Imputación múltiple
Un novedoso método para el tratamiento de valores faltantes

Adriana Pérez, Martín Alonso Rondón · Bogotá

Clínicos, investigadores y en especial epidemiólogos, frecuentemente se enfrentan al problema de tener datos faltantes en las bases de datos de sus estudios. Muchas técnicas estadísticas existen para dar solución a este problema, como por ejemplo el análisis de casos completos y la imputación de datos. Sin embargo, estas técnicas no siempre se implementan en los análisis principalmente por su desconocimiento o por la ausencia de paquetes computacionales con este tipo de aplicaciones. El presente trabajo desarrolla una descripción de tres de las técnicas de manejo de valores faltantes más utilizadas. El objetivo principal es mostrar las ventajas de una técnica novedosa para el tratamiento de los datos faltantes conocida como imputación múltiple. Con el fin de ilustrar al lector, al final se presenta su aplicabilidad mediante el desarrollo de un ejemplo práctico. *(Acta Med Colomb* 2002; **27:** 204-208).

Palabras clave: *valores faltantes, imputación múltiple, análisis de datos.*

Introducción

Se considera un valor faltante aquel valor de una variable de interés para el cual no se tiene información pero que debió ser obtenido por los investigadores. Clínicos, investigadores y en especial epidemiólogos, muchas veces se encuentran con el problema de tener datos faltantes en las bases de datos de sus estudios o investigaciones y no cuentan con los conocimientos o las herramientas para manejar esta ausencia de información.

La literatura médica tampoco ha recomendado ni ha manejado el tema de estimaciones inadecuadas o sesgos que se pueden tener al informar los resultados con valores faltantes. En otros casos, los autores no mencionan si realizaron algún tipo de tratamiento para el manejo de estos datos faltantes. No obstante, al registrar los resultados se observa la ausencia de consistencia en los datos y esto se debe al inadecuado manejo tales valores.

El presente trabajo tiene como objetivo mostrar a nuestros colegas médicos una técnica novedosa para el tratamiento de los datos faltantes en las bases de datos clínicas. Desde el punto de vista estadístico, las técnicas para el manejo de los datos faltantes es conocido y muchas herramientas se han desarrollado en este aspecto, sin embargo, éstas no se han implementado en el análisis de las bases de datos clínicas, muchas veces por desconocimiento de los errores obtenidos en los valores estimados al manejar la información, aunque consideramos que uno de los principales inconvenientes era la ausencia de herramientas computacionales para poder realizar este tipo de análisis.

Análisis de datos disponibles

Dentro de los métodos más comunes para manejar datos faltantes, está el análisis de datos o casos disponibles. En éste, se utiliza la información disponible u observada para realizar las estimaciones, así, a medida que se tiene un mayor número de valores faltantes, hay menor información en comparación con aquellas variables donde la información es completa en sus individuos. Como se observa en la Figura 1, se tienen dos variables de interés X_1 y X_2; ambas presentan valores faltantes en diferentes individuos. Se desea calcular el promedio de ambas variables así como la correlación entre ellas. El análisis de casos disponibles radica en que para el promedio de la variable X,, se utilizan los individuos disponibles para esa variable. Para calcular el promedio de la variable X_2 se utilizan los individuos con información en esa segunda variable. En forma similar, cuando se calcula el coeficiente de correlación entre éstas se hace uso sólo de la información disponible en ambas variables simultáneamente.

Como es de esperarse en estas circunstancias los datos obtenidos de este análisis muy posiblemente no represen-

Este trabajo es un resultado indirecto del proyecto de investigación "Comparación de técnicas de estimación de valores faltantes para cuantificar la severidad fisiológica en pacientes admitidos a cuidado intensivo en Colombia" financiado por Colciencias contrato 1203-04-954-98 y con el apoyo de la Pontificia Universidad Javeriana y la red internacional de epidemiología clínica (INCLEN).

Dra. Adriana Pérez Medina: Unidad de Epidemiología Clínica y Bioestadística, Facultad de Medicina, Pontificia Universidad Javeriana, Bogotá, D.C. y The University of Texas Health Science Center at Houston. School of Public Health at Brownsville; Dr. Martín Alonso Rondón Sepúlveda: Unidad de Epidemiología Clínica y Bioestadística, Facultad de Medicinal Pontificia Universidad Javeriana, Bogotá, D.C.

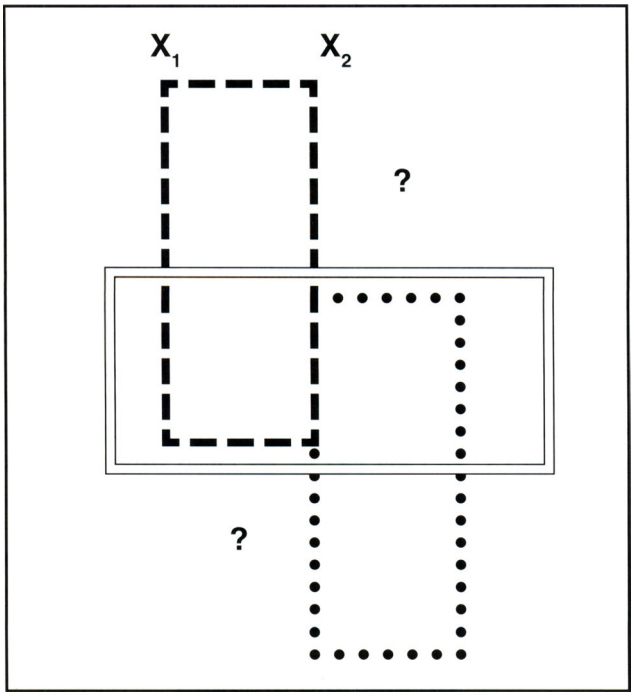

Figura 1. *Uso de la información en el análisis de casos disponibles.*
Identifica la información utilizada para calcular el promedio de X_1,
Identifica la información utilizada para calcular el promedio de X_2,
Identifica la información utilizada para calcular el coeficiente de correlación de X_1 con X_2.

tan la población de estudio ya que no se analizan los mismos individuos o la información consistentemente. Esta técnica es tal vez la más usada por todos los investigadores porque los programas estadísticos la tienen implementada por omisión a cualquier otra.

Imputación simple

Existen circunstancias donde los investigadores no se pueden dar el lujo de "perder individuos" o de reducir su tamaño de muestra. Un ejemplo claro es cuando se desea evaluar una escala de medición y se requieran los valores de todas las variables. En estos casos la técnica de manejo de datos faltantes más frecuentemente usada es el método de imputación de la media o de la moda de la variable a los datos faltantes. Al "imputar" o asignar a cada individuo, con información faltante, el promedio de la variable del resto de los datos observados, no se distorsiona la variabilidad de los datos; pero las asociaciones se distorsionan porque no existe variabilidad residual dada por el proceso de imputación. Entonces, el número de individuos después de la imputación se aumenta en forma ficticia al asignarle el valor del promedio. Esta técnica de manejo de los datos faltantes se conoce bajo el nombre de imputación simple.

Ambas técnicas, el análisis de casos disponibles y la técnica de imputación simple, han sido demostradas como herramientas poco apropiadas para el manejo de los datos faltantes (1-5).

Imputación múltiple

En los últimos diez años, se promueve en la literatura estadística el uso de un método inédito para el manejo de los datos faltantes llamado el método de imputación múltiple. Aunque este método había demostrado tener excelentes propiedades para su uso y utilización en el análisis de datos clínicos (3, 5) sólo hasta en los tres últimos años se tuvo acceso al software que permite utilizarla en forma gratuita (6) o a través de una subrutina dentro del programa S-plus© (7).

A diferencia de las otras técnicas de imputación, el interés principal de ésta no es estimar los valores puntuales de la variable como tal, sino proveer un método que permita encontrar un mejor estimador cercano al verdadero parámetro de interés (por ejemplo el promedio aritmético), con la información recolectada en el estudio (observada y faltante).

Esta técnica de imputación requiere varios supuestos para su correcto uso como son: el mecanismo de pérdida de los valores faltantes tiene que ser en forma aleatoria (FFA); se establece un modelo de imputación de los datos faltantes relacionado con la forma del análisis de los mismos y se tiene un supuesto sobre la distribución *a priori* de los parámetros de interés.

El proceso de generación de los valores imputados debe realizarse más de una vez, es decir, se tendrán al menos dos bases de datos sin información faltante. Posteriormente, con estas bases completas, se debe hacer un proceso de combinación de los resultados, para obtener el análisis sin sesgo por la realización del proceso de imputación (3, 5).

Mecanismo de pérdida de los valores faltantes

Existen tres mecanismos (1,3) de pérdida de los valores faltantes y serán explicados con un ejemplo de un paciente en un estudio. Después de realizar una medición inicial del paciente, se le pide que regrese a una visita de seguimiento. El paciente no regresa a la visita de seguimiento y por tanto la información requerida es faltante.

Esta información es faltante de forma completamente aleatoria (FFCA) si la razón por la cual el paciente faltó, no depende ni de la información del paciente en la visita inicial ni de la información que es faltante. En el ejemplo puede ser porque el paciente cambió de ciudad de residencia y no puede continuar en el estudio, o puede ser porque es el día del cumpleaños del paciente y simplemente no fue a la visita.

De otra parte, esta información es faltante de forma aleatoria (FFA), si la razón por la cual el paciente faltó esta relacionada con la información suministrada en la visita inicial, pero no depende de la información que es faltante. En el ejemplo puede ser que el paciente deje el estudio bajo el consejo del médico, porque la información suministrada en la visita inicial es por ejemplo inadecuada.

Finalmente, esta información puede ser faltante de forma informativa (FFI) o no aleatoria, si la razón por la cual

el paciente faltó está relacionada en la información suministrada en la visita inicial y también está relacionada con la información que es faltante. Un ejemplo de esta circunstancia, puede ser cuando el paciente falta a su visita de seguimiento porque no se siente bien.

Modelo de imputación de los datos faltantes

El modelo de imputación de los datos faltantes hace referencia a la distribución estadística de todas las variables a analizarse (3, 5). Se debe identificar cómo son todas las variables de interés, las cuales pueden ser todas perfectamente continuas (normal) o discretas (binomial, multinormal, etc.) o mixtas (modelo mixto).

Distribución *a priori* de los parámetros de interés

Usualmente, la distribución y los valores iniciales de los parámetros de interés se obtienen de estudios previos en el tema (5).

Número de veces que se debe imputar los valores faltantes

Rubin (3) demuestra matemáticamente que se debe imputar más de dos veces una base de datos. Recomienda cinco veces como un número adecuado para obtener estimaciones insesgadas de la base de datos con valores faltantes. El autor (3) también demuestra cómo la ganancia al realizar más de cinco veces el proceso no es grande. La Figura 2 presenta el caso de cinco bases de datos obtenidas después de realizar proceso de imputación múltiple.

Combinación de las bases de datos imputadas

Después de realizar la generación por imputación múltiple se generan varias bases de datos, por lo tanto es necesario combinar los resultados para el análisis de interés. La combinación puede ser realizada variable por variable o análisis por análisis. Por ejemplo, si se desea el promedio de una variable como el nivel de creatinina (NC) en la sangre y se realizaron cinco imputaciones múltiples, entonces, se calcula el promedio del NC en cada una de las cinco bases de datos y posteriormente se promedian los cinco promedios en uno solo de NC, éste es el NC que se informa.

En contraste, si se está interesado en obtener el coeficiente de correlación (CC) entre dos variables. Entonces se calcula el CC en cada una de las cinco bases de datos imputadas y luego se promedian los cinco CC calculados. Este CC promediado es el estimador de interés. Los lectores interesados en una explicación detallada del tema con un ejemplo real en pacientes de cuidado intensivo en Colombia pueden consultar Pérez et al (8).

Un ejemplo práctico

Los datos presentados en la Tabla 1 son ficticios y corresponden a la edad (X_1), el peso (X_2), y los niveles de colesterol plasmático total (Y) de una población de 25 pacientes quienes sufrían de hiperlipoproteinemia, antes de ser tratados con un medicamento (9).

Estamos interesados en estimar el promedio de la variable Y ante la presencia de valores faltantes. Para ello, se estableció como supuesto que todas las variables siguen

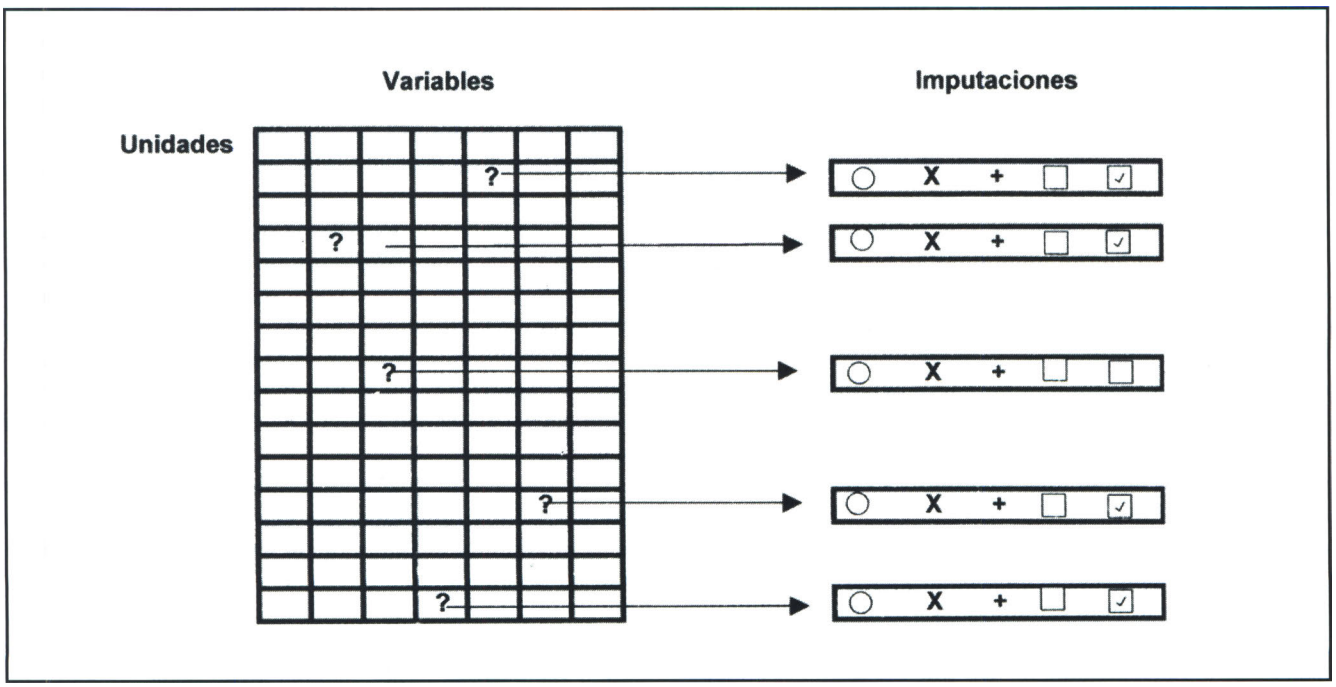

Figura 2. Generación de cinco imputaciones múltiples en una base de datos con valores faltantes, bajo el supuesto que los datos faltantes son en forma aleatoria, un modelo de imputación fijo y una distribución a priori de los parámetros de interés de las variables en esta base de datos.

una distribución normal. Para evaluar el comportamiento, se "crearon" valores a declararse faltantes en la variable Y. Se seleccionaron tres individuos aleatoriamente, como si fueran valores faltantes (Y_2). Igualmente, se seleccionaron seis individuos aleatoriamente como si fueran valores faltantes (Y_2). Estas asignaciones corresponden a 12% y 24% de valores faltantes respectivamente.

La Tabla 1 presenta las dos nuevas variables (Y_1,Y_2) con los valores retirados de los datos originales en los dos escenarios establecidos. Adicionalmente, se presenta el promedio y la desviación estándar (de) de cada una de las variables.

En la Tabla 1 se puede observar, como el promedio en Y cambia a medida que se incrementa el porcentaje de valores faltantes (en nuestro caso aumentó). El promedio real de 310.72 mg/dL (de=77.83 mg/dL) en los niveles de colesterol plasmático cambió a un promedio de 315.59 mg/dL (de=81.46 mg/dL) cuando se presenta 12% de valores faltantes en la variable, y cambió a un valor de 320.68 mg/dL (de=72.03 mg/dL) cuando se presenta 24% de valores faltantes en la variable.

Para tratar de resolver el problema de los valores faltantes, se generaron cinco imputaciones para cada uno de los escenarios antes planteados con ayuda del software NORM© (6). Se tienen como covariables la edad (X_1) y el peso (X_2). Se utilizó una información *a priori* no informativa como distribución de los parámetros. Los valores generados como imputados en cada uno de los escenarios, 12 % y 24% de valores faltantes, se presentan en la Tabla 2, sólo con el objetivo de que el lector pueda obtener los resultados presentados. Los valores en sí no son importantes, lo importante es la combinación de éstos con el fin de obtener el promedio deseado.

Finalmente, la Tabla 3 presenta los promedios y las desviaciones estándar calculados después de combinar los resultados obtenidos en las cinco imputaciones para cada uno de los escenarios. Es decir, se presentan los promedios de las bases de datos completas.

En la Tabla 3 se puede observar como la estimación del promedio de Y mejoró con respecto a los resultados de los datos disponibles mostrados en la Tabla 1. Especialmente, cuando se tiene 24% de valores faltantes, la estimación del promedio de Y con el proceso de imputación múltiple fue de 311.6 mg/dL (de=78.5 mg/dL), realmente muy parecido al verdadero valor de 310.72 mg/dL (de=77.83 mg/dL), pero clínicamente diferente con el valor obtenido de 320.68 mg/dL al usar los casos disponibles.

Tabla 1. *Valores observados de tres variables en estudio, con sus promedios y desviaciones estándar*

Paciente	X_1	X_2	Y	Y_1	Y_2
1	84	46	354	354	354
2	73	20	190	190	-
3	65	52	405	405	405
4	70	30	263	-	-
5	76	57	451	451	-
6	69	25	302	302	302
7	63	28	288	288	288
8	72	36	385	385	385
9	79	57	402	402	402
10	75	44	365	365	365
11	27	24	209	209	-
12	89	31	290	290	290
13	65	52	346	346	346
14	57	23	254	-	-
15	59	60	395	395	395
16	69	48	434	434	434
17	60	34	220	220	220
18	79	51	374	374	374
19	75	50	308	-	-
20	82	34	220	220	220
21	59	46	311	311	311
22	67	23	181	181	181
23	85	37	274	274	274
24	55	40	303	303	303
25	63	30	244	244	244
Promedio	68.68	39.12	310.72	315.59	320.68
Desviación estándar	12.73	12.25	77.83	81.46	72.03

X_1: edad del paciente en años
X_2: peso del paciente en kg
Y: nivel de colesterol total plasmático en mg/dL
Y_1: nivel de colesterol total plasmático en mg/dL con 12% de valores faltantes
Y_2: nivel de colesterol total plasmático en mg/dL con 24% de valores faltantes

Summary

Frequently, clinicians, researchers and especially epidemiologists confront the problem of missing data in their

Tabla 2. *Valores obtenidos después del proceso de imputación múltiple con cinco imputaciones para la variable Y en dos escenarios de porcentajes de valores faltantes.*

Paciente	12% de valores faltantes					24% de valores faltantes				
	Y_{11}	Y_{12}	Y_{13}	Y_{14}	Y_{15}	Y_{21}	Y_{22}	Y_{23}	Y_{24}	Y_{25}
2						301	271	151	178	243
4	306	244	229	294	260	302	252	294	233	379
5						444	411	360	327	424
11						175	215	274	189	167
14	162	120	204	226	199	248	194	145	259	245
19	307	279	407	363	320	283	476	409	325	314

Tabla 3. *Promedios y desviaciones estándar calculadas después de combinar los resultados obtenidos en las cinco imputaciones, para los dos escenarios.*

Imputación	12% de valores faltantes		24% de valores faltantes	
	Promedios	**Varianzas**	**Promedios**	**Varianzas**
1	308.7	6747.0	313.8	5670.4
2	303.4	7515.9	316.5	6651.3
3	311.3	6966.8	309.0	6745.6
4	313.0	6247.1	304.2	5659.0
5	308.9	6455.4	314.6	5915.3
Promedio	309.1	6786.5	311.6	6128.3
Desviación estándar	82.5	-	78.5	-

study data set. This kind of problem can be solved by many statistical techniques. For example, the complete case analysis or the multiple imputation technique. However, these techniques are not common to be implemented mainly because the users do not know how to use them or there was not statistical software available with these tools. This article describes three of the most frequently used imputation techniques. The main aim is to show the advantages and disadvantages of a new statistical technique known under the name of multiple imputation to handle missing data. A practical example is presented showing the applicability of this technique.

Key words: *missing data, multiple imputation, data analysis.*

Referencias

1. **Little RJA, Rubin DB.** Statistical Analysis with Missing Data. New York. John Wiley & Sons, 1987.
2. **Särndal CE, Swensson B, Wretman J.** Model Assisted Survey Sampling. New York. Springer Verlag, 1992.
3. **Rubin DB.** Multiple Imputation for Nonresponse in Surveys. New York, John Wiley & Sons, 1987.
4. **Dempster AP**, **Laird NM**, **Rubin DB.** Maximum Likelihood Estimation from Incomplete data via the EM Algoritm (with discussion). *Journal of the Royal Statistical Society, Series B* 1977; **39**:1-38.
5. **Schafer JL.** Analysis of Incomplete Multivariate Data. New York. Chapman Hall, 1997.
6. **Schäfer JL.** RM Version 2.02 for Windows 95/98/NT. Multiple Imputation of Multivariate Continuous Data Under a Normal Model. Department of Statistics, The Pennsylvania State University,http://www.stat.psu.edu/~jls/. 1999.
7. **Schimer J, Schafer LJ, Hesterberg T, Fraley C, Clarkson DB.** Analyzing data with missing values in S-Plus. Insightful Corporation. Seattle.
8. **Pérez A, Dennis RJ, Gil JFA, Rondón MA.** Informe Técnico del Proyecto: Comparación de técnicas de estimación de valores faltantes para cuantificar la severidad fisiológica en pacientes admitidos a cuidado intensivo en Colombia. Presentado a Colcienicas. Junio 2000.
9. **Kleinbaum D, Kupper L, Muller K, Nizam A.** Applied Regression Analysis and Multivariable Methods. Tercera edición. Duxbury Press. Pacific Grove, 1998.