Gac Sanit. 2017;31(4):342–345

Nota metodológica

Imputación de valores ausentes en salud pública: conceptos generales y aplicación en variables dicotómicas

Gilma Hernández[a,b], David Moriña[c,d] y Albert Navarro[d,*]

[a] Instituto de Investigaciones Médicas, Universidad de Antioquia, Medellín, Colombia
[b] Programa de Doctorado en Metodología de la Investigación Biomédica y Salud Pública, Departament de Pediatria, d'Obstetricia i Ginecologia i de Medicina Preventiva, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès, Barcelona), España
[c] Unitat d'Infeccions i Càncer (UNIC), Programa d'Investigació en Epidemiologia del Càncer (PREC), Institut Català d'Oncologia (ICO)-IDIBELL, L'Hospitalet de Llobregat (Barcelona), España
[d] GRAAL-Unitat de Bioestadística, Facultat de Medicina, Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès, Barcelona), España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 23 de noviembre de 2016
Aceptado el 9 de enero de 2017
On-line el 15 de marzo de 2017

Palabras clave:
Valores ausentes
Imputación
Salud pública
Epidemiología

RESUMEN

Que haya valores ausentes en variables registradas en encuestas de salud es habitual, pero no lo es imputarlos posteriormente cuando se realiza el análisis. Trabajar con datos imputados puede tener ventajas en términos de precisión de los estimadores y de identificación sin sesgos de la asociación entre variables. Probablemente, el proceso de imputación sigue siendo desconocido para muchos profesionales no estadísticos, que le atribuyen una alta complejidad y quizás un objetivo que no es exactamente el que persigue. Para aclarar estas cuestiones, esta nota pretende ofrecer una visión amena, no exhaustiva, del proceso de imputación, que permita conocer sus bondades para el trabajo de un salubrista. Todo ello en el marco de variables dicotómicas, habituales en salud pública. Para ilustrar los conceptos se usa un ejemplo en el cual se trabaja con datos con valores ausentes, imputados de forma simple y múltiple.

Imputing missing data in public health: general concepts and application to dichotomous variables

Keywords:
Missing data
Imputation
Public health
Epidemiology

ABSTRACT

The presence of missing data in collected variables is common in health surveys, but the subsequent imputation thereof at the time of analysis is not. Working with imputed data may have certain benefits regarding the precision of the estimators and the unbiased identification of associations between variables. The imputation process is probably still little understood by many non-statisticians, who view this process as highly complex and with an uncertain goal. To clarify these questions, this note aims to provide a straightforward, non-exhaustive overview of the imputation process to enable public health researchers ascertain its strengths. All this in the context of dichotomous variables which are commonplace in public health. To illustrate these concepts, an example in which missing data is handled by means of simple and multiple imputation is introduced.

Introducción

Que haya valores ausentes es frecuente en salud pública. Ignorarlos conlleva la pérdida de potencia del estudio y la obtención de estimadores ineficientes y posiblemente sesgados. Los valores ausentes representan falta de información en el contenido de una o varias variables en un conjunto de datos, y pueden deberse a factores como la no respuesta en una encuesta, la falta de alguna medición, la pérdida en el proceso de recolección, etc. Algunos ejemplos en el ámbito de la salud pública son la imputación del instante de seroconversión al virus de la inmunodeficiencia humana[1] o el estado físico y mental en las personas mayores[2].

El abordaje más frecuente consiste en ignorar los valores ausentes y usar la variable sin mayor consideración. Al hacerlo conjuntamente con otra variable sin valores ausentes, el análisis tiene en cuenta solo aquellos casos completos (*listwise deletion* [LW]), descartando información disponible. Con esta estrategia, si el análisis es multivariado, incluso en situaciones en que el porcentaje de valores ausentes sea bajo en cada variable, puede suponer que el número de casos analizados sea sensiblemente inferior al tamaño muestral con el que se creía trabajar[3]. Ello implica estimaciones ineficientes y, a veces, sesgadas[4–7].

La alternativa consiste en imputar los valores ausentes, consiguiendo que no se descarten casos. Si bien se dispone de programas estándar, como SAS, R, Stata o SPSS, que cuentan con algoritmos de imputación, diríamos que su uso no es habitual.

Existe literatura sobre imputación en el ámbito de la salud, pero la mayoría se ocupa de la imputación de variables continuas[7,8] y

* Autor para correspondencia.
Correo electrónico: albert.navarro@uab.cat (A. Navarro).

http://dx.doi.org/10.1016/j.gaceta.2017.01.001

no dicotómicas, muy habituales en salud pública. El propósito de esta nota es ofrecer a profesionales no estadísticos una descripción general de la imputación de valores ausentes, enfatizando en variables de naturaleza dicotómica.

Mecanismos de pérdida

Existen tres mecanismos:

- *Missing Completely At Random (MCAR):* la probabilidad de observar un valor ausente en una variable no depende de las otras variables ni de ella misma. Los sujetos con y sin valores ausentes tienen las mismas características.
- *Missing At Random (MAR):* la probabilidad de observar un valor ausente depende de otras variables, no de los valores de la propia variable.
- *Missing Not At Random (MNAR):* la probabilidad de observar un valor ausente depende de los valores de la propia variable, una vez controladas el resto de las variables. En esta situación no pueden imputarse los valores ausentes.

Es importante identificar el patrón en que aparecen los datos ausentes, ya que esto puede determinar la viabilidad de imputar y, en caso afirmativo, el método más eficiente[3,5,7].

Imputación simple

Consiste en asignar un valor al valor ausente, que posteriormente es analizado exactamente igual que los realmente observados. Para variables dicotómicas existen varios métodos: entre otros, generar una nueva categoría que agrupe los valores ausentes; asignar el valor del vecino más cercano; o el método Hot-Deck, que consiste en extraer al azar, del grupo de sujetos con las mismas características que el que presenta el valor ausente, uno de los valores observados (donador). El lector interesado puede profundizar en imputación simple consultando varios trabajos[4,5].

Imputación múltiple

Su objetivo primario es mantener la variabilidad de la población preservando las relaciones entre variables. Tiene tres fases (fig. 1):

1. *Imputation step:* se crean *m >1* conjuntos de datos completos donde en cada uno se mantienen fijos los valores observados ($x1_i$), imputando los valores ausentes $x1_{i\_imp\_k}$. El valor imputado para una misma observación en cada conjunto no tiene por qué ser el mismo, lo cual incorpora variabilidad a estos valores (de los cuales nunca conoceremos el valor real). La obtención de valores plausibles se consigue mediante un modelo de imputación, que debería contener las variables que se analizarán posteriormente, incluida la respuesta, más aquellas que ayuden a explicar los valores ausentes.
2. *Completed-data analysis step:* cada conjunto de datos es analizado individualmente mediante procedimientos estándar, obteniendo estimadores particulares en cada conjunto $\left(\hat{\beta}_{X1\_k}\right)$ y $\left(\hat{\sigma}_{X1\_k}\right)$. Los estimadores diferirán en cada conjunto a causa de la variación introducida en la imputación de los valores ausentes.
3. *Pooling step:* combinando las estimaciones de los diversos conjuntos de datos mediante reglas simples[6] se obtienen los estimadores definitivos ($\hat{\beta}_{X1\_imp}$), así como los errores ($\hat{\sigma}_{X1\_imp}$) que incorporan la incertidumbre de los valores ausentes.

Para profundizar en la imputación múltiple pueden consultarse Rubin[6] y Van der Palm et al.[2].

Ejemplo

Tenemos una población con tres variables dicotómicas: la dependiente, Y∼Bin(N, π=0,207); la variable con valores ausentes, X_1∼Bin(N, π=0,399); y una sin valores ausentes, X_2∼Bin(N,

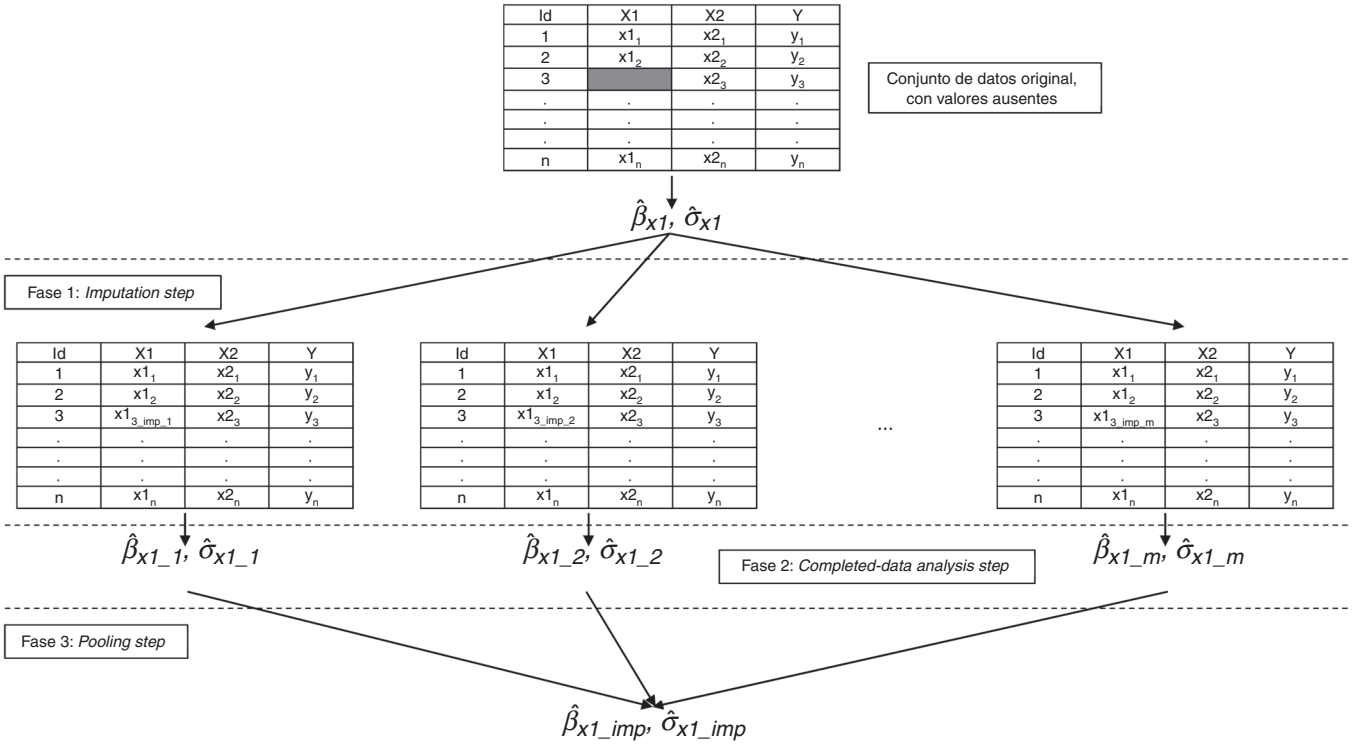

Figura 1. Esquema del proceso de imputación múltiple para una variable X1, con dos covariables sin valores ausentes (X2 e Y).

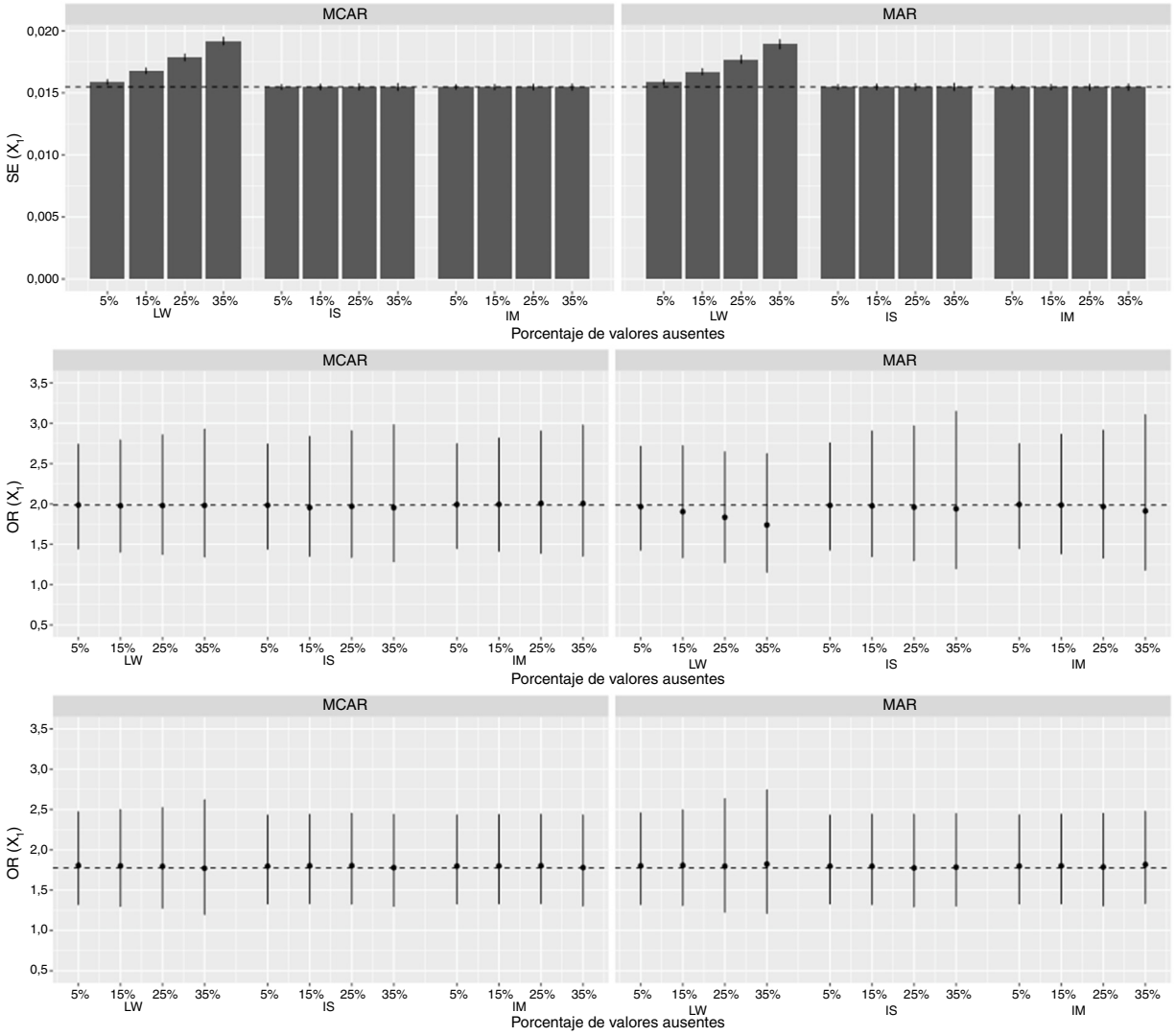

Figura 2. Resultados de las simulaciones: error estándar de X_1 (SE(X_1)), OR(X_1) y OR(X_2). La línea discontinua indica el valor poblacional.

π=0,442). Seleccionamos muestras de tamaño n = 1000 con diferentes porcentajes de pérdidas según MCAR y MAR (véase el Apéndice disponible *online* como Material suplementario). Se estiman los coeficientes de una regresión logística según LW, imputación simple (método Hot-Deck, librería R HotDeckImputation[9]) e imputación múltiple, mediante ecuaciones encadenadas[2,10] (librería R mice[10]). Se comparan los resultados en términos de precisión para la estimación de X_1 y de la asociación entre X_1 y X_2 con Y.

En la figura 2 se presentan los resultados de las simulaciones. En términos de precisión de X_1 puede observarse que, con LW, a mayor porcentaje de pérdida, peor precisión, mientras que al trabajar de forma imputada esta se mantiene. En términos de asociación de X_1 con Y se observa que, cuando el patrón de pérdidas es MCAR, todos los métodos realizan estimaciones cercanas al valor real. Sin embargo, cuando el patrón es MAR, LW obtiene estimadores con mayor sesgo al aumentar el porcentaje de valores ausentes. La imputación simple y la imputación múltiple arrojan estimadores cercanos al valor real en todos los casos, ligeramente con menor variabilidad con la imputación múltiple.

Discusión y conclusiones

En nuestra opinión, hay tres razones fundamentales por las que el uso de la imputación múltiple sigue siendo poco frecuente: 1) porque se cree que su objetivo consiste simplemente en sustituir un valor ausente por uno imputado; 2) por la percepción de que es una técnica compleja; y 3) por la creencia de que ante la incertidumbre que provoca un valor ausente lo más prudente es dejarlo como tal. La primera es falsa; sobre la segunda, creemos que puede afirmarse que hay técnicas más complejas cuyo uso está generalizado; y para la última opinamos que, a menudo, imputar puede ser más prudente que no hacerlo (con la información disponible e imputando podemos lograr estimadores más eficientes y menos sesgados, si no insesgados).

Trabajar con LW aumenta la imprecisión, y si el mecanismo de pérdida es MAR, generará estimadores sesgados[5,7]. Hay que distinguir entre imputación simple e imputación múltiple: la primera solo sustituye el valor ausente por otro que es tratado exactamente igual que uno observado; la segunda consiste en un proceso más elaborado que permite capturar la incertidumbre de los valores ausentes. A diferencia de cuando se trabaja con una variable continua, donde la imputación simple suele subestimar el error[5–7], según nuestros resultados para variables dicotómicas parecería que las diferencias entre imputación simple e imputación múltiple no son tan sensibles, siempre que el mecanismo de imputación reproduzca el patrón de pérdida. Y es que la validez de los resultados depende de que, en el caso de la imputación múltiple, el modelo de imputación se realice adecuadamente[3].

Nótese que la magnitud y la dirección del sesgo no siempre coincidirán con lo mostrado en nuestro ejemplo; dependerá de la relación entre las variables estudiadas. Siguiendo a Sterne et al.,[3] en la actualidad los procedimientos de imputación son ampliamente accesibles, por lo que no existe excusa para que los análisis potencialmente engañosos e ineficientes basados en LW sean considerados adecuados sin mayor atención.

Editora responsable del artículo

María Victoria Zunzunegui.

Contribuciones de autoría

Todas las personas firmantes contribuyeron a la concepción y el diseño del trabajo, el diseño de las simulaciones, el análisis y la interpretación de los datos, la escritura del documento y su revisión crítica con contribuciones intelectuales importantes, y aprobaron la versión final para su publicación.

Financiación

Si bien este trabajo no ha tenido financiación directa, el segundo autor ha sido parcialmente apoyado por becas del Instituto de Salud Carlos III (Gobierno de España), cofinanciado por fondos FEDER (Fondos para el Desarrollo Regional Europeo) - Una forma de hacer Europa (referencias: RD12/0036/0056, PI11/02090) y por la Agència de Gestió d'Ajuts Universitaris i de Recerca (2014SGR 756) y RecerCaixa 2015 (MD088652).

Conflicto de intereses

Ninguno.

Agradecimientos

Queremos agradecer a la Dra. Valeria Stuardo MA la lectura crítica y los posteriores comentarios a una de las versiones de este manuscrito.

Anexo. Material adicional

Se puede consultar material adicional a este artículo en su versión electrónica disponible en doi:10.1016/j.gaceta.2017.01.001

Bibliografía

1. Pérez-Hoyos S, Ferreros I, del Amo J, et al. Imputación del instante de seroconversión al VIH en cohortes de hemofílicos. Gac Sanit. 2003;17:474–82.
2. Van der Palm DW, van der Ark LA, Vermunt JK. A comparison of incomplete-data methods for categorical data. Stat Methods Med Res. 2016;25:754–74.
3. Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
4. Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley; 2002.
5. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002;7:147–77.
6. Rubin DB. Multiple imputation for nonresponse in surveys. New York: Wiley-Interscience; 2004.
7. Donders ART, van der Heijden GJMG, Stijnen T, et al. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087–91.
8. Cañizares M, Barroso I, Alfonso K. Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. Gac Sanit. 2004;18:58–63.
9. Joenssen DW. HotDeckImputation. Hot Deck Imputation Methods for Missing Data. 2015.
10. Van Buuren S, Groothuis-Oudshoorn K. MICE. Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45:1–67.