

Información faltante en experimentos biológicos. Una solución mediante imputación múltiple a través del programa SAS®

Badler, C. E.; Alsina, S. M.; Puigsubirá, C. R.; Vitelleschi, M. S.*

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario. Bvrd. Oroño 1261. (2000) Rosario, Santa Fe, Argentina.
Teléfono: 0341-4802793-int.151. E-mail: salsina@unr.edu.ar

RESUMEN: La existencia de información faltante en bases de datos provenientes de experimentos diseñados en el área biológica, en general, conducen a análisis sesgados e inferencias incorrectas. Una alternativa para solucionar esta problemática es la técnica de imputación múltiple, que consiste en reemplazar cada valor faltante por un vector de posibles valores y combinar las estimaciones obtenidas a partir de cada uno de ellos. La misma permite incorporar información auxiliar y una medida de variabilidad a partir de la información proporcionada por los distintos valores imputados.

En los últimos años la disponibilidad de dicha técnica en los programas computacionales, entre ellos el Statistical Analysis System (SAS), permite su aplicación en forma operativa y evaluar la eficiencia del método utilizado.

En este trabajo se presenta una aplicación a datos de un experimento del área biológica, generando pérdidas según el mecanismo de pérdida al azar (MAR). Se aplica la técnica de imputación múltiple a través de las diferentes opciones propuestas por el programa SAS y se comparan las eficiencias relativas obtenidas a partir de cada una de ellas.

De esta manera se pretende proporcionar un procedimiento válido para enfrentar la falta de información que limita el análisis y puede distorsionar las conclusiones del mismo.

Palabras claves: Imputación Múltiple, Información Faltante, SAS.

SUMMARY: Missing information in biological experiments. An approach through multiple imputation with the statistical analysis system. Badler, C. E.; Alsina, S. M.; Puigsubirá, C. R.; Vitelleschi, M. S. Missing information in data sets from designed experiments is a commonly occurring complication for scientific investigation in the biological area, because it usually produces loss of efficiency and biased analysis. One approach to the problem of incomplete data is multiple imputation. This statistical technique replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute, and combines the estimates that were obtained from the application of the statistical analysis on each imputed data set.

In recent years a number of software packages, as the Statistical Analysis System (SAS), have facilitated the implementation of this procedure and the evaluation of the method.

An application is presented using information from a data set of an experiment in the biological area in which missing values are generated according to a MAR mechanism. The SAS different procedures are used to implement multiple imputation by different imputation methods and the relative efficiencies obtained from each one are compared.

This paper attempts to present a valid procedure to deal with missing information, a real problem that limits the analysis.

key words: Multiple imputation; Missing data, SAS.

Introducción

Las bases de datos obtenidas a partir de experimentos diseñados, frecuentemente, presentan falta de información que generalmente afecta los resultados. Las continuas propuestas para su solución responden al creciente interés de incrementar la calidad de la información.

La técnica de imputación múltiple constituye un aporte metodológico importante en el tratamiento de dichas bases, pudiendo ser aplicada a todo tipo de datos y para cualquier análisis estadístico (1-2-3). La aparición de recursos informáticos para implementarla permite la reconstrucción de las bases mediante diferentes métodos según la característica de los esquemas de pérdida que presenten.

A través de la técnica mencionada, y luego de seleccionar el método de imputación a aplicar, se obtienen varias imputaciones, se evalúa la variabilidad que la imputación incorpora en las estimaciones

* Este trabajo forma parte del Proyecto de Investigación ECO9, acreditado en el 2003 en la Universidad Nacional de Rosario.

y se combinan las estimaciones obtenidas a partir de cada una de ellas.

En este trabajo se ilustra sobre la operatividad del programa SAS para la aplicación de la técnica de imputación múltiple a través de distintos procedimientos de imputación y se evalúa el efecto de las diferentes opciones utilizando un indicador de la eficiencia de las estimaciones. Se emplea una base de datos proveniente del área de biología experimental en la que se simulan pérdidas mediante el mecanismo MAR.

Material

A partir de un experimento realizado con el objeto de evaluar distintos aspectos de la fertilidad en ratas, se trabaja con una base de datos que incluye distintas características de 41 parejas de ratas provenientes de dos líneas endocriadas¹.

Las variables seleccionadas son:

- Coeficiente de Supervivencia (CS), definido como el cociente entre el número de crías que sobreviven a los 21 días y el tamaño de la camada.
- Peso de la madre en gramos al día de la puesta en el plantel (PeM).
- Edad de la madre (EdM), definida como el número de días transcurridos desde el nacimiento de la madre hasta la puesta en plantel.
- Número de crías nacidas vivas (Crs).

Metodología

Se reproduce la situación del investigador que se enfrenta con información faltante a partir del experimento diseñado y que en la etapa de análisis desea realizar inferencias. A tal fin, se generan pérdidas en una de las variables, en un porcentaje alto para evidenciar su efecto en las estimaciones.

Imputación múltiple

A partir de la base de datos incompleta obtenida se implementa la técnica de imputación múltiple para su tratamiento.

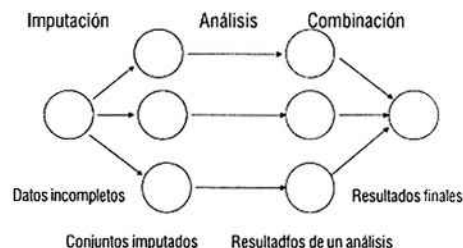
La misma consiste en reemplazar cada valor faltante por un conjunto de posibles valores, mediante la utilización reiterada de un determinado método de imputación. A partir de los varios conjuntos "com-

pletos" se puede aplicar el análisis estadístico elegido y sus resultados son combinados para la estimación de los parámetros propuestos. La técnica de imputación múltiple puede ser aplicada a todo tipo de datos, en forma previa a cualquier análisis estadístico y permite incorporar información auxiliar en el modelo de imputación. Trabaja bajo el supuesto que los datos perdidos son perdidos al azar (MAR), es decir que la probabilidad de que una observación esté perdida depende de los valores observados pero no de los valores perdidos de la unidad correspondiente.

Dicha técnica comprende tres etapas (Figura 1):

- **Imputación:** los valores perdidos son imputados m veces, generando m conjuntos de datos "completos".
- **Análisis:** cada uno de los m conjuntos de datos "completos" es analizado a través de métodos estadísticos estándares.
- **Combinación:** los resultados obtenidos a partir de cada uno de los m conjuntos son combinados para realizar la inferencia.

Figura 1: Etapas en la aplicación de la técnica de Imputación Múltiple



Inferencia a partir de la aplicación de imputación múltiple

Se desean realizar estimaciones de los parámetros y obtener una medida de la variabilidad asociada a partir de los m conjuntos de datos.

Sea θ_j ($j = 1, \dots, m$) el estimador del parámetro θ y su variancia asociada U_j ($j = 1, \dots, m$), calculados en cada uno de los m conjuntos "completos" en la etapa de análisis. En la etapa de combinación se

1. Las líneas de ratas provenían del Instituto de investigaciones Médicas de Rosario.

define al estimador de θ , a través de los m conjuntos, como:

$$\hat{\theta}_m = \sum_{j=1}^m \frac{\hat{\theta}_j}{m}$$

y la variancia asociada a dicho estimador como:

$$\hat{T}_m = \hat{U}_m + (1 + m^{-1}) \hat{B}_m$$

donde:

$$\hat{U}_m = \sum_{j=1}^m \frac{\hat{U}_j}{m}$$

mide la variabilidad dentro de las imputaciones, y

$$\hat{B}_m = \sum_{j=1}^m \frac{(\hat{\theta}_j - \hat{\theta}_m)(\hat{\theta}_j - \hat{\theta}_m)'}{(m-1)}$$

refleja la variabilidad entre imputaciones.

Además, $(1 + m^{-1})$ es el factor de ajuste por trabajar con un número finito de imputaciones.

Si el parámetro es un escalar las estimaciones por intervalo y los tests de hipótesis se basan en una distribución t-Student:

$$(\theta - \hat{\theta}_m) \hat{T}_m^{-1/2} \sim t_{v_m}$$

donde los grados de libertad se calculan como:

$$v_m = (m-1) \left\{ 1 + \left[\frac{(1 + m^{-1}) \hat{B}_m}{\hat{U}_m} \right]^{-1} \right\}^2$$

basados en la aproximación de Satterthwaite.

Cuando los grados de libertad del conjunto de datos "completos" (v_0) y la proporción de datos perdidos son pequeños, el cálculo de los grados de libertad del conjunto de datos "completos", v_m , puede resultar mucho más grande que v_0 , lo cual es inapropiado. Barnard y Rubin (4) recomiendan el uso de:

$$v_m^* = \left[\frac{1}{v_m} + \frac{1}{v_{obs}} \right]^{-1}$$

donde:

$$v_{obs} = \frac{v_0 + 1}{v_0 + 3} v_0 (1 - \gamma)$$

y

$$\gamma = \frac{(1 + m^{-1}) B}{T}$$

Cuando el parámetro de interés es un vector de p componentes se utiliza la contrapartida multivariada de las expresiones previas.

Estas expresiones incorporan la variabilidad de las imputaciones y proveen estimadores consistentes de los parámetros y sus errores estándares, bajo el supuesto de que el modelo de imputación sea el correcto.

Eficiencia de la imputación múltiple

Rubin (1) propone una medida de la eficiencia de la estimación basada en m imputaciones a través de la siguiente expresión:

$$\left(1 + \frac{\hat{\lambda}}{m} \right)^{-1}$$

donde:

$$\hat{\lambda} = \frac{r + 2/(v_m + 3)}{r + 1}$$

es la fracción de información faltante para la cantidad que está siendo estimada y cuantifica cuánto más precisa podría haber sido la estimación si no hubieran habido pérdidas, y

$$r = \frac{(1 + m^{-1}) \hat{B}_m}{\hat{U}_m}$$

es el incremento relativo en la variancia debido a la no respuesta, que se anula cuando no existe información perdida. Tanto $\hat{\lambda}$ como r son medidas utilizadas para diagnóstico ya que evalúan el grado de influencia de la información faltante en la estimación del parámetro.

La Tabla 1 permite observar los valores de la eficiencia entre combinaciones de m y $\hat{\lambda}$.

Se observa que en la mayoría de las situaciones no se justifica trabajar con un gran número de conjuntos imputados ya que la ganancia en eficiencia disminuye luego de los primeros valores de m .

Puede obtenerse una mayor eficiencia incrementando el número de imputaciones si los grados de libertad (v_m) son muy pequeños (menos de 10),

Tabla 1: Eficiencia de la estimación a través de imputación múltiple según el número de imputaciones (m) y la fracción de información perdida $\hat{\lambda}$

m	$\hat{\lambda}$				
	0.10	0.20	0.30	0.40	0.70
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

en caso contrario no se logrará ninguna ganancia en la precisión de las estimaciones haciendo crecer m . Dado que v_m depende de m y de r , si el valor de m es grande o el de r es pequeño, los grados de libertad serán grandes y la distribución se aproximará a la normal. Cuando hay mayor influencia de

$$\hat{B}_m \text{ con respecto a } \hat{U}_m$$

los grados de libertad se acercan al valor mínimo de $(m-1)$, pero cuando ocurre lo inverso los mismos tienden a infinito.

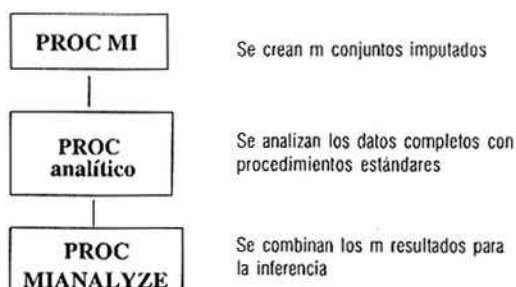
Imputación múltiple en SAS®

La implementación de esta técnica se ha visto facilitada varios años después de su formulación por el desarrollo de herramientas computacionales, constituyendo una activa área de investigación y pro-

duciendo un incremento notable de su utilización. Las mismas son de fácil acceso y se encuentran disponibles por diferentes medios, facilitando la aplicación y una inmediata evaluación (5).

Entre los programas de análisis estadístico, el programa SAS/STAT en su Versión 8.2 (6-7), presenta el procedimiento PROC MI que permite obtener valores para las variables no observadas en las bases de datos multivariadas incompletas, mediante la utilización de distintos métodos, creando m conjuntos imputados. Luego que cada uno de los m conjuntos de datos "completos" son analizados con métodos estadísticos estándares, otro procedimiento, PROC MIANALYZE, puede ser usado para generar inferencias válidas de los parámetros de interés, combinando los resultados a partir de los m conjuntos de datos que han sido completados (Figura 2).

Figura 2: Procedimientos de SAS para Imputación Múltiple

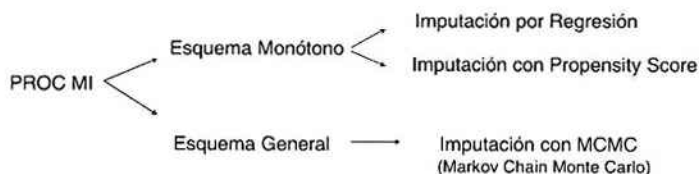


PROC MI

El PROC MI posibilita la realización de imputaciones a través de tres métodos. Para esquemas monótonos, el método paramétrico de regresión, bajo el supuesto de distribución normal multivariada o el

de "propensity score" para el caso no paramétrico. Para esquemas de pérdida arbitrarios, dispone del método de "Monte Carlo con Cadenas de Markov" (MCMC), bajo el supuesto de distribución normal multivariada (Figura 3).

Figura 3: Métodos de imputación del PROC MI, según esquema de pérdida



Otra alternativa para el manejo de datos con un esquema arbitrario de pérdida es usar el método MCMC para imputar valores de tal forma de obtener un esquema monótono y así poder utilizar los métodos de imputación más flexibles.

Aunque el PROC MI para regresión o MCMC supone distribución normal multivariada, las inferencias por imputación múltiple podrían ser robustas para desviaciones de la normalidad si no es grande la cantidad de información perdida.

Para cada método requerido se genera una salida con información sobre el método y las opciones utilizadas en el procedimiento de imputación y la descripción del esquema de pérdida con sus correspondientes frecuencias, porcentajes y promedios de cada variable por grupo de unidades igualmente observados. Mediante el PROC PRINT se pueden visualizar los valores imputados para cada variable.

Método de imputación por regresión

El método ajusta un modelo de regresión para cada variable con valores faltantes, utilizando como covariables las variables previas en el esquema monótono. Basándose en el modelo resultante, se simula un nuevo modelo de regresión, el cual es utilizado para imputar los valores faltantes de cada variable. Como el conjunto de datos presenta un esquema monótono de pérdida, el proceso se repite secuencialmente para las variables con valores faltantes.

Para una variable y_i con valores faltantes, el modelo:

$$E(y_i) = \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \dots + \beta_{(i-1)} y_{(i-1)}$$

se ajusta con sólo los valores observados.

Para cada imputación, los nuevos parámetros $(\beta_0^*, \beta_1^*, \dots, \beta_{(i-1)}^*)$ son simulados a través de $(b_0, b_1, \dots, b_{(i-1)})$. Los valores perdidos son reemplazados por:

$$\beta_0^* + \beta_1^* y_1 + \beta_2^* y_2 + \dots + \beta_{(i-1)}^* y_{(i-1)} + z_i \sigma^*$$

donde $y_1, y_2, \dots, y_{(i-1)}$ son los valores de las covariables de las primeras $(j-1)$ variables y z_i es el valor resultante de la simulación de un desvío normal.

Método de imputación por "propensity score"

El "propensity score" es la probabilidad condicional de asignación de un tratamiento particular dado un vector de covariables observables.

En éste método se genera un "propensity score" para cada variable con información faltante para estimar la probabilidad que la observación ha sido perdida. Luego, las observaciones son agrupadas en relación a este "propensity score" y se aplica para cada grupo una imputación con aproximación bootstrap bayesiana.

Bajo un esquema monótono el método sigue los siguientes pasos para imputar los valores faltantes de cada variable y_i :

- crea una variable indicadora R_i que toma el valor cero si existe algún dato faltante en la variable y_i y 1 en otro caso;

- ajusta un modelo de regresión logit:

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \dots + \beta_{(i-1)} y_{(i-1)}$$

donde $\pi_i = \Pr(R_i=0 / y_1, y_2, \dots, y_{(i-1)})$ y $\text{logit}(\pi) = \log(p/(1-p))$;

- crea un “propensity score” para cada observación para estimar la probabilidad de que ella esté perdida;

- divide las observaciones en un número fijo de grupos basado en estos “propensity score”;

- aplica para cada grupo una imputación con aproximación bootstrap bayesiana. En el grupo k -ésimo, se denota con Y_{obs} a las n_1 observaciones sin valores faltantes para y_i y con Y_{per} a las n_0 observaciones con información faltante. La imputación con aproximación bootstrap bayesiana primero extrae aleatoriamente con reemplazo las n_1 observaciones a partir de Y_{obs} para crear el nuevo conjunto de datos Y_{obs}^* . Este es un procedimiento no paramétrico análogo al de extraer los parámetros a partir de la distribución a posteriori predictiva de los datos perdidos. El proceso, luego, extrae aleatoriamente con reemplazo los n_0 valores de Y_{per} a partir de Y_{obs}^* . Se repite el proceso secuencialmente para cada variable con información faltante.

Este método es efectivo para inferencias a partir de distribuciones individuales de variables imputadas, tal como un análisis univariado pero no lo es para análisis que involucran relaciones entre variables, tal como un análisis de covarianza.

Método de imputación por MCMC

El mismo consta de un conjunto de técnicas para generar extracciones pseudo-aleatorias a partir de distribuciones multidimensionales (8).

En este método se construye una cadena de Markov lo suficientemente extensa para que la distribución de los elementos se establezca en una distribución estacionaria, la cual es de interés.

MCMC ha sido aplicado como un método para explorar las distribuciones a posteriori en la inferencia bayesiana, es decir, se simula la distribución

conjunta a posteriori de los parámetros desconocidos y se obtienen estimaciones, basadas en simulaciones de los parámetros a posteriori, que son de interés.

Asumiendo que los datos siguen una distribución normal multivariada el método consta de la repetición de dos pasos:

Imputación: a partir de la estimación del vector de promedios y la matriz de covariancias se simulan, independientemente, valores perdidos para cada una de las observaciones. Si se denota con $y_{i(\text{perf})}$ las variables con valores faltantes y con $y_{i(\text{obs})}$ las variables con valores observados para la unidad i -ésima, en este paso se extraen los valores para $y_{i(\text{perf})}$ a partir de la distribución condicional de $y_{i(\text{perf})}$ dada $y_{i(\text{obs})}$.

A posteriori: se simulan el vector de promedios y la matriz de covariancias poblacionales a posteriori a partir de la muestra “completa”. Estos nuevos estimadores son usados en el paso de Imputación. Se puede usar como información a priori, por ejemplo, la matriz de covariancias ya que es útil para estabilizar la inferencia sobre el vector de promedios para una matriz de covariancias singular.

Estos dos pasos son iterados hasta lograr resultados confiables para un conjunto de datos imputados en forma múltiple. El objetivo es que las iteraciones converjan a una distribución estacionaria y luego simular una extracción aproximadamente independiente de los valores faltantes.

proc mianalyze

El PROC MIANALYZE combina los resultados del análisis estadístico aplicado en cada uno de los m conjuntos “completos” creados por imputación, para generar una única estimación de los parámetros.

En todos los casos citados el procedimiento brinda los mismos resultados para cada uno de los parámetros estimados: las variancias intra ($\hat{\Sigma}_m$), entre ($\hat{\Sigma}_m$) y Total ($\hat{\Sigma}_m$), los grados de libertad para la variancia total (ν), el incremento relativo en la variancia (r) y la fracción de información faltante ($\hat{\lambda}$), la estimación de cada parámetro ($\hat{\theta}_m$), su desvío estándar, los límites del intervalo de confianza del 95% y los grados de libertad. A través de las opciones del PROC MIANALYZE se puede, además, especificar: el nivel de significación y los grados de libertad para la estimación del parámetro a partir de los datos “completos”, permitiendo el cálculo de los

grados de libertad ajustados, según la expresión de Barnard y Rubin.

Resultados

En este experimento el investigador se propone ajustar el siguiente modelo de regresión:

$$CS = \beta_0 + \beta_1 PeM + \beta_2 EdM + \varepsilon$$

El modelo ajustado a partir de la base de datos sin pérdidas resultaría:

$$\hat{CS} = 1.38588 - 0.00468 PeM + 0.00355 EdM$$

Se generan pérdidas en la variable PeM en un porcentaje cercano al 50%, bajo un mecanismo MAR, ya que los valores de la variable PeM se consideran perdidos en las unidades en las que la variable EdM es menor que 91 días.

Mediante el PROC MI y PROC MIANALYZE de SAS se aplica la técnica de imputación múltiple a

través de los tres métodos disponibles, considerando las variables CS, EdM, Crs y PeM.

En cuanto a la cantidad de imputaciones solicitadas en cada procedimiento, cada método fue aplicado con distintos valores de m a partir de $m=3$. De esta manera, se dispone de una herramienta práctica para elegir el valor definitivo de m , a través de la observación de la estabilización del valor de las estimaciones. Para visualizar la utilización de una medida de la eficiencia en las estimaciones, también utilizada para la elección definitiva de m , se presenta para $m=3$ y $m=20$, la siguiente información: para cada método de imputación las estimaciones de los coeficientes de regresión correspondiente a cada variable del modelo con sus desvíos, $\hat{\lambda}$, r y la medida de eficiencia como elementos para una evaluación de los resultados y de los métodos (Tablas 2 y 3).

Tabla 2: Estimación de los coeficientes de regresión y eficiencia asociada según método de imputación ($m=3$)

Método de Imputación	Variabes	Estimación de los coeficientes de regresión	r	$\hat{\lambda}$	Eficiencia de la estimación
REGRESIÓN	PeM	-0.0019 (0.0019)	1.7587	0.7290	0.8045
		0.0020 (0.0037)	0.3881	0.3300	0.9009
PROPENSITY SCORE	EdM	-0.0029 (0.0019)	0.0672	0.0664	0.9783
	PeM	0.0018 (0.0052)	1.1850	0.6357	0.8251
MCMC	EdM	-0.0046 (0.0016)	0.8019	0.5297	0.8499
	PeM	-0.0055 (0.0061)	1.9424	0.7497	0.8001
	EdM				

Tabla 3: Estimación de los coeficientes de regresión y eficiencia asociada según método de imputación (m=20)

Método de Imputación	Variabes	Estimación de los coeficientes de regresión	r	$\hat{\lambda}$	Eficiencia de la estimación
REGRESIÓN	PeM	-0.0021 (0.0018)	1.2457	0.5685	0.9724
	EdM	0.0012 (0.0040)	0.4854	0.3342	0.9836
PROPENSITY SCORE	PeM	-0.0024 (0.0022)	0.15514	0.1359	0.9933
	EdM	0.0343 (0.0035)	0.16601	0.1442	0.9928
MCMC	PeM	-0.0048 (0.0019)	0.9274	0.4934	0.9759
	EdM	-0.0041 (0.0049)	1.0991	0.5367	0.9738

Para la elección del método de imputación y de la cantidad de replicaciones de las imputaciones se deben tener en cuenta las estimaciones y los indicadores obtenidos.

Se observa que, bajo un mecanismo de pérdida MAR y para todos los métodos de imputación aplicados, la eficiencia de la estimación supera los valores esperados en la Tabla 1, para el caso en que m=20.

Con tres replicaciones de la imputación, la eficiencia relativa en la estimación es mayor al utilizar propensity score y regresión respectivamente para PeM y EdM.

La eficiencia relativa crece al aumentar el m dado que la fracción de información que se considera perdida es alta.

Discusión

El investigador que aplica un análisis estadístico sobre datos con información faltante, que requieren la estimación de parámetros, dispone de una herramienta operativa en el programa SAS para incorporar en el análisis los valores desconocidos mediante la aplicación de la técnica de imputación

múltiple. El procedimiento proporciona distintos métodos de imputación, dependiendo su elección del comportamiento de las variables y de las características del esquema de pérdida y brinda medidas para evaluar la eficiencia de las estimaciones. La evaluación de los resultados deberá profundizarse a través del análisis de las medidas estadísticas asociadas a la estimación.

Los resultados obtenidos a partir de la generación de pérdidas son de validez para las frecuentes situaciones que se presentan en la práctica experimental con pérdidas reales en las variables de interés de los experimentos biológicos.

Bibliografía

- 1- Rubin, D.. 1987. "Multiple imputation for nonresponse in surveys". John Wiley & Sons.
- 2- Rubin, D.. 1996. "Multiple imputation after 18+ years". *Journal of the American Statistical Association*, **91**, N° 434, 473-489.
- 3- Allison, PD.. 2000. "Multiple imputation for missing data: A cautionary tale". *Sociological Methods and Research*, vol. 28, 301-309.

4- Barnard, J and Rubin D.. 1999. "Small-sample degrees of freedom with multiple imputation". *Biometrika*, **86**, 948-955.

5- Horton, N. and Lipsitz, S.. 2001. "Multiple imputation in practice: comparison of software packages for regression models with missing variables". *American Statistician Association*, **55**, N°3.

6- SAS® Institute Inc., Statistics and Operations Research. "What's new in data analysis. multiple imputation for missing data", 2002. <http://www.sas.com/rnd/app/da/new/dami.htm> , (Mayo 2002).

7- Yuan, Y.C..2001. "Multiple imputation for missing data: concepts and new development". *SUGI Proceedings*. <http://www.ats.ucla.edu/stat/sas/library/default.htm>, (Octubre 2001).

8- Schaffer, J. 1997. "Analysis of incomplete multivariate data". Chapman and Hall.