Divulgaciones

Información faltante en bases de datos biológicos. Estimación máximo verosímil a través del operador "sweep"*

RECIBIDO: 20/5/06
ACEPTADO: 26/10/06

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@fcecon.unr.edu.ar

RESUMEN: Cuando una base de datos presenta pérdidas parciales en ciertas variables y se desea estimar el vector de promedios y la matriz de covariancias, algunos de los procedimientos habituales para el tratamiento de datos faltantes se caracterizan por su simplicidad pero descartan información de las unidades incompletas.
Una alternativa es la de calcular dichas estimaciones a través del método de máxima verosimilitud, el cual utiliza la información tanto de las observaciones completas como de aquellas que tienen datos faltantes en alguna de las variables. En situaciones en que el mecanismo de pérdida de la información es ignorable, el esquema de pérdida es monótono y se cumple el supuesto de distribución normal multivariada, las estimaciones pueden ser obtenidas a través de la maximización de una secuencia de predicciones por regresión, que pueden ser realizadas mediante la aplicación del operador "Sweep".
En este trabajo se aplica la metodología propuesta para calcular las estimaciones máximo verosímiles del vector de promedios y la matriz de covariancias de variables correspondientes a una base de datos de niños y adolescentes diabéticos atendidos en establecimientos asistenciales públicos y privados de la ciudad de Rosario en el año 2003.
PALABRAS CLAVE: Información faltante, Estimación máximo verosímil, Operador "Sweep"

SUMMARY: *Missing information in biological data sets. Maximum likelihood estimation using the sweep operator.*
When a database presents missing information for some units of certain

* Este trabajo forma parte de un Proyecto de Investigación acreditado en la Universidad Nacional de Rosario.

variables and the purpose is to estimate the mean and convariance matrix, some of the classic methods of treatment are easy to apply but do not use the information of the variables not completely observed.
If those estimations are calculated by the likelihood method, all the available information can be incorporated for the estimations.
When the missing data mechanism is ignorable, the missing pattern is monotone and the variables follow a multivariate normal distribution, the maximum likelihood estimates can be obtained as the maximization of a sequence of regression predictions, which can be performed by the Sweep operator.
In this work, the methodology is applied to find maximum likelihood estimates of the mean and covariance matrix of the variables of a data set of diabetic children and young adults that have been assisted in public and private institutes of Rosario city, during the year 2003.
KEY WORDS: Missing information, Maximum Likelihood Estimation, Sweep Operator.

Introducción

La estimación de parámetros a partir de bases de datos con información faltante es considerada como un problema primordial en la inferencia estadística. Una alternativa es la de realizar estimaciones máximo verosímiles utilizando la información tanto de las unidades completas como la de las que presentan valores faltantes en alguna de las variables. Dicha estimación se ve facilitada ante el cumplimiento de ciertos supuestos y puede ser realizada mediante la aplicación del operador "Sweep". Este algoritmo provee una forma simple y conveniente para la obtención de las estimaciones máximo verosímiles a partir de una base de datos incompleta cuyas variables tienen una distribución normal multivariada.

En este trabajo se utiliza una base de datos biológicos, en las que se generan pérdidas en algunas variables según un mecanismo al azar, se utilizan transformaciones para lograr el ajuste del conjunto de datos multivariados a la distribución normal multivariada y se reordenan las variables para obtener un esquema de pérdidas monótono, a fin de cumplir los supuestos para realizar las estimaciones máximo verosímiles mediante la utilización del operador "Sweep".

Material

A partir de una base de datos correspondiente a 94 niños y adolescentes diabéticos atendidos en establecimientos asistenciales públicos y privados de la ciudad de Rosario en el año 2003, se trabaja con las variables:

- Peso (**y1**)
- Glucemia (**y2**)
- Edad (**y3**)
- Dosis (**y4**)
- Altura (**y5**)

Metodología

Factorización de la función de verosimilitud de una distribución normal multivariada ante la presencia de información faltante

La estimación máximo verosímil de los parámetros de una cierta distribución a través del método de máxima verosimilitud, puede ser compleja ante la existencia de información faltante en la base de datos. Esto se

debe al hecho que el logaritmo de dicha función puede tomar una forma complicada sin un máximo fácilmente obtenible. Dicha maximización se ve facilitada mediante una forma de parametrización alternativa que permite la descomposición del logaritmo de dicha función en una suma de términos (1 - 2).

Cuando las variables no son completamente observadas según lo preveía el diseño, la matriz de datos **Y** de orden (nxJ), donde n es el número de unidades y J el de variables, se puede particionar en dos componentes de la forma $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{per})$ donde $\mathbf{Y}_{obs}$ representa a los valores observados e $\mathbf{Y}_{per}$ a los perdidos.

La función de densidad o probabilidad conjunta de la matriz Y particionada es:

$$f(Y/\theta) \equiv f(Y_{obs}, Y_{per}/\theta)$$

donde θ es el vector de los parámetros de la distribución.

La función de densidad o probabilidad marginal de $\mathbf{Y}_{obs}$ se obtiene integrando la función conjunta a través de los datos perdidos:

$$f(Y_{obs}/\theta) = \int_{-\infty}^{+\infty} f(Y_{obs}, Y_{per}/\theta)dY_{per}$$

La función de verosimilitud de θ basada en $\mathbf{Y}_{obs}$ para el mecanismo de pérdida ignorable puede ser cualquier función de θ proporcional a la función marginal de Y_{obs}: $L(\theta/Y_{obs}) \propto f(Y_{obs}/\theta)$; y su logaritmo log

$$\{L(\theta/Y_{obs})\} = l(\theta/Y_{obs})$$

Para realizar la descomposición del logaritmo de la función de verosimilitud en una suma de términos se utiliza alguna función de θ. Sea $\phi=\phi(\theta)$ una función biyectiva de θ, la descomposición resulta:

$$l(\phi/Y_{obs}) = l_1(\phi_1/Y_{obs}) + l_2(\phi_2/Y_{obs}) + ... + l_J(\phi_J/Y_{obs})$$

El espacio paramétrico conjunto de $\phi=(\phi_1,\phi_2,..., \phi_J)$ es el resultado del producto de los espacios paramétricos individuales para cada ϕ_j, j=1,...,J y los términos $l_j(\phi_j/Y_{obs})$ corresponden al logaritmo de funciones de verosimilitud para casos completos o en forma más general, para casos con datos incompletos más sencillos. En aquellos casos donde es posible hallar una descomposición que cumpla con estas características, la maximización de $l(\phi/Y_{obs})$ puede ser obtenida maximizando cada uno de los términos $lj(\phi j/Y_{obs})$ que conforman la suma.

Cuando el esquema de pérdida es monótono la descomposición de la factorización toma una forma particular. Dicho esquema se presenta cuando para cada observación i, si el valor de $y_{i,j+1}$ está registrado, todas las variables previas también lo están para dicha observación y alternativamente, si $y_{i,j+1}$ no está registrado todas las subsiguientes variables para dicha observación tampoco lo estarán (Figura 1).

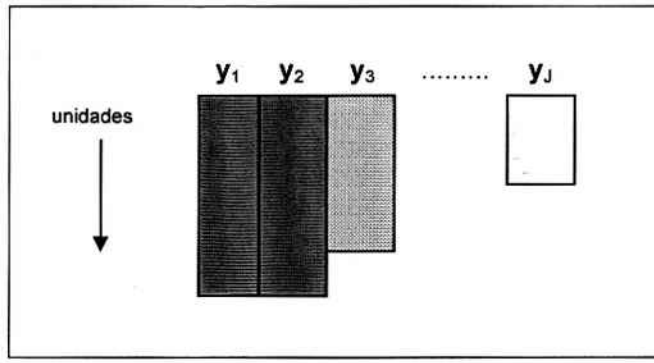
De las n unidades de la base de datos se considera que m_J (J identifica la variable) están completamente observadas, bajo el esquema mencionado, existen m_{J-1} unidades observadas para todas las variables excepto para $\mathbf{y}_J$, m_{J-2} unidades observadas para todas las variables excepto para $\mathbf{y}_J$ e $\mathbf{y}_{J-1}$ y así sucesivamente.

La factorización apropiada para este esquema es:

$$\prod_{i=1}^{n} f(y_{i1},...,y_{iJ}/\phi) = \prod_{i=1}^{n} f(y_{i1}/\phi_1)\prod_{i=1}^{m_2} f(y_{i2}/y_{i1},\phi_2)...\prod_{i=1}^{m_J} f(y_{iJ}/y_{i1},...,y_{iJ-1}/\phi_J)$$

siendo $f(y_{ij} / y_{i1}, ..., y_{i,j-1}, \phi_j)$ la distribución condicional de y_{ij} dado $y_{i1}, y_{i2}, ..., y_{ij-1}$, indexada por el parámetro ϕj, j=1, ..., J. Si $(y_{i1}, y_{i2}, ..., y_{iJ})$ siguen una distribución normal multivariada, luego $f(y_{ij} / y_{i1}, ..., y_{i,j-1}, \phi_j)$ tiene una distribución normal con promedio que es función lineal en $(y_{i1}, y_{i2}, ..., y_{i,j-1})$ y varian-

Figura 1: Esquema de pérdida monótono



cia constante. En el espacio paramétrico irrestricto de ϕ, los ϕ_j son distintos y de este modo los estimadores máximo verosímiles de los ϕj son obtenidos a través de las regresiones de y_{ij} sobre $y_{i1}, y_{i2}, \ldots, y_{i,j-1}$ usando el conjunto de observaciones para las cuales $y_{i1}, y_{i2}, \ldots, y_{ij}$ están todas registradas.

Los estimadores máximo verosímiles de los promedios y covariancias de todas las variables pueden ser obtenidos como funciones de los estimadores de los parámetros de las regresiones mencionadas (3). Dicho cálculo se ve facilitado mediante la utilización del operador "Sweep".

Operador "SWEEP"

Sea G una matriz simétrica de dimensión pxp. Para cualquier k ε {1,...,p} el operador "Sweep" (1 – 2) en posición k, SWP[k], produce otra matriz H simétrica pxp, cuyos elementos son de la forma:

$$h_{kk} = -1/g_{kk}$$
$$h_{jk} = h_{kj} = g_{jk}/g_{kk} \qquad k \neq j$$
$$h_{jl} = g_{jl} - g_{jk}g_{kl}/g_{kk} \qquad k \neq j, k \neq l$$

Aplicar el operador Sweep recorriendo todas las posiciones k de la matriz **G**, es equivalente al cálculo de $-\mathbf{G}^{-1}$. Esta inversa existe sí y sólo sí ninguno de los barridos involucra la división por 0. Es decir:

$$SW\,P[1,...,p]\mathbf{G} = SW\,P[1]...\,SW\,P[p]\mathbf{G} = -\mathbf{G}^{-1}$$

Cuando se realizan barridos en varias posiciones no es necesario llevarlo a cabo en un orden particular, dado que el operador "Sweep" es conmutativo: $SWP[j,k]\mathbf{G} = SWP[k,j]\mathbf{G}$ para cualquier $j \neq k$ con j, k ε {1,...,p}.

Se define el operador "Sweep" inverso en posición k, y se lo denota con $\mathbf{H} = RSW[k]$. Sus componentes son de la forma:

$$h_k = -1/g_k$$
$$h_{jk} = h_{kj} = -g_{jk}/g_k \qquad k \neq j$$
$$h_{jl} = g_{jl} - g_{jk}g_{kl}/g_k \qquad k \neq j, l \neq j$$

El operador "Sweep" inverso es también conmutativo y además, es el inverso del operador "Sweep", es decir $RSW[k]SWP[k]\mathbf{G} = \mathbf{G}$, para cualquier k ε {1,...,p}.

Se presenta el caso cuando el operador "Sweep" y el "Sweep" inverso pueden ser aplicados para encontrar las estimaciones máximo verosímiles del vector de promedios y la matriz de covariancias de una distribución normal multivariada a partir de un

conjunto de datos incompletos en el que las variables han sido convenientemente agrupadas en bloques con igual número de observaciones dentro de cada uno de ellos, para obtener un esquema de pérdidas monótono. Por simplicidad se consideran tres bloques de variables (Z_1, Z_2 y Z_3). La extensión a más de tres bloques es inmediata. Los pasos a seguir son:

• **Paso 1:** encontrar los estimadores máximo verosímiles $\hat{\mu}_1$ y $\hat{\Sigma}_{11}$ del vector de promedios μ_1 y de la matriz de covariancias Σ_{11} del primer bloque de variables, las cuales están completamente observadas. Estas estimaciones son simplemente el vector de promedios y la matriz de covariancias de Z_1, calculados a partir de todas las observaciones muestrales.

• **Paso 2:** encontrar los estimadores máximo verosímiles, $\hat{\beta}_{20.1}$, $\hat{\beta}_{21.1}$ y $\hat{\Sigma}_{22.1}$ de los interceptos, los coeficientes de regresión y la matriz de covariancias residual de la regresión de Z_2 en Z_1. Estas pueden ser encontradas barriendo las variables Z_1 fuera de la matriz de covariancias ampliada de Z_1 y Z_2 basadas en las observaciones con Z_1 y Z_2 ambas observadas.

• **Paso 3:** encontrar los estimadores máximo verosímiles $\hat{\beta}_{30.12}$, $\hat{\beta}_{31.12}$, $\hat{\beta}_{32.12}$ y $\hat{\Sigma}_{33.12}$ de los interceptos, los coeficientes de regresión y la matriz de covariancias residual de la regresión de Z_3 en Z_1 y Z_2. Estas pueden ser obtenidas mediante el barrido de las variables Z_1 y Z_2 fuera de la matriz de covariancias ampliadas de Z_1, Z_2 y Z_3 basadas en las observaciones completas de Z_1, Z_2 y las observadas de Z_3.

• **Paso 4**: calcular la matriz **A** de la forma:

$$\mathbf{A} = \mathrm{SWP}[\mathbf{1}]\begin{bmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

donde SWP[1] indica el barrido a través del conjunto de variables Z_1.

• **Paso 5:** calcular la matriz:

$$\mathbf{B} = \mathrm{SWP}[\mathbf{2}]\begin{bmatrix} a_{11} & a_{12} & \hat{\beta}_{20.1}^T \\ a_{21} & a_{22} & \hat{\beta}_{21.1}^T \\ \hat{\beta}_{20.1} & \hat{\beta}_{21.1} & \hat{\Sigma}_{22.1} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

donde SWP[2] indica el barrido a través del conjunto de variables Z_2.

• **Paso 6:** finalmente, la estimación máximo verosímil de la matriz de covariancias ampliada de Z_1, Z_2 y Z_3 está dada por:

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \mathrm{RSW}[1,2]\begin{bmatrix} c_{11} & c_{12} & c_{13} & \hat{\beta}_{20.1}^T \\ c_{21} & c_{22} & c_{23} & \hat{\beta}_{31.12}^T \\ c_{31} & c_{23} & c_{33} & \hat{\beta}_{32.12}^T \\ \hat{\beta}_{20.1} & \hat{\beta}_{31.12} & \hat{\beta}_{32.12} & \hat{\Sigma}_{33.12} \end{bmatrix}$$

Esta matriz contiene las estimaciones máximo verosímiles del vector de los promedios y la matriz de covariancias de Z_1, Z_2 y Z_3 (4). Los pasos 4 a 6 pueden ser representados concisamente por la ecuación:

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \mathrm{RSW}[1,2]\left[\ \mathrm{SWP}[2]\begin{bmatrix} \mathrm{SWP}[1]\begin{bmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} & \hat{\beta}_{20.1}^T & \hat{\beta}_{30.12}^T \\ & \hat{\beta}_{21.1}^T & \hat{\beta}_{31.12}^T \\ \hat{\beta}_{20.1} \quad \hat{\beta}_{21.1} & \hat{\Sigma}_{22.1} & \hat{\beta}_{32.12}^T \\ \hat{\beta}_{30.12} \quad \hat{\beta}_{31.12} & \hat{\beta}_{32.12} & \hat{\Sigma}_{33.12} \end{bmatrix}\right]$$

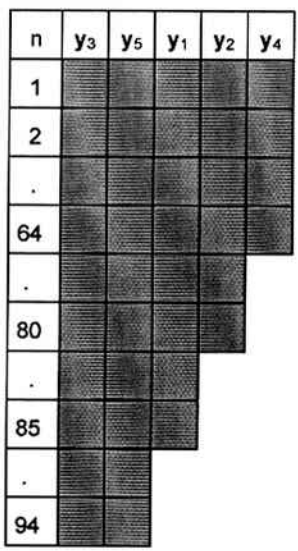
Esta ecuación define la transformación de $\hat{\phi}$ a $\hat{e}$.

Aplicación

Dado que se parte de una base de datos con información completamente observada y el objetivo de este trabajo es realizar la estimación del vector de promedios y la matriz de covariancias de un conjunto de datos con información faltante, se generan pérdidas completamente al azar, con el fin de que el mecanismo de pérdida sea ignorable, en las variables peso, glucemia y dosis, aproximadamente en un 10, 15 y 30 por ciento, respectivamente. Para tal fin se utilizó el software SAS.

Reordenando las variables en la base de datos de manera conveniente se obtiene el siguiente esquema de pérdida monótono (Figura 2):

Figura 2: Esquema de pérdida monótono



Al verificarse el alejamiento de la distribución conjuntamente normal de las variables se aplicaron las siguientes transformaciones: logaritmo neperiano de las variables peso (y^*_1) y glucemia (y^*_2) y potencia de orden tres de la variable altura (y^*_5).

Se siguieron los siguientes pasos para la aplicación del operador "Sweep", mediante la utilización del software S- PLUS.

•**Paso 1**: las estimaciones máximo verosímiles de los parámetros de las distribuciones marginales de (y_3, y^*_5) resultan:

$$\hat{\mu}_3 = 10.32 \qquad \hat{\sigma}_{33} = 18.4726$$

$$\hat{\mu}_5 = 3.26 \quad \hat{\sigma}_{55} = 2.0154 \quad \hat{\sigma}_{35} = 5.45$$

•**Paso 2:** las estimaciones de los coeficientes de regresión y la variancia residual para la regresión de y^*_1 sobre (y_3, y^*_5) resultan:

$$\hat{\beta}_{10.35} = 2.349 \quad \hat{\beta}_{11.35} = 0.025 \quad \hat{\beta}_{12.35} = 0.336$$

$$\hat{\sigma}_{11.35} = 0.1622$$

•**Paso 3:** las estimaciones de los coeficientes de regresión y la variancia residual para la regresión de y^*_2 sobre (y_3, y^*_5, y^*_1) resultan:

$$\hat{\beta}_{20.351} = 6.816 \quad \hat{\beta}_{21.351} = -0.026 \quad \hat{\beta}_{22.351} = 0.085$$

$$\hat{\beta}_{23.351} = -0.1913 \quad \hat{\sigma}_{22.351} = 0.1725$$

Paso 4: las estimaciones de los coeficientes de regresión y la variancia residual para la regresión de y_4 sobre las restantes variables resultan:

$$\hat{\beta}_{40.3512} = -0.148 \quad \hat{\beta}_{41.3512} = 0.056 \quad \hat{\beta}_{42.3512} = -0.060$$

$$\hat{\beta}_{43.3512} = -0.276 \quad \hat{\beta}_{44.3512} = 0.284 \quad \hat{\sigma}_{44.3512} = 0.2156$$

Resumiendo los últimos pasos:

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} =$$

$$\text{RSW(2135)} \left(\text{SWP(2)} \left(\text{SWP(1)} \left(\text{SWP(35)} \begin{bmatrix} -1 & 10.32 & 3.26 & 2.349 \\ 10.32 & 18.47 & 5.45 & 0.025 \\ 3.26 & 5.45 & 2.01 & 0.336 \\ 2.349 & 0.025 & 0.3364 & 0.1622 \end{bmatrix} \begin{bmatrix} 6.816 & -0.148 \\ -0.026 & 0.056 \\ 0.085 & -0.06 \\ -0.913 & -0.276 \end{bmatrix} \right) \right) \right)$$

$$\begin{bmatrix} 6.816 & -0.026 & 0.085 & -0.913 & 0.1725 & 0.281 \\ -0.148 & 0.056 & -0.06 & -0.276 & 0.2813 & 0.2156 \end{bmatrix}$$

Calculando y reordenando las variables se obtienen las siguientes estimaciones máximo verosímiles:

$$\begin{matrix} & y^*_1 & y^*_2 & y^*_3 & y^*_4 & y^*_5 \end{matrix}$$

$$\hat{\mu}^T = [3.70 \quad 6.11 \quad 10.32 \quad 0.95 \quad 3.26]$$

$$\hat{\Sigma} = \begin{bmatrix} 0.493 & -0.085 & 2.293 & -0.080 & 0.813 \\ -0.085 & 0.190 & -0.456 & 0.059 & -0.126 \\ 2.293 & -0.456 & 18.473 & -0.055 & 5.450 \\ -0.080 & 0.059 & -0.055 & 0.256 & -0.080 \\ 0.813 & -0.126 & 5.450 & -0.080 & 2.015 \end{bmatrix}$$

Resulta de utilidad la comparación de las estimaciones obtenidas a partir de la aplicación del operador "Sweep" para las variables en las que se generaron pérdidas con respecto a las que se hubieran obtenido de las variables de la base completa (Tabla 2).

Tabla 2: Estimaciones máximo verosímiles de los promedios y variancias partir de la base completa y de las obtenidas al aplicar el operados "Sweep".

	Promedios		**Variancias**	
Variable	**Base completa**	**"Sweep"**	**Base completa**	**"Sweep"**
y^*_1: ln peso	3.70	3.70	0.479	0.493
y^*_2: ln glucemia	6.09	6.11	0.182	0.190
y^*_4: dosis	0.94	0.95	0.223	0.256

Se puede observar la cercanía entre las estimaciones provenientes de la base completa con las obtenidas al aplicar el operador "Sweep", mostrando así la precisión del método.

Discusión

Cuando se dispone de un conjunto de datos que presenta pérdidas parciales en algunas variables, el método de máxima verosimilitud resulta una opción válida para la estimación de los parámetros. El mismo permite incorporar toda la información disponible de las variables incompletamente observadas. Su implementación se ve facilitada a través de la utilización del operador "Sweep" cuando se cumplen los siguientes supuestos: el mecanismo de pérdida es ignorable, el esquema de pérdida es monótono y las variables se distribuyen conjuntamente normal.

El análisis del mecanismo de pérdida puede ser realizado mediante métodos exploratorios y/o tests probabilísticos; la obtención de esquemas monótonos a través de disposición conveniente de las variables y, si es necesario de la eliminación de algunas unidades y ante la posible falta del cumplimiento del supuesto de distribución de las variables conjuntamente normal, pueden aplicarse transformaciones a las mismas para un acercamiento a dicha distribución.

Bibliografía

1. Little, R. J. and D. B. Rubin. (1987). "Statistical Analysis with Missing Data". John Wiley & Sons. New York.

2. Schafer, J.L . (1997)."Analysis of Incomplete Multivariate Data". Chapman and Hall.

3. DiCesare, J and D. McLeish. (2003). "Imputation, Estimation and Missing Data in Finance". The Canadian Journal of Statistics, Vol. 31.

4. Badler, C; Alsina, S.; Beltrán, C.; Bussi, J.; Puigsubirá, C.; Vitelleschi, M. (2001)."¿Eliminar o utilizar?.Estimación máximo verosímil ante la presencia de información faltante". Revista de la Sociedad Argentina de Estadística. **5** (1-2): 17-32