivel básico, (c) su gran flexibilidad para poder adaptar lo previamente diseñado para un problema a otros parecidos y finalmente, (d) por los beneficios añadidos que puede aportar a los alumnos un mayor conocimiento y dominio de esta herramienta ofimática tanto para otras asignaturas como para su futura vida laboral.

Por consiguiente, el interés pedagógico de las sesiones de simulación, ampliamente utilizadas por el autor en sus cursos de docencia universitaria, radica en los siguientes aspectos: (a) ofrecer una visión más intuitiva, aplicada y complementaria de algunos de los conceptos teóricos habitualmente enseñados en los cursos de inferencia estadística,

lo que facilita su comprensión y asimilación por parte de los alumnos, (b) mostrar una visión introductoria de las técnicas de simulación como herramienta de investigación y análisis, (c) permitir un aprendizaje más activo del alumno en estas materias y, como objetivo de carácter secundario y transversal, (d) posibilitar la mejora en el uso de la hoja de cálculo por el alumno como herramienta avanzada para el diseño, planteamiento y solución de problemas de carácter estadístico.

El resto de esta comunicación se estructura como sigue: en primer lugar se describen el planteamiento y los objetivos perseguidos de un problema concreto al que se van a aplicar las técnicas de simulación, en segundo lugar, se describe el diseño de la hoja de cálculo correspondiente, finalmente, se exponen las conclusiones obtenidas así como posibles propuestas de extensión de esta aplicación a ámbitos diferentes al presentado.

2.- PLANTEAMIENTO DE LA SESIÓN DE SIMULACIÓN

El ejemplo de simulación desarrollado en esta comunicación está centrado en el análisis empírico del desempeño de dos diferentes estimadores de la media de una población normal con varianza conocida, a través de la estimación del mismo en un número muy grande de muestras aleatorias simuladas en hoja de cálculo.

El objetivo pretendido consiste en comparar ambos estimadores respecto de algunas de las propiedades clásicas de los estimadores: insesgo, eficiencia (relativa) y consistencia, así como comparar de forma gráfica las distribuciones empíricas de ambos, como aproximación a sus distribuciones de probabilidad teóricas.

Los estimadores propuestos son la media muestral, a_x , y un estimador *naive*, que será denominado h_x , definido como la semisuma del primer y el último valor de la muestra,

es decir:
$$h_x = \frac{x_1 + x_n}{2}$$

Se han considerado dos razones para la elección de estos dos estimadores:

- 1) La existencia de grandes diferencias respecto a su desempeño. Si bien ambos son estimadores insesgados, la media muestral presenta una varianza mucho menor que h_x , siendo por tanto eficiente en términos relativos y además, h_x no es un estimador consistente de la media poblacional mientras que a_x sí.

- 2) La deducción teórica de todos los resultados es bastante sencilla, pudiendo ser realizada previamente incluso por los propios alumnos.

En la hoja de cálculo Excel, una vez fijados los valores de los parámetros poblacionales, se generan muestras aleatorias de diferentes tamaños, en este caso de tamaños 10, 20 y 100 procedentes de dicha población. Para cada una de ellas se calculan los valores de cada uno de los dos estimadores. Tras obtener los valores correspondientes a un número elevado de muestras (se han generado 10000 para cada uno de los tamaños muestrales considerados), se procede al cálculo de sus valores resumen y a la representación gráfica de la distribución empírica de ambos estimadores.

3.- DISEÑO DE LA HOJA DE CÁLCULO

En primer lugar, se establecen en las celdas correspondientes, los valores de los parámetros poblacionales: media y desviación típica. En este caso, y con el fin de que el ejemplo no resulte demasiado abstracto para los alumnos, se ha pensado que la población represente el “*peso de una naranja*”, fijándose un valor medio de 120 gramos y una desviación típica de 20 gramos.

Seguidamente se procede a la generación de una muestra aleatoria de tamaño 10, otra de tamaño 20 y una tercera de tamaño 100. Esto se realiza con facilidad usando las funciones estadísticas de las que dispone la hoja de cálculo. Por una parte, la función ALEATORIO() proporciona un número pseudoaleatorio siguiendo la ley uniforme $\{0,1\}$, y por otra, la función DISTR.NORM.INV(*probabilidad acumulada; media; desviación típica*) permite obtener el valor de una distribución normal con valores paramétricos cualesquiera que acumula una cierta probabilidad. La combinación de ambas funciones permite obtener cualquier número aleatorio normal en una celda. Esa fórmula copiada al resto de celdas permite obtener una muestra simulada.


Hay que señalar que la función ALEATORIO() de la hoja de cálculo Excel es de carácter volátil, esto es, su valor es diferente cada vez que se emplea, y además, siempre que se introduce cualquier cambio en la hoja, todas las celdas que dependan de esta función se recalculan. Esta característica, que puede ser engorrosa en ciertos momentos, en especial cuando se desea conservar algún resultado, es la llave que permite realizar la

simulación tipo Montecarlo, al permitir generar diferentes muestras simuladas mediante un simple recálculo de la hoja¹.

Una vez generada una muestra de cada tamaño, se calculan en distintas celdas los valores de las estimaciones realizadas por los dos estimadores.

La Figura 1 muestra la organización de todo lo comentado hasta ahora.

Figura 1. Diseño de la hoja para generar muestras aleatorias.

	A	B	C	D	E	F	G	H	I	J	
1					POBLACIÓN						
2					Peso aleatorio de las naranjas (gr)						
3					distribución NORMAL						
4					MEDIA	120					
5					DESV.TÍPICA	20					
6											
7									estimadores de la media poblacional		
8											
9									n=10	n=20	n=100
10							ax		120,84	123,67	121,50
11						MUESTRA (m.a.s)	h		137,72	117,93	127,40
12											
13					=DISTR.NORM.INV(ALEATORIO(),\$F\$4,\$F\$5)					115,75	
14					113,76	152,42	115,54	97,60	95,46		
15					117,40	125,78	105,69	122,39	117,32		
16					96,61	132,49	132,30	113,57	110,27		
17					123,31	132,06	114,18	124,69	125,58		
18					154,02	128,65	147,93	159,23	130,36		
19					112,46	118,30	122,55	142,90	135,88		
20					129,11	106,86	144,46	90,62	127,91		
21					86,29	121,09	96,84	159,33	115,95		
22					141,10	101,53	123,01	109,12	119,52		
23					155,01	114,81	107,12	161,40	116,05		
24					99,04	130,14	79,87	107,13	116,26		

Si se ha usado la opción de bloquear el cálculo, basta presionar la tecla de función F9 para obtener muestras diferentes y, por consiguiente, nuevos valores de estimación. Parece claro que si los valores de los estimadores se van conservando de alguna forma y se repite este proceso muchas veces, se acabará teniendo una distribución empírica de valores de los mismos que estará cercana a la forma de su distribución teórica de probabilidad.

Como la repetición de este proceso de forma manual es inabordable para un número de iteraciones alto, se ha programado una sencilla macro en el lenguaje de programación Visual Basic que la hoja incorpora, para que todo esto se haga de forma automática. En la Figura 2 se muestra la programación usada para este caso concreto. Básicamente lo

¹ Es posible evitar el recálculo no deseado si en el menú Herramientas – Opciones, en la ventana Calcular se elige el cálculo manual. En este caso, sólo se realizan cálculos en la hoja al presionar la tecla de función F9.

que hace esta macro es realizar 10000 veces la misma operación: recalcular la hoja y copiar los valores de las estimaciones en la columna correspondiente a los resultados de cada estimador, cada vez una fila más abajo. Se ha añadido un contador como mecanismo de control del funcionamiento de la macro.

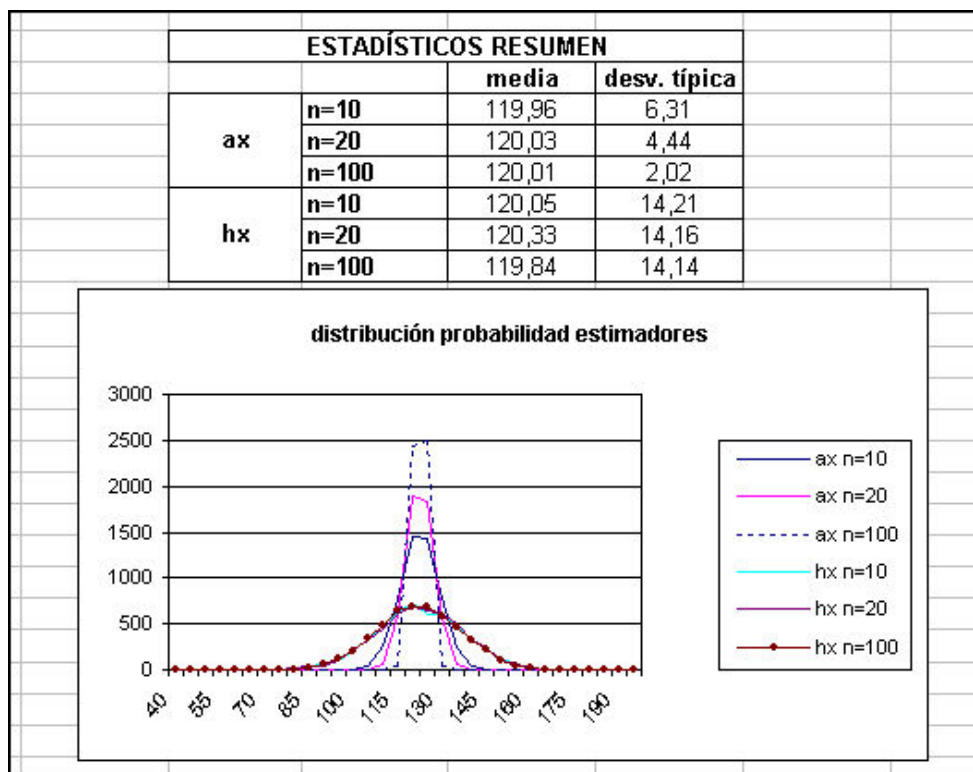
Figura 2. Programación de la macro para realizar las iteraciones.

```
Sub naranjas()  
'  
' naranjas Macro  
' Macro grabada el 10/05/2004 por Carlos  
'  
For k = 1 To 10000  
Calculate  
Cells(k + 2, 12).Value = [h10]  
Cells(k + 2, 13).Value = [i10]  
Cells(k + 2, 14).Value = [j10]  
Cells(k + 2, 15).Value = [h11]  
Cells(k + 2, 16).Value = [i11]  
Cells(k + 2, 17).Value = [j11]  
Cells(8, 18).Value = [k]  
Next k  
  
End Sub
```

Finalmente, se ha elaborado la tabla de frecuencias para cada uno de los estimadores, junto con sus estadísticos resumen (media y desviación típica), así como su representación gráfica. Hay que señalar que la tabla de frecuencias y el gráfico están elaborados con intervalos vinculados a los valores especificados en cada caso para los parámetros poblacionales, de forma que queden centrados en torno a la media de las distribuciones.

La representación gráfica final tras las 10000 iteraciones aparece reflejada en la Figura 3. Se ha considerado conveniente representar las seis distribuciones muestrales en el mismo gráfico, para poder compararlas entre sí, y poder alcanzar más fácilmente los objetivos pedagógicos propuestos.

Figura 3. Resultados de la simulación.



Finalmente, conviene destacar que se ha considerado más didáctico realizar esta sesión de forma que el gráfico de las distribuciones de frecuencias se vaya rehaciendo a medida que se van realizando las iteraciones, en lugar de realizar primero todos los cálculos y posteriormente el gráfico. Esta opción, no obstante, consume más recursos de ordenador y puede hacer que todo el proceso lleve bastante más tiempo, sobre todo en equipos no demasiado potentes.

4.- CONCLUSIONES Y POSIBLES EXTENSIONES

De los resultados obtenidos en esta sesión de simulación, parece interesante destacar a efectos pedagógicos las siguientes cuestiones:

- a) El hecho de que los estimadores, como estadísticos, *i.e.* funciones de la muestra, son aleatorios, es decir, cada muestra diferente da una estimación diferente de un único parámetro poblacional.
- b) La noción intuitiva de estimador insesgado, en el sentido de que la media de “muchísimas” estimaciones coincide aproximadamente con el

verdadero valor del parámetro. En este caso, los dos estimadores propuestos cumplen dicha propiedad.

- c) La noción intuitiva de eficiencia relativa. Entre dos estimadores insesgados, será eficiente en términos relativos aquel cuya distribución presente una menor variabilidad en torno a la media, lo que indica que va a ser menos probable que en otros obtener estimaciones que se alejen demasiado en uno o en otro sentido del verdadero valor del parámetro poblacional. En la sesión planteada se aprecia que para cualquiera de los tres tamaños muestrales considerados, la dispersión de la distribución de valores generados de h_x es muy superior a la de la media muestral, tal y como puede apreciarse muy claramente en las representaciones gráficas.
- d) La noción intuitiva de estimador consistente, en el sentido de que, la distribución muestral del estimador va teniendo menor dispersión respecto de la media a medida que el tamaño muestral va siendo cada vez mayor. Conviene destacar que tener una muestra mayor es disponer de mayor información potencial para poder estimar el parámetro poblacional, sin embargo, el estimador h_x no hace uso de dicha mayor información, pues sea cual sea el tamaño muestral sólo utiliza dos elementos muestrales. La media muestral a_x por el contrario, si aprovecha dicha mayor información.

Como posibles extensiones de esta aplicación, se sugieren entre otras las siguientes variantes: distribución muestral de otros estadísticos o estimadores de distribución conocida o no, realización de contrastes de hipótesis, obteniendo niveles de significación empíricos o construcción de curvas de potencia de contraste, etc.

La experiencia docente del autor muestra que, una vez realizada por el profesor una sesión de simulación, es posible encargar a los alumnos la realización como trabajo personal (de carácter voluntario, obviamente) de algunas de estas extensiones sugeridas.

Los resultados obtenidos son altamente positivos en general: además de servir para una comprensión más profunda y complementaria de los conceptos teóricos, se desarrolla el pensamiento científico y analítico y se mejoran las capacidades de modelización de problemas.

REFERENCIAS BIBLIOGRÁFICAS

PELOSI, M.K. y SANDIFER, T.M. (2000), “ Doing statistics for business with Excel”. Ed Wiley.