Badler, Clara Elisabeth
Alsina, Sara María*
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística - Facultad de Ciencias Económicas y Estadística - Universidad Nacional de Rosario

METODOLOGÍA PARA INFORMACIÓN FALTANTE: ¿ELEGIMOS BIEN?

1. INTRODUCCIÓN

En las últimas décadas es continuo el surgimiento de numerosas propuestas metodológicas para tratamiento de la información faltante. Sin embargo, al momento de su utilización, la elección no siempre es adecuada, pudiendo afectar el resultado de los análisis. Factores como la facilidad de aplicación de las técnicas, su disponibilidad en los softwares estadísticos de mayor utilización y/o la falta de difusión metodológica, pueden ser los causantes.

Resulta entonces de interés la revisión de conceptos teóricos asociados a las técnicas y la incorporación de resultados obtenidos, para su consideración ante la elección de la metodología a ser utilizada.

A una relectura de la metodología tendiente a recalcar los supuestos sobre los que fue formulada, se incorporan las evaluaciones y comentarios de importantes metodólogos que, a partir de distintos ámbitos, han ido sumando experiencias y conclusiones sobre aspectos de la metodología, en un tiempo transcurrido desde el surgimiento de cada una de las propuestas originales.

Distintos pueden ser los destinatarios, tanto analistas que deban introducirse en la temática como aquellos que ya la han incorporado, ya que los conceptos vertidos en este trabajo pueden ser motivadores para la profundización de los aspectos planteados o para acceder a la bibliografía citada, como punto de partida para la construcción o como aporte la reconstrucción de un marco teórico, o proporcionar argumentos iniciales para la elección adecuada de un tratamiento.

2. ¿DÓNDE ESTÁ LA FALTA DE INFORMACIÓN?

En forma general, Meng (2000) se refiere a la información faltante como a los datos que habíamos planeado observar y algún proceso nos lo impide.

La información faltante ha estado siempre presente en los diversos contextos de análisis y la estadística la ha conceptualizado de diversas maneras. Se presenta en los datos provenientes de relevamientos, experimentos o registros, correspondientes a un momento o a través del tiempo y surge por diferentes razones: en la mayoría de las encuestas algunos individuos no proveen información por rechazos, en estudios de medidas repetidas las faltas pueden deberse a abandono de los declarantes que no llegan al final del estudio.

de Leeuw et al. (2003) proveen una clasificación de la información faltante según las causas que la originarían:

---

* Docente-investigador e investigador del Consejo de Investigaciones de la Universidad Nacional de Rosario (CIUNR).

-información que no fue provista por el respondente para ciertas preguntas o ítems.
-información que se perdió; frecuentemente en la carga de datos.
-información provista pero no utilizable (sin sentido, fuera de rango, no codificable, no confiable, cuestionable, etc.), denominada información confusa (Badler et al., 1999).

La fuerza que ha ido adquiriendo este campo metodológico incluso hace que se consideren a variables con determinadas características como faltantes (por ejemplo variables latentes), para ser analizadas en el contexto de la información faltante (Little y Rubin, 2002; Schafer y Graham, 2006).

3.¿CUÁL ES EL PROBLEMA DE LOS DATOS FALTANTES?

Los datos faltantes crean problemas en la investigación científica pues la mayoría de los procedimientos de análisis no han sido diseñados para incorporarlos.

"Las faltas son generalmente una molestia, no el principal objeto de investigación, pero su incorporación produce dificultades conceptuales y desafíos de cálculo. Al carecer de recursos o también de marco teórico, los investigadores, metodólogos y creadores de software recurren a la corrección de los datos para obtener una apariencia de completitud. Desafortunadamente, estas correcciones producen muchas veces más daño que beneficio" (Schafer y Graham, 2002).

Tres típicos problemas surgen con la información faltante: pérdida de eficiencia, complicaciones en el manejo y análisis de los datos, y sesgos debidos a diferencias entre los datos observados y no observados (Barnard y Meng, 1999). Los sesgos que causan los datos faltantes dependen de la forma en que éstos falten o sea del denominado mecanismo; es importante, para prevenir y para tratar, establecer las causas y determinantes de las faltas (de Leeuw et al., 2003). Los datos faltantes pueden sesgar las estimaciones de los párametros, aumentar las tasas de error tipo I y II y alterar los intervalos de confianza. Además, como una pérdida de datos va acompañada en general por pérdida de información, puede reducirse el poder estadístico (Collins, Schafer y Kam, 2001).

4. LOS DATOS Y LAS CARACTERÍSTICAS DE LAS PÉRDIDAS

Es fundamental la caracterización de las pérdidas en los conjuntos de datos a través de sus esquemas y mecanismos, para poder realizar la selección de los tratamientos de acuerdo a los supuestos sobre los que han sido formulados.

El hecho de que se "debería saber más sobre los datos faltantes para seleccionar el mejor método para tratar los datos faltantes" parecería contradictorio (de Leeuw et al., 2003). Ante la dificultad de acceder a este conocimiento, se logra un acercamiento mediante distintas estrategias.

4.1. ¿Cómo se ubican las pérdidas en la base de datos?

El esquema o patrón de pérdida indica qué valores en el conjunto de datos están observados y cuáles son faltantes. El mismo está definido por una matriz indicadora de pérdida (R) (individuos por variables).

Algunos métodos de tratamiento de la falta de información son aplicables a cualquier esquema de pérdida mientras otros requieren esquemas especiales, como el monótono (Little y Rubin, 1987) (cuando la matriz indicadora de pérdida puede ser ordenada de manera tal que cada variable es siempre observada en más individuos que la siguiente), que surge también frecuentemente en datos longitudinales sujetos a abandonos. Como en los datos reales el esquema monótono no es tan frecuente pero facilita la aplicación de muchas técni-

cas, es posible crearlo descartando una proporción pequeña de observaciones o aplicando algún método de tratamiento previo al definitivo (Horton y Lipsitz, 2001;Little y Hyonggin, 2003; SAS, 2005).

El examen del esquema puede proporcionar información de gran utilidad: la mayoría de las faltas corresponden a una variable (que si no es central en el análisis estadístico propuesto puede tal vez eliminarse), o a una combinación de variables (relación que puede ser por ejemplo incorporada en el modelo de imputación), pocas unidades acumulan muchas faltas, las faltas pueden integrar un sistema que se puede probar de incluir en el análisis estadístico y en los procesos de imputación (de Leeuw et al., 2003).

4.2. ¿Con qué se relacionan las pérdidas?

Little y Rubin (2002) afirman que el mecanismo de pérdida "se refiere a la relación entre la falta y los valores de las variables en la matriz de datos"; es visto por Meng (2000) como "el proceso no intencional que nos impide observar los datos que habíamos planeado".

"El mecanismo es un proceso de selección de datos, como un proceso de muestreo, típicamente no controlado o desconocido por el recolector de datos. Así como la teoría general del proceso de muestreo y el impacto del sesgo de selección fueron formalizados hace largo tiempo, la fundamentación teórica del mecanismo de pérdida fue formalmente desarrollado por Rubin en el año 1976" (Meng, 2000).

En los procedimientos para datos faltantes la pérdida se considera como un fenómeno probabilístico. Se trata R como un conjunto de variables aleatorias con una distribución de probabilidad conjunta denominada distribución de pérdida, que podrá o no ser especificada. "En la literatura estadística la distribución de R algunas veces es denominada mecanismo de respuesta o mecanismo de pérdida, lo que puede producir confusión dado que mecanismo sugiere un proceso del mundo real por el cual algunos datos se registraron y otros no. La distribución de R es mejor vista como una estrategia matemática para describir las proporciones y patrones de valores faltantes y para captar aproximadamente posibles relaciones entre las faltas y los propios valores de los items faltantes (Collins, Schafer y Kam, 2001).

Schafer y Graham (2002) consideran esquemáticamente las distribuciones de R de acuerdo a la naturaleza de la relación de las faltas con los datos, basándose en los conceptos de Rubin (1976) y analizan cada categoría relacionándola con distintos esquemas de pérdida. Denominando Y a la matriz de datos, unidades por variables; Y_{obs} la submatriz de datos observados e Y_{miss} la submatriz de datos no observados, las distribuciones de R se clasifican:

- Si la distribución de la falta no depende de Y_{mis}: $P(R/Y) = P(R/Y_{obs})$; la probabilidad de pérdida en la variable no está relacionada al valor de la variable y los datos están faltantes al azar o MAR.
- Si la distribución de la falta no depende tampoco de los datos observados: $P(R/Y) = P(R)$; los datos están faltantes completamente al azar o MCAR, como caso particular de MAR.
- Si la distribución de las faltas depende de Y_{mis}, los datos están faltantes no al azar o MNAR.

Si los datos son MAR y los parámetros que gobiernan el proceso de falta no están relacionados con los parámetros que se desean estimar, el mecanismo se denomina ignorable, lo que significa que no es necesario modelizar el mecanismo como parte del proceso de estimación. Es difícil imaginar en la realidad casos en que el último requerimiento no se cumpla, por lo que se pueden considerar MAR e ignorable como condiciones equivalentes. Si los datos son MNAR es también denominada no respuesta no ignorable (Little y Rubin, 1987; Allison, 2002).

La diferencia entre MCAR, MAR y MNAR no se debe a la aleatoriedad en sí, sino a la relación entre variables de interés y variables que explican las faltas (Collins, Schafer y Kam, 2001; Schafer y Graham, 2002).

"Suponer MCAR significa básicamente que creemos que los datos observados son una submuestra aleatoria de una muestra y que podemos analizarlos como analizamos la muestra, solamente con un tamaño más reducido" (Meng, 2000). Como la probabilidad de que falten valores de Y no está relacionada con los valores de Y misma o con los valores de cualquier otra variable en el conjunto de datos, el supuesto MCAR permite la posibilidad de que la falta en Y se relacione con la falta en alguna otra variable X, por ejemplo, si la gente que se niega a dar su edad invariablemente se niega a dar su ingreso, aún así puede considerarse MCAR. El supuesto MCAR sería violado si la gente que no declaró su ingreso fuera más joven en promedio que la gente que sí lo dio. Si no hay diferencias se dice que los datos fueron observados al azar (OAR). Pero para ser MCAR todavía se tendría que cumplir que no hay relación entre la pérdida de una variable y los valores de esa variable (Allison, 2000).

Es posible probar estadísticamente si el mecanismo es MCAR contra la hipótesis alternativa de que sea MAR (Little,1988; Diggle et al., 2002). MCAR es un supuesto muy exigente y sucede en la realidad que las faltas frecuentemente dependen de las variables registradas. Un supuesto menos restrictivo es que las faltas dependan sólo de los valores observados y no de los valores faltantes como en MAR.

Si los datos son MAR es importante saber qué variables están relacionadas con la falta; para un tratamiento adecuado estas variables deberían ser incluídas en el ajuste o modelo de análisis. Con MAR las faltas son un proceso aleatorio condicional a los datos observados, o sea que los datos faltantes son una muestra aletoria de todos los valores dentro de clases definidas por valores observados (de Leeuw et al., 2003). MAR es un supuesto asumido por muchos métodos de tratamiento, pero cuando la falta no está bajo control del investigador, su distribución es desconocida y MAR es solamente un supuesto; no hay forma de probar el supuesto MAR, a excepción de realizar un seguimiento de los no respondentes o imponiendo un modelo no verificable (Little y Rubin, 1987). En muchos casos se pueden esperar alejamientos de MAR, pero no es fácil establecer la relación de los mismos con la continuidad de la validez de los tratamientos que requieren dicho supuesto. Trabajos como el de Collins, Schafer y Kam (2001) y Schafer y Graham (2002) se ocupan del tema. En algunos casos se sabe que el supuesto MAR se verifica, como en faltas planeadas y luego considerados datos faltantes: diseños de cohortes secuenciales, uso de cuestionarios conteniendo múltiples conjuntos de variables, etc..

Cuando los datos son MNAR los sesgos pueden ser de importancia y debe postularse un modelo para las faltas e incluirlo en el análisis para prevenir los sesgos. Hay casos en que el mecanismo es conocido y en otros, desconocido por el analista.

4.3. Estrategias de acercamiento al conocimiento del mecanismo

El tipo de mecanismo asociado a los datos no es fácil de establecer. Para su aproximación es importante estudiar el esquema como se especificó, aunque muchas veces su inspección no basta para saber con certeza si la falta es independiente de los valores de la variable, en cuyo caso será necesario información extra para acercarse al mecanismo. Esta información adicional puede provenir de otras fuentes distintas a la misma base de datos, como teoría, lógica, datos previos, re entrevistas, opinión de los relevadores, etc. (de Leeuw et al., 2003).

El conocimiento de las razones causantes de la falta de información ayudan a aproximar el mecanismo. A partir de su clasificación, de Leeuw et al. (2003) proponen que:

-la información no provista puede ser asociada a los tres mecanismos según los casos:

cuando es accidental podría considerarse MCAR; si hay deseo pero imposibilidad de respuesta podría considerarse MAR relacionándola a la categoría de unidades a la que se asocia la imposibilidad; en el caso que haya un rechazo o negación a la respuesta probablemente sea MNAR.

-la información que se perdió en general correspondería a un mecanismo MCAR y

-la información confusa resulta problemática respecto al mecanismo pero en general es asociable a MNAR aunque muchas veces con mecanismo conocido.

"Cuando hay datos faltantes debido a razones fuera de nuestro control, debemos hacer supuestos sobre el proceso que los ha producido. Estos supuestos en general no se pueden probar. La práctica científica sugiere que los supuestos sean hechos explícitos y la sensibilidad de los resultados al apartarse de los supuestos, sea investigada. Uno espera que similares conclusiones seguirán a una variedad de supuestos realísticos alternativos; cuando ello no sucede, la sensibilidad debería ser informada" (Collins, Schafer y Kam; 2001). El tipo de aleatoriedad de la pérdida en los datos de una base requerirá estrategias específicas en su tratamiento (de Leeuw et al., 2003).

5. LOS MÉTODOS Y SUS SUPUESTOS

Todos los métodos para tratar pérdida de datos dependen críticamente de importantes supuestos para su validez. Dichos supuestos se refieren fundamentalmente a la distribución de las variables y al esquema y mecanismo de las pérdidas. Según los métodos y su desarrollo histórico, los supuestos ligados a los métodos pueden haber sido explícitamente especificados en su formulación (como MAR para imputación múltiple) o implícitos como condición para su desempeño eficiente (como MCAR para imputación por el promedio).

En general, la no coincidencia entre las características de los datos a tratar y los supuestos a partir de los cuales ha sido formulado un método, conduciría a resultados de los análisis incorrectos o engañosos. La importancia de estos sesgos dependerá de cómo sea el supuesto y de la intensidad con que no se cumpla. Frecuentemente, no hay manera de comprobar si estos supuestos son satisfechos (Allison, 2001)

Sucede a veces que la aplicación del método puede demostrar que el apartarse en alguna medida de las condiciones iniciales, no afecta notablemente a los resultados. En estos casos se habla de robustez del método, vista como su sensibilidad a la violación de los supuestos sobre los cuales fue formulado. Si los supuestos de los métodos de tratamiento aplicados no coinciden con las características de los datos y de las pérdidas observados en la base, se deberían considerar los resultados de aplicaciones y experimentaciones para observar la robustez del método utilizado.

La metodología disponible en la literatura estadística es vasta. En las últimas décadas han ido ganando lugar las numerosas propuestas metodológicas para incorporar el mecanismo de pérdida; mucha investigación se ha dedicado a una mejor comprensión y modelización de mecanismos de datos faltantes de la vida real para controlarlos y para incorporarlos al proceso de estimación (Collins, Schafer y Kam, 2001).

6. EVALUACIÓN PARA LA SELECCIÓN

Un tratamiento para datos faltantes no puede ser correctamente evaluado fuera de la modelización, estimación o procedimiento de pruebas en el cual esté inserto. Si la muestra contiene valores faltantes, el método para tratarlos debería considerarse parte del procedimiento general para calcular las estimaciones de los parámetros de interés. No se debe perder de vista que si el objetivo de un procedimiento estadístico es lograr inferencias válidas y

eficientes, los intentos por recuperar los datos faltantes pueden alterar la inferencia; por ejemplo reemplazar con la media distorsiona la estimación de las variancias y correlaciones (Schafer y Graham, 2000).

En este sentido es fundamental conocer el efecto que pueden producir (o producen) los tratamientos en la estructura de los datos, y por consiguiente en la estimación de parámetros a partir de ellos y como consecuencia en los análisis en los que ellos estén involucrados.

Para la selección del método a aplicar resulta fundamental la compatibilidad entre el mecanismo de pérdida de los datos y el mecanismo sobre el que fue formulado el método. "La performance de cada método de tratamiento depende fuertemente del mecanismo de pérdida, que se refiere a las razones por las cuales los valores faltan, y en particular si la falta depende de los valores de variables en la base de datos. Muchos métodos de tratamiento suponen mecanismos MCAR o MAR y conducen a estimaciones sesgadas cuando los datos no son MAR sino MNAR" (Little y Hyonggin, 2003) .Entonces resulta de utiidad tratar de averiguar: ¿qué consecuencias presenta el no cumplimiento de dichos supuestos cuando se aplica a datos que no los cumplen?, ¿son robustos los métodos?, ¿hay evidencias de resultados de la aplicación del método de tratamiento cuando no se cumplen los supuestos sobre los cuales fue desarrollado?.

6.1. ¿Cómo se opera en la evaluación?

En general, los analistas y autores evalúan los métodos de tratamiento utilizados en forma experimental, a través de los resultados obtenidos al aplicar ténicas de análisis estadístico a posteriori del tratamiento de la información faltante a bases de datos en las que se han generado pérdidas o simulado repetidamente las mismas, en proporciones que permitan observar su incidencia en la estructura de los datos. De esta manera se pueden comparar también los resultados con los obtenidos a partir de la base completa.

La validación de un método se realiza a partir de escenarios definidos por el tamaño de muestra, la distribución probabilística asociada a los datos, las relaciones observadas entre variables, la proporción de pérdidas, el/los mecanismo/s con que se generan las pérdidas y el análisis estadístico replicado sobre las simulaciones. Collins, Schafer y Kam (2001) afirman que las conclusiones que pueden extraerse de su estudio están limitadas a escenarios similares al utilizado por ellos y no pueden ser generalizados a otros escenarios".

Se considera también la robustez como otro factor de validación, en los casos en que se generan las pérdidas según mecanismos y/o características de las variables, diferentes a los que sustentan originalmente al método aplicado. Cuando son varios los mecanismos simulados, se mantienen proporciones similares de pérdida para no introducir un nuevo factor de variación.

Se comparan así métodos entre sí, según los objetivos de cada trabajo: generalmente casos completos con otros métodos, o métodos convencionales respecto a estimaciones directas, imputación múltiple respecto a estimaciones directas o convencionales, o propuestas puntuales provenientes de distintos escenarios respecto a las ya existentes en la bibliografía. Así expresan Ambler et al. (2007): "El objeto de este trabajo es investigar un número de métodos para imputar datos faltantes para evaluar sus efectos en la estimación de modelos de riesgo y en sus predicciones. Se investigan métodos de imputación multiple, incluyendo hotdeck e imputación múltiple con ecuaciones en cadena (MICE) junto con otros métodos de imputación simple…".

La incorporación de la información proporcionada por variables auxiliares es valorada como cualidad importante de los métodos de tratamiento. "Todos los ajustes por datos faltantes requieren modelización de supuestos que relacionan los datos faltantes a las covariables

observadas. La sensibilidad a los supuestos es un asunto particularmente serio para el análisis que incorpora covariables útiles para los ajustes de los datos faltantes" (Little y Hyonggin, 2003; Collins, Schafer y Kam, 2001).

Otro aspecto considerado para la evaluación de los métodos es su operatividad, a la que se une un factor fundamental como es el de la disponibilidad o no de software ad hoc y su acceso: incluidos como procedimientos de softwares de utilización generalizada (SAS, SPSS, etc.), o formulados especialmente para el tratamiento (gratuitos o no), otros con procedimientos computacionales desarrollados y transmitidos por sus autores.

Para el tratamiento de casos reales, como los datos faltantes son realmente desconocidos (?), es fundamental el esfuerzo en lograr un acercamiento al conocimiento del mecanismo de pérdida de los datos, una elección del tratamiento con conocimiento de sus bases teóricas de formulación y la consideración de las conclusiones de las experiencias y evaluaciones del mismo.

6.2. ¿Qué indicadores se utilizan en la evaluación?

El indicador utilizado para la evaluación de los tratamientos, en general surge del análisis estadístico aplicado a posteriori del tratamiento y se relaciona con la estimación de sus parámetros, que en general se reitera para incorporar la variabilidad de las estimaciones.

A partir de los criterios clásicos de evaluación de procedimientos estadísticos, se pretende que el sesgo (diferencia entre el valor promedio del estimador del parámetro y el verdadero valor del parámetro) y la variancia del estimador sean pequeños (Schafer y Graham, 2002). Ambler et al. (2007): "MICE estuvo entre los métodos de imputación múltiple con mejor desempeño con respecto a la calidad de las predicciones. Además produjo las estimaciones menos sesgadas con buena cobertura …"; Brand et al. (2003): "… Describimos un método de simulación que profundiza varios aspectos del sesgo y la eficiencia del proceso de imputación múltiple".

Collins, Schafer y Kam (2002) agregan que, a pesar de que la mayoría de la literatura sobre datos faltantes enfatiza la evaluación de los procedimientos para datos faltantes en términos del sesgo, encuentran importante examinar otros indicadores luego de los experimentos pues los efectos pueden diferir a través de los distintos criterios: el RMSE (raíz del error cuadrático medio); el sesgo estandarizado (expresión del sesgo como un porcentaje del error estándar, es decir cociente entre la diferencia de la estimación promedio del parámetro a través de las simulaciones con el verdadero valor del parámetro y el error estándar); la cobertura como porcentaje de intervalos y la amplitud promedio de los intervalos de confianza.

Otros autores utilizan para la evaluación indicadores específicos definidos en la formulación de la técnica de tratamiento; para la aplicación de IM, la eficiencia relativa (RE) (Yuan, 2000) o los estimadores que el mismo método define como variación intra imputación, entre imputación o la cobertura del intervalo de confianza del 95% resultante de imputaciones repetidas (Brand et al., 2003).

6.3. Los autores recomiendan a partir de sus evaluaciones

En este trabajo se consideran métodos que son más frecuentemente utilizados por los autores en sus aplicaciones y evaluaciones (Horton y Kleinman, 2007) y que los softwares presentan como opciones de distintos análisis, con procedimientos bastante estandarizados. Sin embargo no se deja de tener en cuenta que continuamente surgen propuestas para el tratamiento de la información faltante en escenarios definidos por el tipo de variable, las distribuciones estadísticas y/o la técnica estadística a aplicar sobre los datos y a partir de áreas temáticas donde el problema de los datos faltantes se ha planteado.

Se presentan los métodos considerados mediante una breve descripción y los supuestos sobre el mecanismo asociado. Entre los denominados convencionales: casos completos, casos disponibles, imputación por el promedio y por regresión; entre los más recomendados: imputación múltiple y métodos basados en verosimilitud; como una de las nuevas propuestas, procedimientos de ponderación. Una última categoría se plantea en forma diferente, a partir del mecanismo, no ignorable, de igual manera que en general la presenta la bibliografía.

Se incorporan recomendaciones de distintos autores sobre aspectos de la metodología a partir de experiencias en la aplicación, que pueden ser de utilidad ante la elección adecuada del tratamiento.

*** Relativo a métodos convencionales:**

Casos completos ("Case deletion", "listwise deletion" o "complete-case analysis"): se utilizan sólo los casos de la base que no tienen faltas en ninguna variable. Presenta virtudes y defectos. Permite disponer de una matriz de datos rectangular, sirve para cualquier tipo de variable y análisis estadístico, no requiere métodos computacionales especiales y muchos procedimientos de análisis de los softwares lo consideran como método por defecto.

Disminuye el total de observaciones previstas, pudiendo ser problemático para la inferencia pues si la fracción de faltas es importante, la pérdida de eficiencia del estimador CC puede ser sustancial. Si la pérdida es MCAR, los estimadores CC serán no viciados; tendremos una submuestra aleatoria de la muestra original, las estimaciones que no eran viciadas para el conjunto de datos completos tampoco lo serán con la submuestra, a pesar de que los errores estándar serán mayores por la menor cantidad de información utilizada, incluso en general mayores que usando otros métodos de tratamiento. Si la pérdida es MAR (supuesto más realista) puede conducir a estimaciones viciadas (Allison, 2002; Horton y Kleinman, 2007; Little y Rubin, 2002).

Observa Allison (2002): "Los métodos convencionales como casos completes o imputación por regresión, están afectados por tres serios problemas: uso ineficiente de la información disponible, reduciendo el poder de los tests y errores de tipo II; estimaciones sesgadas de los errores estándar, llevando a valores incorrectos del "p-value"; estimaciones sesgadas de los parámetros, debido a no poder ajustar para la selectividad en datos faltantes".

La pérdida de precisión y sesgo de las estimaciones a partir de **casos completos** dependen no sólo de la fracción perdida, del patrón y de las diferencias entre casos completos e incompletos sino también del análisis estadístico aplicado y de los parámetros estimados. Con MCAR la estimación del promedio es insesgado; la de los coeficientes de regresión son insesgados si la probabilidad de ser observado depende de las variables independientes pero no de la dependiente, incluyendo el mecanismo NMAR en el que la probabilidad de una covariable faltante depende del valor de esa covariable y serán sesgados si la probabilidad de ser completa depende de la variable dependiente (Little y Rubin, 2002).

Allison (2002) también lo analiza en el ámbito del análisis de regresión: "Casos completos es el método más robusto para violaciones de MAR entre variables independientes en un análisis de regresión. Hay casos en que CC da mejores estimaciones incluso que máxima verosimilitud e imputación múltiple…En regresión logística, puede tolerar tanto pérdidas no aleatorias en la variable dependiente o en las variables independientes, pero no en ambas".

Casos completos es descartado muchas veces por los sesgos que produciría en las estimaciones si no se cumple el supuesto MCAR, pero los métodos que se presentan como mejores alternativas (IM y MV) también requieren el cumplimiento del supuesto MAR. Lo que se recomienda en general es que se usen métodos que usen toda la información rele-

vada o disponible y eso es tal vez lo que se puede ver como peor defecto de casos completos (Allison, 2001), mientras los otros dos métodos permiten su incorporación (Collins, Schafer y Kam, 2001).

Schafer y Graham (2002) observan que, aunque se afirme que las estimaciones con **casos completos** son sólo válidas bajo MCAR, se deben considerar distintas situaciones. En primer lugar, la medida en que se aparten los datos de este supuesto tendrá distintos impactos sobre los sesgos (aunque sea difícil mensurarlo); hay situaciones en que produce óptimas inferencias bajo MAR (por ejemplo cuando las faltas son de una variable y se estima la regresión de la variable incompleta con respecto a las completas), aunque esta situación no se repite si se estiman medidas de asociación con la variable incompleta o los parámetros de la marginal de la misma; aunque se cumpla MCAR el método puede aún ser ineficiente en la estimación dependiendo de qué parámetro y de la relación entre las variables completas e incompletas. Repitiendo una experiencia similar concluyen que imputación con el promedio y hotdeck producen estimaciones sesgadas de muchos parámetros para todos los mecanismos; imputación con regresión, sesgos menos importantes.

Casos disponibles ("Available-case analysis") : se utilizan para el análisis los casos que no tienen faltas en las variables involucradas en cada cálculo del estimador que se está calculando. El tamaño de la base cambiará de variable a variable de acuerdo al esquema de pérdida tratando de incorporar más información que casos completos. Muchos procedimientos de análisis en los softwares permiten usarlo para sus estimaciones.

Está formulado en base al supuesto de que las faltas no dependen de los valores de los datos (MCAR) en cuyo caso produce estimaciones consistentes de promedios y variancias, pero se requieren modificaciones para estimar covariancias o correlaciones. Si los datos son sólo MAR pero no observados al azar, las estimaciones pueden ser seriamente viciadas.

Una derivación es el método de casos disponibles "pairwise", en el que las medidas de covariación se basan en casos en que ambas variables involucradas están presentes.

Se esperaría que las estimaciones con **casos disponibles** fueran más eficientes que casos completos**,** conclusión a que se arriba por simulación cuando los datos son MCAR y las correlaciones modestas, pero no es cierto cuando las correlaciones son altas. Los problemas importantes con casos disponibles son que los errores estándar estimados y los test estadísticos producidos por software convencional son viciados, pues el tamaño muestral no es único, sino que depende de las variables que se están usando en ese momento, y además que la matriz de correlaciones puede no estar definida positiva lo que implica que no se pueden calcular las regresiones. En general no es una alternativa válida ante casos completos excepto para la estimación de algunos parámetros que se vean favorecidos por la incorporación de unidades (Little y Rubin, 2002; Allison, 2002).

Imputación simple: los métodos de imputación reemplazan los datos faltantes por estimaciones en base a modelos implícitos o explícitos y son utilizados tanto en datos transversales como longitudinales (Little y Rubin, 2002). Son numerosas las propuestas y un criterio de clasificación las divide en estocásticas y determinísticas, según incorporen la generación de números aleatorios o no (Schulte Nordholt,1998).

El aspecto atractivo de los métodos de imputación es que ellos producen un archivo de datos que es completo, requerida por la mayoría de los métodos estadísticos a aplicar (de Leeuw, 2003). Como principal inconveniente es que subestiman la variancia de los estimadores aún si el modelo de imputación utilizado es el correcto y que muchos de ellos distorsionan la distribución y las relaciones de los datos.

Una propuesta muy utilizada es la imputación determinística **con el promedio**: los valores faltantes de una variable son reemplazados por el promedio de los valores observados para

la misma variable. Bajo MCAR subestima la variancia y las covariancias. Una variación muy utilizada para lograr mejor aproximación, imputa con el promedio de la categoría de la variable a la que pertenece la unidad no observada (Class mean imputation).

Puede verse como una generalización de esta última a la **imputación por regresión** (Conditional mean imputation), en la que se generan los valores perdidos a partir del modelo de regresión múltiple construido con los datos completamente observados, considerando a la variable con pérdidas como dependiente y seleccionando las independientes del resto de las variables. Depende del supuesto de que las faltas se relacionan sólo con los datos observados (MAR), que permite ajustar las estimaciones usando información disponible. La imputación por regresión será estocástica si se agrega al modelo un residuo dependiendo de la distribución.

La imputación con el promedio y por regresión están disponibles en muchos procedimientos de los softwares como soluciones alternativas a casos completos.

Entre otros métodos de imputación simple son muy utilizados hot deck (determinístico y estocástico), cold deck o combinación de métodos.

Muchos de los métodos de imputación producen sesgos en las estimaciones, aún cuando los datos sean MCAR. En este sentido es preferible casos completos, siempre que la fracción perdida no sea importante (Allison, 2001). Utilizando imputación por el promedio o por regresión los errores estándar son subestimados y no reflejan la incertidumbre asociada a los valores faltantes; producen estimaciones viciadas de las variancias y covariancias (Allison, 2002).

"El método de **regresión** se complica cuando hay más de una variable con faltas. Si las imputaciones se basan solamente en otras variables independientes y no en la dependiente (si el análisis a aplicar es de regresión), y si los datos son MCAR, los coeficientes mínimos cuadrados son consistentes, implicando que son aproximadamente no viciados en largas muestras, pero no totalmente eficientes. Se pueden lograr estimadores usando mínimos cuadrados ponderados o generalizados" (Allison, 2002).

La imputación aleatoria puede eliminar los sesgos de la determinística a través de extracciones aleatorias de la distribución residual de cada variable imputada y agregar estos números aleatorios a los valores imputados, permitiendo así el uso de fórmulas convencionales para calcular variancias y covariancias. Sin embargo todos los métodos de imputación presentan un problema: si usamos datos imputados (tanto aleatorios como determinísticos) como si fueran reales, las estimaciones de los errores estándar serán generalmente muy bajos y no reflejan la incertidumbre asociada a los valores faltantes.

"Casos completos es un derroche debido a que descarta información. Casos disponibles puede resultar en totales y matrices de correlación inconsistentes. Ambos métodos es probable que sean sesgados. Imputación simple puede producir estimaciones puntuales sesgadas y subestimará a las verdaderas variancias muestrales, resultando en p-values falsamente pequeños. Además, se debería tomar conciencia de que estas ténicas están todas basadas en supuestos muy fuertes como MCAR" (de Leeuw, 2003).

*** Relativo a Imputación Múltiple (IM):** El método de imputación múltiple (Rubin, 1987), presenta una solución al problema planteado por imputaciones simples. Repite el proceso de imputación, produciendo múltiple conjuntos de datos "completados", incorporando variabilidad entre las estimaciones de parámetros, que puede ser utilizada para ajustar los errores estándar (Rubin, 1987). El método supone en su versión original que los datos son MAR, supuesto más débil que MCAR y bajo este supuesto MI produce estimaciones consistentes y asintóticamente eficientes.

Allison (2000) resume: "IM tiene muchas ventajas: poder ser usado con cualquier tipo de

datos y para cualquier análisis, introducir error aleatorio en el proceso de imputación y hacer posible estimaciones aproximadamente no viciadas de los parámetros, hechos no posibles con los métodos de imputación simples en general. Pero para ello se deben cumplir ciertos requerimientos: mecanismo de pérdida MAR, modelo de imputación correcto y en algún sentido coincidente con el modelo usado por el analista".

Los procedimientos para crear imputaciones múltiples están basados en modelos; debe ser especificado un modelo de imputación especificando una probabilidad conjunta para los datos observados y faltantes, con flexibilidad o generalidad suficiente para preservar los efectos de interés para el análisis subsiguiente (Collins, Schafer y Kam, 2001). MI extrae cada vez una muestra aleatoria de la distribución elegida para realizar la imputación. En la práctica se trabaja en general con modelos estándar, que son fáciles de usar y dan aproximaciones buenas para muchos conjuntos de datos como el normal multivariado, supuesto utilizado por la generalidad de los softwares. Estos también suponen que las faltas son MAR pero el usuario no puede en general introducir ninguna información sobre la forma en que la probabilidad de las faltas puede estar relacionada a las variables en el sistema.

Además MI requiere una distribución a priori para los parámetros, ya que IM está formulada en un marco bayesiano en el que los parámetros desconocidos son vistos como aleatorios. Todos los softwares por defecto establecen una distribución a priori "difusa" para evitar la introducción de sesgos y permitir que los datos hablen por sí mismos (Collins, Schafer y Kam, 2001).

La inferencia con IM comprende sintéticamente tres pasos: los datos faltantes son imputados m veces, creando m conjuntos de datos; cada uno de ellos son analizados con la técnica estadística clásica propuesta y los m resultados obtenidos son combinados para producir los resultados de la inferencia. El número de imputaciones puede ser establecida informalmente replicando imputaciones y controlando si las estimaciones son estables (Horton y Lipsitz, 2001).

Los avances computacionales significativos han ayudado a establecer a la imputación múltiple como un enfoque instalado y útil para una amplia variedad de problemas de datos incompletos (van Buuren y Eisinga, 2003) y numerosos programas computacionales lo implementan. Se enumeran métodos para las imputaciones en el contexto del procedimiento para IM del programa SAS (2002); ellos dependen de los esquemas de pérdida observados y están formulados sobre distintos supuestos distribucionales:

* Para esquemas monótonos y variables continuas, regresión y "predictive mean matching" que suponen una distribución normal multivariada de los datos; también propensity scores que es no paramétrica.

* Para esquemas monótonos y variables ordinales, el método de regresión logística.

* Para esquemas monótonos y variables nominales, el método de función discriminante.

* Para esquemas arbitrarios, MCMC (Monte Carlo con cadenas de Markov) que supone normalidad multivariada (relacionado a la inferencia bayesiana que usa). Puede también ser usado para obtener un esquema monótono y continuar el análisis.

Las inferencias con MI suponen que el modelo que se usa para analizar los datos imputados (el modelo del analista), es el mismo que el modelo que se usa para imputar los valores faltantes con imputación múltiple (modelo del "imputador"), pero en la práctica ambos modelos pueden no ser los mismos; las consecuencias dependen de los diferentes escenarios (que el analista suponga más relaciones que el imputador o viceversa) por lo que se aconseja incluir tantas variables como sea posible para realizar las imputaciones múltiples (Rubin, 1996), pues la precisión que se pierde cuando se incluyen predictores no útiles no es tan importante frente a la pérdida de validez de los análisis con conjuntos de datos imputa-

dos múltiplemente con modelos incorrectos que pueden conducir a conclusiones falsas (SAS, 2002). Al mismo tiempo, es importante mantener el número de variables bajo control (Barnard y Meng,1999). Para la imputación de una variable particular, el modelo debería incluir variables en el modelo de datos completos, que estén correlacionadas con la variable imputada y/o que están asociadas con la pérdida de la variable imputada (Schafer, 1997; van Buuren, Boshuizen y Knook, 1999).

van Buuren (2007): "El proceso de especificación del modelo de imputación es en sí misma una actividad científica de modelización, que viene con sus propios principios de construcción de modelos. El hecho de que estén disponibles procedimientos altamente automatizados y sofisticados, no libera al analista o "imputador" de la responsabilidad de considerer si son apropiados los supuestos en que se basa el modelo de imputación para el problema que está siendo considerado…..los investigadores deberían incluir una pequeña descripción de su método de datos faltantes en sus artículos científicos...en la sección de análisis estadístico".

van Buuren y Eisinga (2003): "El análisis basado en modelos se refiere a la situación donde se necesita un modelo específico para los datos faltantes, por ejemplo si los datos faltantes no son MAR. IM aún funciona en este caso, pero el énfasis se centra en la especificación del modelo que creó los datos faltantes".

Horton y Lipzitz (2001): "Los métodos de IM pueden ser más eficientes (con el costo de hacer supuestos relativos a la distribución de las pérdidas y al modelo de imputación), comparándolo con los estimadores de casos completo (CC) que descartan las unidades parcialmente observadas. Si el supuesto MAR es sostenible, la imputación múltiple puede también presentar menos sesgo que otros enfoques si el modelo de imputación es correctamente especificado… Se necesitaría investigación adicional para analizar el sesgo resultante de un modelo de imputación pobremente especificado (por ejemplo usando una distribución normal cuando la variable posiblemente faltante es Bernoulli).

*** Relativo a estimación directa:** Una clase de procedimientos se genera definiendo un modelo para los datos observados y basando las inferencias en la verosimilitud bajo dicho modelo, estimándose los parámetros con procedimientos como máxima verosimilitud. Se analiza toda la información disponible, en base a la información tanto de las unidades incompletas como la de las completas (Little y Rubin, 2002).

El enfoque máximo verosímil (**MV**), es particularmente apropiado para ser utilizado en casos de información faltante, pero bajo el cumplimiento de ciertos supuestos. Cuando el mecanismo de pérdida es **ignorable**, y por consiguiente **MAR**, se pueden obtener las verosimilitudes simplemente sumándolas (o integrando para variables continuas) a través de todos los posibles valores de los datos faltantes. El problema entonces es encontrar los valores de los parámetros que maximicen la verosimilitud; una variedad de métodos resuelven este problema de optimización. MV es particularmente sencillo cuando el patrón de pérdida es **monótono** (Little y Rubin, 1987 y 2002; Allison 2002).

Para la mayoría de las aplicaciones de MV a los problemas de falta de información no hay soluciones explícitas para las estimaciones por lo que son necesarios métodos iterativos. El algoritmo EM (Dempster et al., 1977) es un método iterativo que permite obtener las estimaciones máximo verosímiles incluso para el caso de esquemas generales. Depende del supuesto MAR y usa toda la información disponible (por ejemplo todas las variables disponibles como predictoras para imputar los datos faltantes), sin embargo presenta desventajas como que los errores estándar y los test estadísticos brindados por los softwares de modelos lineales pueden no ser correctos y estimaciones no completamente eficientes para modelos sobre identificados (o sea con restricciones en la matriz de covariancias) (Little y Rubin, 1987 y 2002; Allison 2002; SPSS, 1997).

MV se basa en supuestos cruciales: la muestra debe tener tamaño suficiente para que las estimaciones sean aproximadmente insesgadas y normalmente distribuidas; dependiendo de la aplicación particular, los métodos verosímiles pueden o no ser robustos cuando se apartan del supuesto del modelo; para algunos casos la estimación puede ser posible al apartarse del modelo pero requerirán el supuesto MCAR (Schafer y Graham, 2002).

Son numerosos los software que permiten implementar procedimientos MV. Obervan Collins, Schafer y Kam (2001): "Con las herramientas computacionales existentes, ajustar un modelo con datos incompletos no es operacionalmente más difícil que ajustarlo con datos completos. La facilidad en la utilización de estas herramientas parecería haber una generalizada falsa idea de que la aplicación de MV releva al usuario de pensar detalladamente en su aplicación pues todos los ajustes necesarios se hacen automáticamente. Los investigadores deberían estar atentos a que estos ajustes son satisfactorios sólo cuando los datos y esquemas de pérdida satisfacen los supuestos subyacentes, en particular el supuesto MAR. Si se omiten importantes causas o correlaciones del modelo, a´´un cuando ellas sean conceptualmente irrelevantes para el propósito del analista, entonces la estimación de los parámetros MV puede ser sesgada"

MV puede ser aplicado a tablas de contingencia o para estimar distintos modelos lineales (regresión lineal, análisis factorial, ecuaciones simultáneas, ecuaciones estructurales con variables latentes, etc.) bajo el supuesto de que los datos se distribuyen según una normal multivariada.

El **EM** evita uno de los problemas de la **imputación por regresión convencional** (decidir qué variables usar como predictoras ante el hecho de que diferentes patrones de faltas tienen diferentes conjuntos de predictores disponibles). Como EM empieza siempre con la matriz de covariancias completa, es posible obtener estimadores de regresión para cualquier conjunto de predictores sin importar qué tan pocos casos pudieran haber en un esquema particular de pérdida pues usa siempre todas las variables disponibles como predictoras para los datos faltantes. Entonces los promedios y covariancias de todas las variables pueden ser obtenidos como funciones de los estimadores de los parámetros de las regresiones mencionadas (Little y Hyonggin, 2003).

*** Relativo a IM y MV**: En general los métodos más recomendados por todos los autores son estimaciones máximo verosímiles e imputación múltiple. "Máxima verosimilitud e imputación múltiple tienen propiedades estadísticas muy similares. Si los supuestos se cumplen, brindan resultados aproximadamente no sesgados y eficientes, o sea con mínima variancia muestral. Lo destacable es que estos métodos dependen de supuestos menos restrictivos que los requeridos por los métodos convencionales. En el presente, máxima verosimilitud es más adecuado para modelos lineales o loglineales para tablas de contingencia; imputación múltiple puede ser utilizado para cualquier problema estadístico" (Allison, 2002).

Rubin (2001) afirma: "En general IM posee más robustez a supuestos distribucionales que los métodos basados en verosimilitud". MV requiere un modelo, pero IM probablemente es menos sensible que MV a la elección del modelo pues éste es utilizado sólo para imputar, no para estimar otros parámetros. Desde un punto de vista operacional IM es algo diferente a MV pues los datos faltantes son tratados en un paso totalmente separado del análisis. Las consecuencias de esta separación son negativas y positivas: el analista puede proceder a analizar los datos imputados sin tener en cuenta la forma en que las imputaciones fueron creadas, pudiéndose producir sesgos: por ejemplo la imputación pudo haber producio interacciones más débiles que las de la población. En el lado positivo, al imputar separadamente se facilita la incorporación de variables auxiliares que son automáticamente consideradas en subsiguientes análisis (Collins, Schafer y Kam, 2001).

Hay muchas situaciones en que IM y MV producen resultados similares, dependiendo de la

forma en que interactúan tres modelos: el modelo que subyace en el procedimiento MV, el modelo que subyace en el procedimiento IM (usado para generar la imputaciones) y el modelo utilizado para analizar los conjuntos de datos imputados. Ellos difieren en su alcance; el último usa supuestos distribucionales sobre la población de los datos completos, mientras los dos primeros hacen supuestos adicionales sobre las faltas (típicamente MAR) y IM incluye además la distribución a priori de los parámetros. Collins, Schafer y Kam (2001) fomulan este planteo y, a partir de él, experimentan sobre la interactuación, analizando los diferentes resultados obtenidos.

Las propiedades teóricas atractivas de MV e IM no necesariamente se trasladan en un buen desempeño de estos dos métodos en el análisis de datos reales, si no son aplicados a datos que cumplan con los requerimientos de cada uno o si los mecanismos que generan tanto los datos como las pérdidas se apartan sustancialmente de los supuestos estadísticos subyacentes. Ambos se han ido haciendo estándar debido a la implementación con software comercial y gratis, pero desde un punto de vista de operatividad parecería mejor imputación múltiple pues los softwares que permiten implementarlo contemplan más fácilmente la incorporación de variables auxiliares (Allison, 2002; Schafer y Graham, 2002; Con muestras pequeñas parecería que IM tiene un mejor desempeño que MV. Todos los trabajos realizados con IM han supuesto mecanismo MAR aunque hay algunos trabajos con MNAR. Cuando se utiliza el mismo modelo para imputación y análisis , IM produce resultados similares a lo de MV bajo el mismo modelo. Mucho de la fuerza, flexibilidad y tal vez peligro de IM, se debe a la posibilidad de usar dos modelos distintos (Schafer y Graham, 2002; Collins, Schafer y Kam, 2001).

Schafer y Graham (2002) destacan que tanto IM como MV son recomendables para tratar faltas en estudios longitudinales ya que se recomiendan procedimientos que usen toda los datos disponibles que en estos casos pueden ser obtenidos de ondas anteriores o posteriores.

*** Relativo al supuesto de normalidad:** En general los conjuntos de datos no tienen variables todas normales sino algunas muy asimétricas, otras categóricas. Qué pasa entonces con los métodos que hemos ido viendo basados en modelos normales? Para las variables sin faltas, no hay problemas porque nada se les imputa; para las variables con faltas hay cuantiosas evidencias de que los métodos de imputación pueden trabajar bastante bien aún cuando la distribución es no normal, aunque hay algunas técnicas que pueden mejorar el desempeño del modelo normal para imputar variables no-normales: para variables continuas muy asimétricas, es de utilidad en general transformar la variable para reducir la asimetría antes de hacer la imputación. Luego de imputar se puede aplicar la trasformación inversa para volver a la variable original.

van Buuren y Eisinga (2003)*:* "A pesar de que son condiciones fuertes, en la práctica el modelo normal multivariado parece hacer un buen trabajo de imputación aún cuando algunas de las variables tengan distribuciones manifiestamente no normales. Es completamente inocuo para las variables sin falta y para las que tienen, las transformaciones pueden mejorar mucho la calidad de las imputaciones". "Se puede usar una sentencia para transformar las variables y acercarse al supuesto de normalidad" (Horton y Lipzitz 2001).

"A pesar de que MCMC y regresión requieren normalidad multivariada, las inferencias basadas en imputación múltiple pueden ser robustas cuando se aparta la distribución de la normalidad multivariada, si la cantidad de información faltante no es grande. Frecuentemente tiene sentido usar un modelo normal para crear imputaciones múltiples aún cuando los datos observados no sean normales, hecho observado en estudios de simulación descriptos por Schafer (1997) y los que él cita.

Little y Smith (1987): "Aún cuando se conozca que algunas variables con faltas tengan dis-

tribuciones no normales (por ej variables dummy), las estimaciones ML bajo el supuesto normal multivariado frecuentemente tiene buenas propiedades, especialmente si los datos son MCAR".

*** Relativo a procedimientos de ponderación:** Las inferencias aleatorias a partir de datos de encuestas por muestreo sin no respuesta comunmente ponderan las unidades muestreadas con sus ponderaciones de diseño, que son inversamente proporcionales a sus probabilidades de selección. Para no repuesta se modifican las ponderaciones para realizar un ajuste por no respuesta como si ella fuera parte del diseño muestral.(Little y Rubin, 2002; Collins, Schafer y Kam, 2001).

La primer motivación de estos métodos es lograr robustez, sin embargo sus estimadores semiparamétricos pueden ser menos eficientes y con menos poder que estimadores MV o bayesianos con un modelo paramétrico bien especificado, con cualquier distribución de pérdida MAR, que el usuario no necesita especificar (Schafer y Graham, 2002), aunque se reconoce que los procedimientos de ponderación pueden ser útiles en determinadas circunstancias.

Han sido aplicados con distintas técnicas estadísticas y se espera que en muchos casos produzcan resultados similares a IM, bajo un adecuado modelo conjunto para la respuesta y covariables.

***Relativo a tratamientos para datos MNAR:** El tratamiento de datos con mecanismos no ignorables es particularmente dificultoso por la falta de uniformidad entre las características de las pérdidas que hacen de cada análisis un caso particular. Con mecanismo ignorable no es necesario formular un modelo para el proceso de pérdida; la estrategia general es realizar un ajuste para todas las diferencias observables entre casos faltantes y no faltantes y suponer que todas las demás diferencias son asistemáticas.

La gran diferencia es que "dado un modelo para los datos, hay un solo mecanismo ignorable para los datos perdidos pero hay infinitos mecanismos no ignorables para datos perdidos", por lo que no es casual la no proliferación de software "pues es difícil escribir programas que ejecuten aunque sea una fracción de las posibilidades" (Allison, 2002; Little y Rubin, 2002).

Dos enfoques diferentes son los planteados en general para agregar al modelo para los datos completos una distribución para la pérdida específicamente explícita: "selection models" (Amemiya, 1984; Heckman, 1976) y "pattern-mixture models" (Little, 1993).

IM y MV pueden ser utilizados para mecanismo no ignorable, pero requieren un modelo correcto (y reiterado) para el proceso por el cual los datos están faltando, lo que es usualmente difícil de lograr. Los dos métodos son muy sensibles a los supuestos hechos sobre el mecanismo de pérdida o sobre las distribuciones de las variables con faltas y no hay forma de probar estos supuestos.

El requerimiento más importante para datos MNAR es un conocimiento previo del mecanismo que genera las faltas (Allison, 2002; Schafer y Graham, 2002; Collins, Schafer y Kam, 2001).

La clave más recomendada para tratamientos e datos MNAR es medir covariables que sean predictivas de los valores faltantes y modelizar cuidadosamente la relación entre las variables faltantes y estas covariables. Métodos basados en verosimilitud, basados en modelos multivariados para los datos son herramientas útiles para hacer uso de los datos disponibles, pero modelos estándar como la normal multivariada implican relaciones aditivas entre las variables que pueden ser demasiado simplísticas en ciertos casos, por lo que Little y Hyonggin (2003) proponen métodos más específicos como modelos "spline", para obtener predicciones de regresión más robustas a la no linealidad en la relación entre las variables

faltantes y las covariables bajo el supuesto MAR. “Cuando el mecanismo es desconocido y NMAR, las opciones metodológicas son limitadas. En estudios en los cuales es probable que se presenten datos faltantes, se deberían hacer esfuerzos en pos de lograr un mecanismo MAR, evaluando covariables que caractericen a los no respondentes (Little y Rubin, 2002)”. Little y Hyonggin (2003), a partir de su propuesta para datos MNAR, recomiendan un acercamiento a las propuestas que requieren supuesto MAR: “A partir de lo visto, son útiles métodos basados en MAR y en modelos que hagan supuestos relativamente débiles relacionando las covariables con los datos faltantes. O sea un enfoque alternativo es intentar medir covariables que capturen diferencias entre respondentes y no respondentes, de manera que el mecanismo de pérdida pueda ser considerado MAR”.

Cualquier método para datos faltantes no ignorables debería acompañarse por un análisis de sensibilidad: como los resultados pueden variar mucho dependiendo del modelo supuesto, es importante probar distintos modelos posibles y analizar sus resultados. Por lo que es fundamental la elección del modelo correcto, lo que requiere el conocimiento más profundo que sea posible sobre el fenómeno analizado en los datos (Allison, 2002).

La modelización de MNAR se ve más adecuada para los casos (por ejemplo estudios clínicos) en que la falta o el abandono se ve claramente relacionada con la variable medida. En otras situaciones se ha visto con simulaciones (Collins et al., 2001), que cuando la causa verdadera de la falta es la variable en sí misma se puede introducir un sesgo en las estimaciones de parámetros, al fracasar en la forma de incorporación de MNAR (Schafer y Graham, 2002) .

Sucede en ciertos casos, como en los cuestionarios autogestionados, que las faltas MNAR suelen obedecer a un comportamiento no normativo o sea diferente para cada persona y es imposible expresarlo en un comportamiento común. En otros casos en que es previsible, los investigadores pueden mitigar sus efectos mediante cambios en el diseño del análisis (Schafer y Graham, 2002).

*** Relativo a la inclusión de variables auxiliares:** Las variables auxiliares pueden ser incluidas por dos razones: para incluir variables que potencialmente sean causa o estén correlacionadas con la falta (la variable edad si la falta se relaciona con los jóvenes) y aquellas que esén correlacionadas con las variables con faltas, estén o no relacionadas con el mecanismo (si hay faltas en ingreso y nivel educacional está completa, como ambas están correlacionadas, se puede incluir la última para recuperar en parte la información faltante; en estudios longitudinales es posible incluir como auxiliar una variable con faltas que fue observada en tiempos anteriores) (Collins, Schafer y Kam, 2001).

Collins, Schafer y Kam (2001) evalúan estrategias restrictivas (que incluyen pocas o ninguna variables auxiliares en el análisis) versus inclusivas (que incluyen numerosas variables auxiliares) y concluyen: “Con una estrategia de inclusión, no solamente hay una posibilidad reducida de omitir inadvertidamente una importante causa de falta, sino también la posibilidad de ganancias notables en términos de eficiencia incrementada y sesgo reducido. En teoría tanto MV como IM pueden ser usados para implementar esta estrategia. En la práctica, por el diseño de los software actuales, es más sencillo usar MI para esta estrategia; los investigadores se beneficiarían si el software para MV fuera revisado para facilitar el uso de esta estrategia”.

En el marco de la IM, mejoras en la eficiencia y en los sesgos pueden lograrse cuando se agregan variables auxiliares al proceso de imputación múltiple, aún cuando no sean incluidas en el posterior análisis de los datos imputados (Meng, 1994; Rubin, 1996). Es posible agregar variables auxiliares para MV, pero se debe tener en cuenta que agregarlas, si no se hace cuidadosamente, puede producir efectos no deseados (por ejemplo, alterar el signifi-

cado del modelo y redefinir los coeficientes poblacionales que se estiman) (Collins, Schafer y Kam, 2001).

*** En un análisis a partir de los mecanismos:** El Manual de SPSS (1997) resume convenientemente: "Bajo supuesto MCAR, casos completos, disponibles, EM y regresión dan estimaciones consistentes y no viciadas de las correlaciones y covariancias. Bajo MAR, casos disponibles, EM y regresión dan buenas estimaciones. Por ejemplo, si los datos faltantes de una variable pertenecen a una determinada categoría de otra variable y si para esa categoría las faltas de la otra variable son MCAR, casos completos, EM y regresión aún pueden dar estimaciones buenas. Si los datos son MAR y se cumple el supuesto de distribución normal, normal mixto o t con grados de libertad especificos, EM da estimaciones máximo verosímiles de los promedios, desvíos estándar, covariancias y correlaciones".

De Leuw et al. (2003) sintetizan de forma similar: "Cuando los datos pueden considerarse MCAR, las soluciones simples como casos completos y disponibles no resultan en sesgos por no respuesta. Para MAR existen métodos efectivos... por ejemplo, analizar los datos incompletos condicionando a información de covariables es un procedimiento útil para reducir el sesgo de no respuesta. Rubin, Stern y Vehovar (1995) recomiendan usar el supuesto MAR como un punto de partida para análisis en grandes encuestas. .. Si se suponen los datos MAR, dos estrategias de análisis pueden ser usadas: estimación directa e imputación".

*** Observaciones más globales:**

"Prevención y ajuste son dos caras de una moneda... Cuanto mayor cantidad de información se tenga, mejor se puede investigar el mecanismo y mejor se puede ajustar. Por lo tanto la mejor política es primero prevenir las faltas todo lo que se pueda y además recoger información auxiliar; segundo, debería ser utilizada una estrategia analítica de dos pasos: usar toda la información disponible para investigar los esquemas de datos faltantes y analizar el conjunto de datos incompletos realizando los ajustes para datos faltantes necesarios (y correctos)" (de Leeuw et al., 2003).

Se observa en las recomendaciones de muchos autores la tendencia a utilizar estrategias de adaptación de los datos a los supuestos de los métodos (transformaciones de variables, supresión de unidades, incorporación de información auxiliar, etc.) y de adaptación de los mecanismos de los métodos en general hacia MAR.

"Todos los métodos para tratar pérdida de datos, incluyendo imputación múltiple y estimación máximo verosímil, dependen críticamente de importantes supuestos para su validez, y frecuentemente, no hay manera de comprobar si estos supuestos son satisfechos. Lo mejor sigue siendo no tener datos faltantes o sea trabajar mucho en el diseño y en la prevención" (Allison, 2000).

7. CONCLUSIÓN

Si se deben analizar estadísticamente bases de datos con información faltante es fundamental estudiar:

* Las características de las variables y el esquema y mecanismo de pérdida de los datos, incorporando todas las variables de la base al análisis.

* Los supuestos sobre los cuales la metodología de tratamiento está enunciada y sobre los que se desarrollan los procedimientos computacionales.

* Las consecuencias de la no coincidencia de dichos supuestos con las características de los datos a analizar.

*Los antecedentes en la utilización de la metodología y la actualización en un campo meto-

dológico continuamente cambiante.

La cuidadosa elección del tratamiento se sumará al objetivo de disminuir los sesgos e incrementar la eficiencia de los resultados del análisis propuesto.

REFERENCIAS BIBLIOGRÁFICAS

*Allison, P.(2000). "Multiple imputation for missing data: a cautionary tale", Sociological Methods and Research, 28, 301-309.
*Allison, P. D. (2001) "Missing Data", Thousand Oaks, CA: Sage Publications.
*Allison, P.(2002). "Missing Data", Sage Publications. http//: www.lafollette.wisc.edu. Marzo 2007.
*Ambler, G.; Omar, R y Royston, P.(2007). "A comparison of imputation tecniques for handling missing predictor values in a risk model with a binary outcome", Statistical Methods in Medical Research, Vol. 16, No. 3, 277-298.
*Amemiya, T.(1984). "Tobit models: a survey", Journal of Econometrics, 24, 3-61.
*Badler,C.; Alsina,S.; Arnesi, N.;Puigsubirá, C.;Vitelleschi, M..(1999)."Detection of missing information and identification of the mechanism in a household survey". Actas de la 52ª Sesión del International Statistical Institute, 59-60.
*Badler, C.; Alsina, S; Beltrán, C.; Puigsubirá, C.; Vitelleschi, M.S.. (2000). "Simulación de pérdida de información generada por distintos mecanismos, en datos provenientes de la Encuesta Permanente de Hogares, para la evaluación del supuesto MCAR", Rev.del Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE), UNR, Nº7.
*Badler, C.; Alsina, S.; Beltrán, C.; Puigsubirá, C.; Vitelleschi, M. S. .(2000)."Importancia de la evaluación del mecanismo de pérdida y de la estructura de los datos en la EPH del Aglomerado Gran Rosario", Actas del XXVIII Coloquio de la Sociedad Argentina de Estadística.
*Barnard, J., and Meng, X.L. (1999). "Applications of multiple imputation in medical studies: from AIDS to NHANES," Statistical Methods in Medical Research, 8, 17 – 36.
*Brand, J.P.L. (1999). "Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets," Ph.D.Thesis, Erasmus University Rotterdam.
*Brand, J.P.L.; van Buuren, S; Groothiu-Oudshoorn, K.; Gelsema, E.. (2003). "A toolkit in SAS for the evaluation of multiple imputation methods", Statistica Neerlandica ,Vol. 57, nr. 1, pp. 36–45.
*Collins, L.; Schafer, J.; Kam, C.. (2001). "A comparison of inclusive and restrictive sytrategies in modern missing data procedures", Psychological Methods, Vol. 6, Nº 4, 330-351.
*Dempster, A.P.; Laird, N.M. y Rubin, D.B..(1977). "Maximum likelihood estimation from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B 39, 1-38.
*Diggle, P.; Heagerty, P.; Liang, K. and Zeger, S..(2002). "Analysis of longitudinal data", Clarendon, TX; Clarendon Press.
*Gerber, K. " Missing values analysis and imputation", Research Computing Support Center, http://www.itc.virginia.edu/research/talks/missing.ppt. Junio 2007.
*Heckman, J.. (1976). "The common structure of statistical models of truncation, sample election and limited dependent variables, and a simple estimator for such models"., Annals of Economic and Social Mesurement, 5, 475-492.
*Howell, D.C.."Treatment of missing data", www.uvm.edu/~dhowell. Julio 2007.
*Horton, N.; Kleinman, K.. (2007)."Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models", The American Statistician, Vol.61, Nº 1 .
*Horton, N. ; Lipsitz, S.. (2001)."Multiple imputation in practice: comparison of software packages or regression models with missing variables". The American Statistician, Vol.55, Nº3.

Nº3.
*Leeuw, E. de, Hox, J. and Huisman, M.. (2003).”Prevention and treatment of item nonresponse”, Journal of Offcial Statistics (Statistics Sweden), Vol.19, Nº2, 153-176.
*Little, R.. (1988). “Missing-data adjustments in large surveys”, Journal of Business and Economic tatistics, 6, 287-296.
*Little, R. (1993). “Pattern-mixture models for incomplete data”, Journal of the American Statistical Associstion, 88, 125-134.
*Little, R. y Hyonggin, A.. (2003). "Robust likelihood-based analysis of multivariate data with missing values", The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 5. http://www.bepress.com/umichbiostat/paper5. Agosto 2007.
*Little, R.J.A. y Rubin, D.B. (1987), “Statistical analysis with missing data”, New York: John Wiley & Sons Inc..
*Little, R.J.A. y Rubin, D.B. (2002), “Statistical analysis with missing data” (2ª Ed.), New York: John Wiley & Sons, Inc..
*Little, R.J.A. y Smith, P.J.. (1987).”Editing and Imputation for quantitative data”, Journal of the American Statistical Association 82, 58.69.
*Meng, X. (1994). “Multiple-imputation inferences with uncongenial sources of input (with discussion”, Statistical Sciences, 10, 538-579.
*Meng, X.. (2000). “Missing data: dial M for ???”, Journal of the American Statistical Association, 95, 1325-1330.
*Montgomery, and Peck, . (1982). “Introduction to Linear Regression Analysis” EdXXX
*Paulin, G. ; Tsai, S.; Grance, M.. “Model-Based Multiple Imputation”,º SUGI Paper 210-29, http://www2.sas.com/proceedings/sugi29/210-29.pdf. Junio 2007.
*Reiter, J:; Raghunathan, T. ; Kinney, S.. “The importance of modeling the sampling design in multiple imputation for missing data”, Survey Methodology, 32.2, 143 - 150.
*Rubin, D.B.. (1976). “Inference and missing data” , Biometrika, 63, 581-192.
*Rubin, D.B. (1987). “Multiple imputation for nonresponse in surveys”, New York: John Wiley & Sons, Inc.
*Rubin, D.B. (1996), “Multiple imputation after 18+ years,” Journal of the American Statistical Association, 91, 473–489.
*Rubin, D. (2001).”The role of direct likelihood methods in statistical software to avoid missing data patterns”, Symposium on Incomplete Data, Utrecht.
*Rubin, D.; Stern, H. y Vehovar, V.. (1995). “Handling “Don’ know” survey responses: the case of the solvenian plebiscite”, Journal of the American Statistical Association, 91, 622-828.
*SAS. “What´s new in data analysis: multiple imputation for missing data”. http://www. sas. com, 2002.
*SAS Institute Inc. (2005), *“The MI Procedure”. SAS Procedures Guide, Version 8, Cary, NC:SAS Inst.Inc.. http://support.sas.com. Julio 2005.
*Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.
*Schafer, J.L.; Graham, J..(2002). “Missing data: our view of the state of the art”, Psychological Methods, Vol. 7, Nº2, 147-177.
*Schulte Nordholt, E..(1998).”Imputation: methods, simulation experiments and practical examples”, International Statistical Review, 66, 157-180.
*SPSS Inc. (1997).”SPSS missing value analysis”.
*van Buuren S, Boshuizen HC, Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18, 681-694.
*van Buuren, S. “Multiple imputation of discrete and continuous data by fully conditional specification”, Statistical Methods in Medical Research 2007; 16, 219–242.
*van Buuren, S.; Eisinga, R..” Editorial”, Statistica Neerlandica (2003)Vol. 57, 1, 1–2.
*Yang, C. Y.. (2000).”Multiple imputation for missing data: concepts and new development”,

Paper 267-25, SAS Users Group International Proceedings.

Nota: Se incluyen los nombres en idioma original inglés de algunos de los métodos para hacer más precisa su identificación en la bibliiografía.