

# Métodos autosuficientes de estimación y contraste de hipótesis. Utilización de la simulación y el método de Monte Carlo en bioestadística

Preparado por Luis M. Molinero (Alce Ingeniería)

CorreoE: [bioestadistica@alceingenieria.net](mailto:bioestadistica@alceingenieria.net)

[Artículo en formato PDF](#)

Septiembre 2002

[www.seh-lelha.org/stat1.htm](http://www.seh-lelha.org/stat1.htm)

En el artículo de este mes vamos a presentar otras técnicas estadísticas de estimación y contraste de hipótesis, diferentes de la archiconocida metodología paramétrica y no paramétrica (aunque ésta última quizás algo menos utilizada). La idea es ampliar la cultura estadística del lector que desconoce su existencia, o que ocasionalmente alguna vez lo leyó en algún artículo y le sonó a chino, y recordar las ideas a aquellos que ya las conocen, aunque solo sea porque es una saludable inquietud intelectual mirar las cosas desde una nueva óptica.

Recordemos que los dos objetivos principales de la estadística son, por un lado la **estimación** de un parámetro de una población, a partir de los valores obtenidos en una muestra, y por otro el **contraste de hipótesis** a partir de los resultados observados en una muestra.

Plantaremos hoy otro enfoque distinto del habitual a ese tipo de problemas, y para ello comenzaremos sirviéndonos de un pequeño ejemplo.

En la siguiente tabla se presenta el valor de la PAS (media de tres lecturas) en dos grupos de pacientes diabéticos, con 10 pacientes por grupo

<b>A</b>	134	155	129	135	117	118	136	139	126	129
<b>B</b>	147	118	139	120	152	163	145	153	142	157

	<b>Medias</b>
<b>A</b>	131.8
<b>B</b>	143.6
<b>Diferencia</b>	-11.8

Para comparar ambos grupos la metodología tradicional se basa en calcular la probabilidad de observar una diferencia tan grande o mayor que la actualmente encontrada (-11.8), suponiendo que ambas muestras se hayan extraído aleatoriamente de la misma población. Es lo que se conoce como hipótesis nula. Para ello suponemos una distribución de probabilidad para los datos, la distribución normal o de Gauss. En realidad al no conocer tampoco la desviación típica de la población y estimarla también a partir de los datos de ambas muestras, la distribución de probabilidad que utilizamos es la **t de Student**, y a partir de ella determinaremos

que la probabilidad de una diferencia igual o más extrema que la observada es 0.058. Cuando esta probabilidad es muy pequeña (tradicionalmente menor que 0.05), en lugar de aceptar esa hipótesis de igualdad –muestras extraídas de la misma población– y por lo tanto obtención de un resultado improbable en nuestro estudio, preferimos rechazarla y concluir que ambas muestras no proceden de la misma población, o lo que es equivalente, que la diferencia obtenida no parece razonable atribuirla al mero azar.

En este caso hemos planteado un contraste bilateral: cuando no hacemos ninguna presunción del sentido de la diferencia y consideramos inicialmente igual de probable las diferencias en un sentido o como en el otro.

También se puede expresar el resultado calculando el intervalo de confianza del 95 % para la diferencia de medias entre los grupos, que en este ejemplo resulta ser  $-24.02$  a  $0.42$ .

Pero vamos a ver otras posibles formas de contrastar este tipo de hipótesis. Podemos mezclar los 20 valores; supongamos que los escribimos en 20 papeletas y los introducimos en una urna, y extraemos al azar 10 papeletas por un lado y otras 10 por otro, calculamos la diferencia de medias entre estas dos muestras y anotamos si ésta nueva diferencia es igual o mayor que la obtenida originalmente o si, por el contrario, es menor. Volvemos a introducir las papeletas en la urna y repetimos el proceso un gran número de veces. Calculamos, repitiendo ese procedimiento, el porcentaje de las veces que encontramos una diferencia igual o más extrema que la obtenida realmente en nuestro experimento. Si ese número es alto concluimos que la etiqueta grupo A o B verdaderamente puede deberse al azar y que por lo tanto no hay diferencias estadísticamente significativas entre los grupos, mientras que si ese número de veces es pequeño, nuestro resultado es infrecuente, y por lo tanto preferiremos rechazar la hipótesis de igualdad, y pensar que la presencia de diferencias es más plausible. Esta metodología se conoce como **pruebas de aleatorización**.

Mediante un programa de ordenador se puede simular ese proceso de colocar los 20 valores en una urna y extraer papeletas al azar.

Pero además, en este ejemplo concreto es posible considerar todos los conjuntos diferentes de 2 muestras de 10 elementos, corresponde a  $C_{10}^{20}$  combinaciones de 20 elementos tomados de 10 en 10, que constituyen 184756 posibilidades, y podemos calcular el número de entre ellas en las que la diferencia es igual o mayor que la observada. Obviamente, tanto en el caso anterior en el que constituíamos las muestras de forma aleatoria, como en el caso de que enumeremos todas las posibilidades, lo haremos utilizando el ordenador. Pero incluso con el ordenador si las muestras son grandes, el tiempo de cálculo para enumerar todas las posibilidades puede ser enorme y habrá que acudir a la primera situación en la que generamos un gran número, aunque limitado, de muestras aleatorias.

Utilizando el software que se referencia en uno los [enlaces](#), generando 10000 muestras aleatorias, obtenemos para este ejemplo que en el 5.5 % de esas muestras se encuentra una diferencia igual o mayor que la realmente observada, lo que corresponde a un valor muy parecido al obtenido mediante el contraste paramétrico (5.8 %).

También es posible con esta filosofía calcular un **intervalo de confianza** para la diferencia de medias, obteniendo a partir de esas 10000 muestras aleatorias el siguiente intervalo ( $-23.81$  :  $0.33$ )

Hago notar que ahora hemos utilizado una técnica estadística para comparar las dos muestras que no hace ninguna suposición en cuanto a la distribución de probabilidad de la que proceden los datos.

---

Otro de los conceptos centrales de la estadística es el de **estimación**: cómo obtener conclusiones generales de un colectivo a partir de una muestra y por lo tanto disponiendo sólo de parte de la información.

Una vez definido el parámetro que deseamos conocer, ya sea éste la media, la mediana u otro, se trata de calcularlo a partir de los datos observados en nuestra muestra y de cuantificar de alguna manera cómo de precisa es esa estimación; o dicho de otra forma, determinar hasta qué punto podemos garantizar que lo que

nosotros hemos observado es extrapolable al colectivo del que se extrajo esa muestra.

La forma más conocida y habitual de expresar la precisión de un parámetro consiste en indicar un **intervalo de confianza** para un determinado nivel de probabilidad. Así el intervalo de confianza del 95% son aquellos dos valores, uno inferior y otro superior, entre los que cabe esperar que se encontrará el parámetro calculado en 95 de cada 100 veces que se repita el experimento, con una muestra del mismo tamaño extraída de forma aleatoria de la misma población. En el cálculo de ese intervalo de confianza interviene el tamaño de muestra, la dispersión de los datos en la población (si ésta es pequeña la precisión será mayor). Por otro lado hay que tener en cuenta la distribución de probabilidad del parámetro, utilizándose habitualmente, aunque no siempre es adecuado, la distribución normal.

Así en el caso de que se estime una media en una muestra de tamaño  $n$ , el procedimiento para el cálculo del **intervalo de confianza basado en la  $t$  de Student**, consiste en estimar la desviación típica de los datos  $s$  y calcular el error estándar de la media

$$\frac{s}{\sqrt{n}}$$

siendo entonces el intervalo de confianza para la media

$$m \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Para la estimación de un parámetro, es posible emplear una metodología similar a la explicada anteriormente para el contraste de hipótesis, mediante una prueba aleatorizada, consistente en calcular las diferencias de cada valor con respecto de la media y generar muestras de esas diferencias, en las que a cada valor se le da signo positivo o negativo de forma aleatoria, y a partir de la distribución obtenida se calcula un intervalo de confianza, los valores que dejan el 2.5 % a cada lado de esa distribución.

En esta línea basada en considerar que, en ausencia de cualquier otra información respecto a la población que no sea la contenida en la propia muestra, la distribución de los valores encontrados en una muestra aleatoria constituyen la mejor orientación en cuanto a la distribución de esa población, encontramos otros métodos de conseguir estimaciones, de los que vamos a comentar los que se conocen en inglés como "bootstrap" y "jackknife".

El método **bootstrap** aunque ya usado anteriormente, fue descrito de forma sistemática por Efron en 1979. El nombre alude al cordón de los zapatos, recordando la imagen de alguien intentando salir del barro tirando del cordón de sus propios zapatos, y consiste, si tenemos una muestra de tamaño  $N$ , en generar un gran número de muestras de tamaño  $N$  efectuando un muestreo con reemplazamiento de esos valores. Es como si metiésemos los valores en una urna, extraemos una papeleta, anotamos el resultado, y volvemos a colocarlo en la urna, y así hasta obtener  $N$  valores. En esa muestra calculamos el valor del parámetro que estamos estimando. Y así repetimos el proceso un gran número  $B$  de veces (por ejemplo 10000 o más), con lo que obtenemos una distribución de valores para el parámetro en la que podemos calcular su dispersión (análogo del error estándar) y determinar unos límites de confianza utilizando esa distribución.

El método **jackknife** fue propuesto por Tukey en 1958 y en este caso el nombre alude a esas navajas multiuso que son a la vez sacacorchos, abrelatas, destornillador..., buenas para todo si no se tiene algo mejor a mano –siempre es mejor abrir una lata con ellas que con una piedra, pero donde esté un abrelatas del

"explorador"...-

Calculamos ahora el parámetro de interés  $\vartheta$  en la muestra total de tamaño  $N$ ; en este caso las diferentes muestras se van construyendo eliminando cada vez una de las observaciones  $X_i$  y en ellas se calcula de nuevo el valor del parámetro de interés  $\vartheta_i$ , repitiendo el proceso para  $i=1$  hasta  $N$ . Se obtiene lo que se denomina pseudovalores:

$$J_i = N \cdot \vartheta - (N - 1) \cdot \vartheta_i$$

Se llama estimador jackknife a la media de esos pseudovalores, y a partir de la distribución obtenida se calcula también su varianza y un intervalo de confianza para el mismo.

Al no utilizar más que los valores observados en la muestra, podemos denominar a estos métodos como autosuficientes.

La idea que subyace en alguno de los métodos autosuficientes descritos es la de generar a partir de los valores observados muestras aleatorias de acuerdo a un modelo específico que se está contrastando. Vamos a comentar esto un poco más.

Supongamos que no disponemos de la fórmula para calcular intervalos de confianza para la media  $m$ , calculada en una muestra de tamaño  $N$  extraída de forma aleatoria de una población cuyos valores se distribuyen según una función de probabilidad normal o gaussiana de media desconocida, que es la que pretendemos estimar. Estamos suponiendo por tanto un modelo de distribución normal con media  $M$  desconocida, y si pudiéramos generar mediante un programa, de acuerdo a ese modelo, un gran número de muestras aleatorias de tamaño  $N$ , a partir de la distribución de esas muestras podríamos calcular los percentiles del 2.5 %, a un lado y a otro, para determinar así un intervalo de confianza el 95 %. Si lo hacemos comprobaremos que los valores obtenidos concuerdan con los que se calculan utilizando la fórmula teórica basada en las propiedades de la distribución normal.

La pregunta es ¿cuál es entonces la utilidad de estos métodos, si disponemos de una fórmula teórica?. La respuesta: que esa fórmula teórica la tenemos para unos casos muy concretos, en nuestro ejemplo la media de datos extraídos de una población normal, mientras que el método de simulación nos permite obtener estimaciones de forma más general, para cualquier parámetro y cualquier modelo que estemos planteando, siempre que a partir de él podamos generar valores de forma aleatoria.

Esta metodología de simulación, generando a partir de un modelo valores aleatorios, se conoce como **método de Monte Carlo** (por los casinos y las ruletas de ese bonito lugar de la costa mediterránea), y las pruebas aleatorizadas y las estimaciones "bootstrap" se pueden considerar casos particulares de esta metodología.

Cuando se dispone de una respuesta teórica a un problema no tiene sentido utilizar la simulación de Monte Carlo, pero en ocasiones la solución analítica de algunos problemas es muy compleja y el único camino posible es la utilización de técnicas de simulación. Esto es algo que se viene empleando desde hace mucho y con éxito en problemas de ingeniería y también en estudios de biología, ciencias naturales, economía, y bastante menos en medicina.

Debido a que se trata de métodos que sólo son posibles gracias a la moderna potencia de cálculo de los ordenadores, ya que es preciso generar un elevado número de muestras aleatorias, también se conocen como **métodos de uso intensivo del ordenador** (computer intensive methods).

---

Por curiosidad he efectuado una serie de búsquedas en Medline (29-9-2002) obteniendo los siguientes resultados

Búsqueda	Nº de referencias
"monte carlo"	7432
"monte carlo" AND hypertension	29
bootstrap	1204
bootstrap AND hypertension	10
jackknife	10
jackknife AND hypertension	0
resampling	333
resampling AND hypertension	1
"computer intensive method"	4

También por curiosidad he efectuado una búsqueda simple en [Lancet](#) del término "monte carlo", encontrando 18 artículos, y bootstrap 29 artículos.

En [BMJ](#) buscando "[monte carlo](#)" aparecen 33 artículos, y buscando [bootstrap](#) aparecen 34.

Aunque son en general cifras pequeñas vemos, una vez más, que hay gente y artículos para todo...

## Enlaces interesantes

- [Download Trial version of RT 2.1](#)  
A DOS program for randomization methods. Written by Dr. Bryan F. Manly to accompany his book, "Randomization and Monte Carlo Methods in Biology" published by Chapman and Hall.
- [Resampling: The New Statistics](#)  
Julian L. Simon  
Second Edition published October 1997
- [The Philosophy and Practice of Resampling Statistics](#)  
Julian L. Simon
- [Introduction to Monte Carlo Methods](#)  
Computational Science Education Project
- [Stanford Statistics department course on the bootstrap](#)  
Susan Holmes
- [Resampling Methods](#)  
Phillip Good
- [Resampling Statistics: Randomization and the Bootstrap](#)  
David C. Howell University of Vermont
-

## [NonParametric STATistical Testing](#)

NPSTAT and NPFACT are a subset of programs developed to carry out computer intensive analysis (such as randomization tests but not bootstrapping). They provide nonparametric methods of null hypothesis testing.

## [Resampling methods](#)

Chong-ho Yu



[Indice de artículos](#)

[Principio de la página](#) ▲