*Prc Mathematica : Vol. II, No.4, 1988*

# USO DE ESTIMADORES ROBUSTOS PARA IMPUTACION DE DATOS FALTANTES EN ENCUESTAS

**Oscar H. BUSTOS***
**Pedro L. DO NASCIMENTO SILVA****

*Se describe someramente un trabajo en el que se examina con cierto detalle un conjunto de métodos propuestos en Little, R. J.A. y Smith, P.J. (1987), "Editing and Imputation for Quantitative Survey Data", Journal of the American Statistical Association, Vol. 82, pags. 58 - 68. El referido trabajo es la tesis de Mestrado em Estadística presentada, y aprobada, en IMPA por Pedro Luis do Nascimento Silva.*

* Instituto de Matemática Pura e Aplicada (IMPA), Conselho Nacional de Pesquisas - R. de J. Brasil.

** Escola Nacional de Ciencias Estatísticas (ENCE), Instituto Brasileiro de Geografia Econômica (IBGE), R. de J. - Brasil.

## 1. Introducción

En Estadística, la palabra "error" es usada en el sentido de "desvío o diferencia entre una estimativa y el verdadero valor que se desea estimar". Con este significado usaremos aquí la palabra "error".

Los errores que se pueden cometer en una encuesta estadística son de dos tipos: errores muestrales y errores ajenos al muestreo o errores no-muestrales.

Los "errores muestrales" de una encuesta por muestreo, son causados por el hecho de que solamente una muestra de la población es observada, y no la totalidad de la misma.

Los "errores no-muestrales" son todos aquéllos que se cometen en una encuesta que no son derivados de la selección de una muestra de las unidades poblacionales. Pueden ser causados por diferentes razones, por ejemplo: preguntas mal formuladas en un cuestionario, respuestas falsas, incorrectas o faltantes por parte de ciertos encuestados, errores durante la digitación o transcripción de datos, etc.

Los errores muestrales pueden, en general, ser controlados o especificados antes de que la encuesta sea realizada. En tal caso, la especificación del diseño muestral y la fijación del tamaño de la muestra, son las herramientas de las que dispone el estadístico para ejercer ese control. Por otra parte, los errores no-muestrales son de muy difícil control.

Si todos los datos suministrados en los cuestionarios de una encuesta estuviesen completos y fuesen internamente coherentes, no sería necesario hacer crítica de tales datos. Sin embargo, la realidad cotidiana de las encuestas nos dice que, aún en investigaciones cuidadosamente planeadas y ejecutadas, siempre hay casos o cuestionarios con datos incompletos, inconsistentes y, a veces,

obviamente errados.

Los principales errores no-muestrales que pueden ocurrir en una encuesta son: errores de cubrimiento, de no-respuesta (total o parcial) y de respuesta o contenido.

Los errores de cubrimiento ocurren cuando, por alguna razón, no se consigue llegar a todas las unidades que los diseñadores del muestreo habían propuesto alcanzar.

Los errores de no-respuesta aparecen cuando algún encuestado se rehusa a responder una cierta pregunta o cuando es imposible obtener algunos datos referentes a ciertas unidades de la población incluídas en la encuesta.

Los errores de respuesta o contenido aparecen cuando hay inconsistencias o errores en las informaciones suministradas por los encuestados o anotadas por los encuestadores. Estos errores podrían ser introducidos también en el momento de procesar los datos de la encuesta (codificación, digitación, etc.).

Todos estos tipos de errores pueden afectar considerablemente los resultados obtenidos en una encuesta, o hacer la tarea de estimación difícil y cara. Para superar tales problemas, las agencias encargadas de la ejecución de encuestas, acostumbran aplicar procedimientos para la crítica (detección e identificación de problemas) e imputación (corrección, sustitución) de los datos de manera tal de "eliminar" los errores e inconsistencias encontrados, y a "completar" los datos faltantes.

Para resolver los problemas causados por los errores de cubrimiento y por las no-respuestas totales, los métodos preferidos para la compensación o ajuste durante el proceso de estimación han sido los llamados de "reponderación de las respuestas de las unidades encuestadas", conforme lo señalan Kalton y Kasprzyk (1982) y Duncan y Kalton (1987), inclusive con datos provenientes

de otras fuentes. Así que, de un modo general, el tratamiento de estos tipos de errores no se basa en la crítica e imputación de los datos levantados en la propia encuesta; por lo tanto no está dentro del objetivo del trabajo que se describe.

Para el tratamiento de la no-respuesta parcial y de la falta de información debida al rechazo de los valores sospechosos, inconsistentes o errados, se acostumbra emplear métodos de imputación para completar o sustituir los datos faltantes o rechazados, por valores considerados plausibles. Los métodos estudiados en este trabajo caen dentro de esta categoría.

Los métodos de crítica e imputación se vienen aplicando en la práctica del relevamiento por encuestas, desde hace bastante tiempo. Esto se hace, principalmente, en las encuestas de tamaño grande y en las llevadas a cabo por organismos oficiales de producción de estadísticas, como instrumentos que ayudan a evitar que errores groseros en los datos afecten negativamente la calidad de los resultados producidos a partir de las encuestas.

En las encuestas donde las respuestas son cualitativas, las omisiones y errores de respuesta, cuando ocurren en poca cantidad, en general afectan muy poco a la calidad de las estimativas producidas. Lo mismo no ocurre en encuestas cuyas respuestas son cuantitativas. En estos casos, es frecuente que apenas una única respuesta errada cause graves distorsiones en los resultados de las estimativas.

Veamos un ejemplo ilustrativo de lo afirmado al final del parágrafo anterior. Jabine (1987) comenta un error cometido en una encuesta sobre las finanzas de los consumidores en Estados Unidos (1983 Survey of Consumer Finances), que estaba destinada a evaluar la distribución de la renta o de la riqueza en los domicilios de aquel país. Según nos dice este autor, los primeros resultados divulgados a partir de esa encuesta, daban cuenta de que un 35% de la renta total del país, estaba concentrada en poder

de solamente un 0.5% de los domicilios más ricos, revelando un aumento significativo en relación a 1963, cuando una encuesta similar calculó ese porcentaje en apenas un 25%. Debido a la polémica causada por esa información, fue ejecutada una verificación adicional de los datos, descubriéndose un único error de registro en la riqueza de un domicilio que, incorrectamente, fue registrado como poseedor de una fortuna de 200 millones de dólares, cuando en la verdad, era tan sólo de 2 millones. Cuando se corrigió tal error, la estimativa de la parcela de riqueza en manos de los 0.5% más ricos, cayó de 35% a 27%. Ese error, catastrófico, pone de manifiesto claramente la naturaleza y la dificultad en el proceso de crítica de los datos, principalmente en pesquisas que analizan variables cuantitativas.

Es así imprescindible que toda encuesta sea sometida a algún tipo de crítica de los datos, controlando la calidad de todos los pasos en el proceso de adquisición de los datos, desde la recolección hasta el fin del proceso y del análisis.

Como el objetivo básico de la crítica de los datos debe ser el mantenimiento de la calidad, eliminando las inconsistencias y/u omisiones en la mayor medida posible, tal crítica debe llevar a cabo, por lo menos dos tareas de igual importancia: la detección o identificación de los errores, y la adopción de medidas necesarias para su corrección e imputación.

Habitualmente, los métodos desarrollados para la imputación de valores faltantes por no-respuesta parcial, son diferentes de los destinados a substituir datos rechazados como sospechosos, y los métodos para identificación de los errores de contenido, acostumbran suponer que los datos están completos. Por este motivo, se juzga importante el examen detallado del conjunto de métodos propuestos por Little y Smith (1987) para la crítica e imputación de datos cuantitativos que pueden estar incompletos.

En Silva (1989) se realiza un tal examen. Se expone, con

En el mencionado Silva (1989) son presentados algoritmos computacionales codificados en el lenguaje del SAS (Statistical Analysis System). Hemos planeado continuar trabajando en esta línea realizando un estudio Monte Carlo.

En este trabajo abordaremos los siguientes aspectos: en la Sección 2 veremos la Metodología CIDAC (que es la propuesta por Little y Smith (1987)), en la Sección 3 definiremos el modelo matemático focalizado, en la Sección 4 estudiaremos cómo usar una modificación del algoritmo EM, propuesto en 1977 por Dempster, para calcular ciertos estimadores robustos de los parámetros del modelo definido en la Sección 3, en la Sección 5 atacaremos la cuestión referente a la identificación de casos y valores sospechosos, en la Sección 6 expondremos ciertas técnicas de imputación de valores faltantes y terminaremos en la Sección 7 con algunos comentarios.

## 2. Metodología CIDAC (Crítica e Imputación de Datos Cuantitativos)

Para el tratamiento de los problemas causados por la no-respuesta parcial o por datos incompletos, los métodos basados en la imputación de respuestas para los no-respondentes, suponen que los datos de los respondentes están correctos, es decir, libres de errores de contenido.

Un problema frecuente en la práctica cotidiana de la rea-

lización de encuestas, manifestado en forma notable en el IBGE, es la falta de conexión lógica entre los procedimientos de crítica (identificación de errores, inconsistencias y datos "outliers") y los procedimientos de imputación (corrección o sustitución) de los datos considerados sospechosos, errados o inconsistentes.

Este problema ha causado grandes atrasos en la obtención de los resultados finales, pues las correciones efectuadas en los datos generan nuevas inconsistencias y sospechas sobre los datos, obligando a someter el mismo lote de cuestionarios una y otra vez a un mismo procedimiento de crítica. Esa misma operación llegó a ser repetida más de 20 veces en algunos cuestionarios cuando se hizo la crítica cualitativa del Censo Industrial de 1980.

Las limitaciones de las soluciones existentes conducen a la búsqueda de métodos que incorporen, simultáneamente y de forma satisfactoria, diversos atributos y cualidades deseables. Entre otros:

a) integración de la metodología usada para detección de los errores con la metodología usada para la imputación de los valores rechazados;

b) imputación para no-respuesta parcial y para valores rechazados como sospechosos, inconsistentes o errados;

c) incorporación de estructura multivariada en los datos y en los problemas de datos incompletos en la construcción de un modelo;

d) explicitación del modelo usado para la crítica e imputación de los datos.

La única metodología encontrada en la literatura que en mayor medida satisface esos requerimientos, es la propuesta por Little y Smith (1987). Esto no significa que esta metodología es perfecta o libre de problemas.

Esta metodología la denominaremos CIDAC (Crítica e Imputación de Datos Cuantitativos). Combina técnicas e ideas desarrolladas en: 1) estimación robusta en datos multivariados, 2) detección de "outliers" o de valores sospechosos con datos faltantes en problemas multivariados.

La metodología CIDAC comprende básicamente 6 pasos:

1) Organización y transformación de datos para aproximarlos a una normal multivariada.

2) Estimación robusta del vector de medias y de la matriz de covarianza a partir de los datos transformados que pueden estar incompletos o pueden contener casos sospechosos ("outliers").

3) Identificación de casos que contengan valores sospechosos.

4) Identificación de valores sospechosos dentro de los casos identificados como sospechosos.

5) Cálculo de valores a ser imputados en los casos incompletos por no-respuesta parcial, y/o para sustitución de los valores sospechosos rechazados.

6) Transformación inversa de los datos después de imputados para la presentación de los datos en la escala de medida habitual.

El papel de los métodos aquí estudiados será siempre el de intentar impedir que datos incompletos o con errores groseros puedan distorsionar los resultados obtenidos en una encuesta y garantizar la obtención de un archivo de datos final, libre de omisiones e inconsistencias que puedan dificultar los trabajos de análisis de datos que frecuentemente se realizan a seguir.

De ninguna manera se recomienda la aplicación de procedimientos que conducen a la "imputación ciega" o "semi-informada" de datos. Por otra parte, la crítica e imputación jamás serán capaces de mejorar la cualidad de los datos alcanzada en la recolección misma de tales datos.

Antes de concluir esta sección, es importante que hagamos unos comentarios sobre la primera etapa de la CIDAC.

La operación de organización tiene por objetivo explicitar el padrón de ocurrencia de los datos ausentes o incompletos existentes en el conjunto de los datos, pues ciertas hipótesis en las que se basan algunos de los métodos de estimación que más adelante serán usados, se refieren exactamente al mecanismo de generación de los datos faltantes. El análisis exploratorio de datos tiene por objetivo familiarizar al analista con esos datos y deberá contemplar, por lo menos, los siguientes aspectos: a) análisis univariados habituales para cada una de las variables, intentando formarse una idea sobre las distribuciones marginales subyacentes; b) análisis de las relaciones de correlación entre pares de variables; c) análisis de la "gaussianidad" de la distribución multivariada de los datos; d) análisis de la frecuencia de valores faltantes por cada variable.

## 3. Modelo Matemático.

Para permitir la identificación de observaciones sospechosas en un conjunto de datos que se vaya a analizar, es necesario, generalmente, ajustar algún modelo que parezca razonable para describir esos datos y luego analizar los residuos resultantes de tal ajuste buscando las observaciones o casos cuyos residuos se presenten como sospechosos o "outliers". El concepto de "outlier" u observación sospechosa que fue adoptada en este trabajo, está basado en la definición adoptada por Barnet y Lewis (1984). Para que una observación sea declarada "sospechosa" se tiene presente la "sorpresa" causada por su presencia en relación a un modelo formulado "a priori" para describir el mecanismo de generación

de los datos.

El primer paso consiste siempre en la estimación de parámetros de un cierto modelo que, tentativamente, describa los datos. Hay dos posibilidades:

a) Usar estimadores clásicos de máxima verosimilitud, ignorando deliberadamente, la posibilidad de ocurrencia de contaminación en los datos.

b) Usar estimadores robustos para el ajuste del modelo, de modo de minimizar el impacto de la posible contaminación sobre el modelo hipotético.

Los parámetros estimados serán posteriormente empleados en la etapa siguiente para el cálculo de residuos del ajuste y para la identificación de los casos u observaciones sospechosas por medio del análisis de esos residuos. Little y Smith (1987) optaron por un método de estimación robusto.

En la situación típica de la producción de encuestas a cargo de agencias gubernamentales de estadística, es conveniente la confección de archivos finales de datos completos, en el sentido de que cada encuestado tenga las respuestas registradas a todas las variables analizadas. Eso facilita los trabajos que frecuentemente se realizan posteriormente sobre tales archivos. En esto se basa la justificación para que en esas encuestas, se imputen los valores faltantes y se corrijan los "outliers" o valores "sospechosos" que hayan sido declarados como tales. Una vez más, juzgamos que es muy importante eliminar la posible incompatibilidad entre el uso de métodos robustos de estimación y los métodos de detección e imputación de "outliers".

Vamos a ver a continuación una propuesta de formalización del modelo matemático un poco diferente de la dada en Silva (1989), pues se piensa que se adaptará mejor para la demostración

de resultados matemáticos. Este será el material básico de un futuro trabajo.

Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad.
$q \geq 1$ entero (número de variables a estudiar).
$\underset{\sim}{Y}^q = (Y_1, \ldots, Y_q)'$ vector aleatorio q-dimensional definido sobre $\Omega$ (representa el vector aleatorio cuya distribución queremos estudiar).
$\underset{\sim}{X}^q = (X_1, \ldots, X_q)'$ vector aleatorio q-dimensional definido sobre $\Omega$ (representa el vector aleatorio que realmente se observa).
Sea

$$2^q = \{A : A \subseteq \{1, \ldots, q\}\},$$

con la $\sigma$-álgebra discreta.
$\wp$ la variable aleatoria discreta definida sobre $\Omega$ con valores en $2^q$ (representa las variables presentes en cada observación).

*Notación:* Para cada $\underset{\sim}{Y}^q = (Y_1, \ldots, Y_q)'$ y cada $p \in 2^q$ sea

$$\underset{\sim}{Y}^A = (Y_{p^1}, \ldots, Y_{p^t})',$$

donde $p = \{p^1, \ldots, p^t\}$ con $p^1 < \cdots < p^t (p \neq \emptyset)$.
Con esta notación y los elementos hasta ahora definidos, tenemos que en el caso de datos faltantes, lo que realmente observamos es **la función aleatoria:**

$$\underset{\sim}{Z}(\omega) = \underset{\sim}{X}(\omega)_{\wp(\omega)}$$

Las hipótesis de trabajo habituales son:

1) $\underset{\sim}{Y}$ **tiene distribución normal multivariada con media** $\underset{\sim}{\mu}$ **y matriz de covarianza** $V$.

2) La distribución de los índices correspondientes a los datos faltantes no depende de los datos observados ni de los datos verdaderos correspondientes a los datos ausentes. En términos precisos:

$$P(\underset{\sim}{Y} \in B, \wp = p) = P(\underset{\sim}{Y} \in B)P(\wp = p),$$

cualesquiera sean $B$, boreliano de $\Re^q$, y $p \in 2^q$.

3) $P(\wp = \emptyset) = 0$. Es decir, suponemos que cada cuestionario tiene al menos una variable observada.
Sea $n$ entero positivo (representa el número total de casos o cuestionarios a ser trabajados en la encuesta).
Sean
$\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_n$ independientes e idénticamente distribuídas como $\underset{\sim}{Y}$;
$\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ independientes e idénticamente distribuídas como $\underset{\sim}{X}$;
$\underset{\sim}{Z}_1, \ldots, \underset{\sim}{Z}_n$ independientes e idénticamente distribuídas como $\underset{\sim}{Z}$;
$\wp_1, \ldots, \wp_n$ independientes e idénticamente distribuídas como $\wp$.

## 4. Estimación Robusta Usando el Algoritmo ER (EM modificado)

Para la estimación robusta de $\underset{\sim}{\mu}$ y $V$ debemos considerar la estimación de tales parámetros en un problema con datos faltantes. En ese caso, el método usual es la ejecución de un algoritmo , llamado *algoritmoEM*. Este algoritmo se basa en la iteración alternada de dos pasos, llamados $E$ y $M$, iteración que se realiza hasta alcanzar la convergencia. La idea de Little y Smith (1987) fue la de modificar el paso $M$ a fin de disminuir el peso de las observaciones más distantes en el cálculo de las estimativas de $\underset{\sim}{\mu}$ y $V$, y de allí definir un nuevo algoritmo iterativo para la estimación robusta de $\underset{\sim}{\mu}$ y $V$, denominado *AlgoritmoER*.

A continuación vamos á ver en forma muy esquemática el *algoritmoEM*, dejando el *ER* apenas indicado en términos no matemáticos.

Suponemos que $\underset{\sim}{X} = \underset{\sim}{Y}$ (o más débilmente: $\underset{\sim}{X}_p = \underset{\sim}{Y}_p$).

Sean $\underset{\sim}{\mu}^{(o)}$ y $V^{(o)}$ estimadores iniciales para $\underset{\sim}{\mu}$ y $V$.

Para cada $t = 0, 1, \ldots, T$ donde $T$ es el máximo número de iteraciones permitidas, pongamos:

$$\underset{\sim}{\Theta}^{(t)} = (\underset{\sim}{\mu}^{(t)}, V^{(t)}).$$

*Paso E* (Estimación de valores faltantes).

En este paso se calculan los valores esperados de ciertas estadísticas suficientes para los datos completos, fijados los datos presentes observados y dadas las estimativas actuales de $\underset{\sim}{\Theta}$.

Vamos a suprimir el subíndice "$i$" para simplificar la notación.

Pongamos: $\wp = \{p_1, \ldots, p_k\}$ con $p_1 < \cdots < p_k$.

Se define:

$(X_1^{(t)}, \ldots, X_q^{(t)})$ por:

$$X_j^{(t)} = \begin{cases} X_j^{(t)} & \text{si } j \in \wp\,, \\ E(X_j^{(t)} \mid \underset{\sim}{X}_{\wp}, \underset{\sim}{\Theta}^{(t)}) & \text{si } j \notin \wp\,. \end{cases}$$

donde:

$$E(X_j^{(t)} \mid \underset{\sim}{X}_{\wp}, \underset{\sim}{\Theta}^{(t)}) = \mu_j^{(t)} + (V_{jp_1}^{(t)}, \ldots, V_{jp_k}^{(t)})[V_{p_u p_v}^{(t)}]^{-1}(\underset{\sim}{X}_{\wp} - \underset{\sim}{\mu}_{\wp}^{(t)}), \text{ si } j \notin \wp.$$

También se define:

$$C_{rs}^{(t)} = \begin{cases} V_{rs}^{(t)} - (V_{rp_1}^{(t)}, \ldots, V_{rp_k}^{(t)}) \\ \times [V_{p_u p_v}^{(t)}]^{-1} (V_{p_1 s}^{(t)}, \ldots, V_{p_k s}^{(t)})' & \text{si } r \text{ y } s \notin \wp\,, \\ 0 & \text{caso contrario.} \end{cases}$$

En la práctica, el *PasoE*, equivale a imputar los datos faltantes en la $t$-ésima iteración, usando la regresión de los valores ausentes sobre los valores presentes estimada en base a los valores $\underset{\sim}{\mu}^{(t)}$ y $V^{(t)}$.

*Paso M* (Actualización de las estimativas de $\underset{\sim}{\Theta}$ ).

Se define:

$$\underset{\sim}{\mu}^{(t+1)} = (1/n)\sum_{i=1}^{n} \underset{\sim}{X}_i^{(t)}$$

$$V^{(t+1)} = (1/n)\sum_{i=1}^{n}[\underset{\sim}{X}_i^{(t)} * \underset{\sim}{X}_i^{(t)'} + C_i^{(t)}] - n\underset{\sim}{\mu}^{(t+1)} * \underset{\sim}{\mu}^{(t+1)'}$$

donde:

$$C_i^{(t)} = [C_{irs}^{(t)}]$$

definidos en el *PasoE*.

El *algoritmoEM* formaliza una idea bastante antigua para el tratamiento de falta de información o el problema de datos faltantes. Esa idea consiste en:

a) Sustituir los valores ausentes por estimativas.

b) Estimar los parámetros como si los datos estuviesen completos.

c) Reestimar los valores ausentes, suponiendo que las nuevas estimativas de los parámetros están correctas.

d) Reestimar los parámetros y proseguir iterando estas etapas alternadamente hasta que las estimativas obtenidas hayan convergido.

El *algoritmoER* consiste, básicamente, en una modificación simple del *algoritmoEM*. El *PasoE* es el mismo de antes y el

*PasoM* se modifica para estimar en forma robusta los parámetros $\underset{\sim}{\mu}$ y $V$ . Es por esta razón que este paso se llama $R$ en vez de $M$.

Otro aspecto importante a ser considerado es el de la obtención de los estimadores iniciales $\underset{\sim}{\mu}^{(0)}$ y $V^{(0)}$.

Little y Smith (1987) sugieren usar solamente los casos completos para obtener los estimadores iniciales $\underset{\sim}{\mu}^{(0)}$ y $V^{(0)}$. La desventaja de esta propuesta aparece cuando el número de variables es grande, situación en la que el número de casos completos puede ser significativamente pequeño.

Otra estrategia para resolver este problema podría ser la utilización de todos los valores posibles. Aquí la desventaja se debe al hecho de producir un estimador $V^{(0)}$ que puede no ser definida positiva, lo que acarrea problemas en la primera iteración de los algoritmos.

Un último aspecto digno de ser destacado es que el *algoritmo ER* tiene aplicación en sí mismo, como un algoritmo adecuado para la estimación robusta del vector de medias $\underset{\sim}{\mu}$ y de la matriz de covarianzas V con datos incompletos, cuando los datos pudieran ser considerados aproximadamente normales.

## 5. Identificación de casos y valores sospechosos (detección de "outliers")

Para la identificación de los casos sospechosos, el procedimiento sugerido por Little y Smith (1987), se basa en el cálculo de las distancias de las observaciones al "centro" estimado de los datos.

Vamos a ver una justificación intuitiva de ese procedimiento.

Supongamos estar en el contexto de las hipótesis habituales de trabajo, destacadas al final de la **Sección 3**.

Si $\underset{\sim}{\mu}$ y $V$ fuesen conocidos, tenemos que:

$$\Delta_i^2 = (\underset{\sim}{X_i} - \underset{\sim}{\mu})' V^{-1} (\underset{\sim}{X_i} - \underset{\sim}{\mu})$$

es la distancia de Mahalanobis del caso $i$-ésimo a la medida de la distribución multivariada. Como suponemos que $\underset{\sim}{Y_i}$ tiene distribución normal multivariada, si además suponemos que

$$\underset{\sim}{X_i} = \underset{\sim}{Y_i}, \forall i = 1, \ldots, n. \tag{5.1}$$

entonces $\Delta_i^2$ tiene distribución $X_q^2, \forall i = 1, \ldots, n$.

Los valores calculados de estos $\Delta_i^2$ , podrían ser usados para identificar casos "sospechosos", dado que los valores grandes de esa medida, podrían estar indicando que la hipótesis (5.1) no es verdadera.

Pero observemos que esos valores no pueden ser calculados a partir de las observaciones. Y tal cosa no sólo sucede porque, en general, $\underset{\sim}{\mu}$ y $V$ son desconocidos, sino también porque puede suceder que para ciertos casos estén faltando los valores de una o más variables. Para resolver este último problema, se considera una modificación de la distancia de Mahalanobis, incorporando solamente los valores realmente observados:

$$\mathcal{D}_i^2 = ((\underset{\sim}{X_i})_{\mathcal{P}_i} - (\underset{\sim}{\mu})_{\mathcal{P}_i})' V_{\mathcal{P}_i}^{-1} ((\underset{\sim}{X_i})_{\mathcal{P}_i} - (\underset{\sim}{\mu})_{\mathcal{P}_i}). \tag{5.2}$$

En esta última fórmula estamos usando la siguiente notación:

Si

$$\mathcal{P}_i = \{p_{i1}, \ldots, p_{ik_i}\},$$

entonces

$$V_{\mathcal{P}_i} = [V_{p_{iu}p_{iv}}]_{1 \leq u,v \leq k_i}.$$

La distancia $\mathcal{D}_i^2$ es la distancia de Mahalanobis de los valores realmente observados en el caso $i$-ésimo a la medida de las correspondientes variables.

Es bastante fácil probar que, bajo la suposición 2) del final de la **Sección 3**, se tiene que la distribución de $\mathcal{D}_i^2$ condicional a $\wp_i = \{p_{i1}, \ldots, p_{ik_i}\}$, es $X^2_{k_i}$.

De esa forma, análogamente a lo que vimos para la $\Delta_i^2$, la $\mathcal{D}_i^2$ se puede usar para detectar casos "sospechosos". Mas, nuevamente, en la fórmula (5.2) aparecen los (generalmente) desconocidos $\underset{\sim}{\mu}$ y $V$.

Para resolver el problema recién destacado, Little y Smith (1987), proponen que los valores estimados para $\underset{\sim}{\mu}$ y $V$, encontrados usando el *algoritmoER*, digamos $\hat{\underset{\sim}{\mu}}$ y $\hat{V}$ , sean usados en lugar de $\underset{\sim}{\mu}$ y $V$, en la (5.2). Obtenemos así las nuevas distancias:

$$(5.3) \qquad D_i^2 = ((\underset{\sim}{X_i})_{\wp_i} - (\hat{\underset{\sim}{\mu}})_{\wp_i})' \hat{V}_{\wp_i}^{-1} ((\underset{\sim}{X_i})_{\wp_i} - (\hat{\underset{\sim}{\mu}})_{\wp_i}).$$

Valores grandes de $D_i^2$ , indican que en el caso $i$-ésimo existen valores de variables "sospechosos". Claro, el problema ahora es: cuál es la distribución de $D_i^2$ ?. Little y Smith (1987) proponen usar un procedimiento gráfico informal para la determinación de esos "valores grandes". Este procedimiento consiste en calcular las cantidades:

(5.4)

$$Z_i = [(D_i^2/k_i)^{1/3} - 1 + (2/(9k_i))]/[2/(9k_i)]^{1/2}, \forall i = 1, \ldots, n,$$

y hacer un diagrama de dispersión de los valores de $Z_{(i)}$ (donde $Z_{(1)}, \ldots, Z_{(n)}$, son las estadísticas de orden de $Z_1, \ldots, Z_n$), contra las estadísticas de orden de la distribución normal padrón:

$$\phi_{(i)} = \Phi^{-1}((i - 3/8)/(n + 1/4)),$$

donde $\Phi$ es la función de distribución de la normal padrón.

Ese gráfico es llamado "Q-Q plot" o gráfico de probabilidad normal.

La justificativa de tal procedimeinto se basa en que si los datos observados no estuviesen contaminados, si los datos ausentes fuesen generados por un mecanismo tal como el supuesto en la suposición 2) de la **Sección 3**, y si los datos verdaderos fuesen normales, entonces la distribución asintótica de $D_i^2$ sería una $X^2_{k_i}$. Siendo así, la transformación de Wilson-Hilferty aplicada a los valores de $D_i^2$, dada por $(D_i^2/k_i)^{1/3}$, debería tener una distribución aproximadamente normal con media $1-2/(9k_i)$ y varianza $2/(9k_i)$, luego, la distribución de $Z_i$ sería $N(0,1)$.

Además de ese procedimiento gráfico, los valores de $Z_i$ podrían ser usados directamente para la detección de los casos "sospechosos". Por ejemplo, una regla del tipo $Z_i \geq z$, donde $z$ es fijado por el analista, podría usarse con ese propósito. Aunque sin constituir un test formal de la hipótesis

$$H_0 : (\underset{\sim}{X_i})_{p_i} = (\underset{\sim}{Y_i})_{p_i}, \forall i = 1, \ldots, n,$$

esto es, de que los datos observados no fueron contaminados, ese procedimiento tendría la ventaja de simplificar y permitir automatizar la aplicación de una regla de detección de casos sospechosos.

Little y Smith (1987), dejan entrever una propuesta para el valor de $z$, simplemente $z = 3$. Así, el caso $i$-ésimo con $Z_i \geq 3$ sería considerado "sospechoso".

Esos autores presentan también un procedimiento que, según ellos, podría usarse para la realización de un test más formal de identificación de casos sospechosos.

Tal procedimiento usa como estadística del test, la variable:

$$F_i = [(n_c - k_i)n_c D_i^2]/[(n_c - 1)(n_c + 1)k_c], \forall i = 1, \ldots, n,$$

donde $n_c$ es el número de casos completos en el conjunto de datos, esto es:

$$n_c = \#\{i/k_i = q\}.$$

A continuación, los autores conjeturan que $F_i$, bajo la hipótesis $H_0$, sería una $F(k_i, n_c - k_i)$, donde $F(g, h)$ denota la distribución $F$ de Snedecor con $g$ y $h$ grados de libertad.

Con todo, debido a las dificultades teóricas inherentes a la demostración de esta última conjetura, Little y Smith, prefirieron recomendar el uso del procedimiento gráfico informal ya visto.

Una vez detectados los casos o cuestionarios sospechosos, el analista puede decidir si desea ejecutar una verificación manual de los datos, intentando localizar, dentro de cada uno de esos casos, los valores o respuestas más sospechosos o con mayor chance de estar incorrectamente registrados.

En lugar de usar un procedimiento manual, se podría usar un algoritmo de selección de variables, como el sugerido por esos autores.

La principal ventaja de contar con un tal algoritmo, es la de posibilitar la automatización de un procedimiento de detección de errores y correción de datos. No obstante, es preciso tener en cuenta una vez más lo que dijimos al final de la **Sección 2** acerca de la no recomendación de ninguna metodología que lleve a la "imputación ciega" o "semi-informada" de datos. De todas maneras, un algoritmo de ese tipo, serviría como una orientación al analista para mejorar un procedimiento de verificación manual de datos, sugiriendo una lista de variables cuyos valores son más sospechosos de estar errados en el cuestionario a analizar.

Veamos pues, el algoritmo propuesto por Little y Smith.

En términos poco precisos, esa regla consiste en ordenar cada variable con valor o respuesta observado en caso $i$, según el decrecimiento marginal de la distancia $D_i^2$ cuando el valor de esta variable y de los de todas las variables con mayor influencia en tal distancia son removidos para el cálculo de la misma.

En términos más formales, ese algoritmo consta de las siguientes etapas:

1) para cada caso $i$, marcado como "sospechoso" en base al valor de $Z_i$, calcular, para cada variable, $j$ en el conjunto $\{p_{i1}, \ldots, p_{ik_i}\}$, la distancia $D_i^2(j)$, usando la fórmula (5.3) con el valor de la variable $j$ omitido;

2) encontrar $j_1$ tal que

$$D_i^2(j_1) \leq D_i^2(j), \forall j \in \{p_{i1}, \ldots, p_{ik_i}\}$$

esto es, encontrar la variable $j_1$ cuya remoción provoca la mayor reducción en el valor de la distancia $D_i^2$;

3) usar el valor $D_i^2(j_1)$ en (5.4), para calcular el valor de la transformada $Z_i(j_1)$ correspondiente; si el valor de $Z_i(j_1)$ fuere mayor o igual que $z$, continuar con la etapa 4), caso contrario, interrumpir el ordenamiento y rechazar que el valor observado de la variable $j_1$ sea sospechoso de estar errado en el caso $i$;

4) para cada $j \in (\{p_{i1}, \ldots, p_{ik_i}\} - \{j_1\})$, calcular la distancia $D_i^2(j_1, j)$ usando (5.3) con los valores de las variables $j_1$ y $j$ omitidos;

5) encontrar $j_2$ tal que

$$D_i^2(j_1, j_2) \leq D_i^2(j_1, j), \forall j \in (\{p_{i1}, \ldots, p_{ik_i}\} - \{j_1\}),$$

esto es, encontrar la variable $j_2$ cuya remoción provoca la mayor reducción en el valor de la distancia $D_i^2(j_1)$;

6) usar el valor de $D_i^2(j_1, j_2)$ en (5.4), para calcular el valor de la transformada $Z_i(j_1, j_2)$ correspondiente; si el valor de $Z_i(j_1, j_2)$ fuere mayor o igual que $z$, continuar con el proceso de ordenamiento hasta que el valor de tales transformadas sea menor que $z$, caso contrario, interrumpir el ordenamiento y rechazar que el

valor observado de la variable $j_2$ sea sospechoso de estar errado en el caso $i$.

El algoritmo aquí enunciado es, en realidad, una modificación del propuesto por Little y Smith (1987). Esta modificación se hizo para hacer la búsqueda y ordenación más eficiente y rápida, sacrificando al visualización de los valores de la transformada $Z$ correspondiente a cada uno de los conjuntos de variables presentes obtenidos con la eliminación de todas las variables una por una, siguiendo la ordenación efectuada, cosa que puede ser útil durante una fase experimental para la elección del valor crítico de rechazo a ser aplicado en cada pesquisa.

En las aplicaciones prácticas de la metodología quedó claro que, generalmente, el algoritmo para detección funciona bien, y que los valores rechazados por sospechosos son los que deberían ser imputados o, por lo menos deberían ser objeto de un análisis posterior.

## 6. Imputación

En este trabajo, el término "imputación" es usado para indicar los métodos empleados tanto para estimar datos faltantes como para estimar los valores rechazados por sospechosos.

Insistimos una vez más que, siempre que fuere posible, es más recomendable el contacto con los encuestados para intentar completar o corregir información, que la imputación de los datos en base a otras técnicas. Sin embargo, razones de tiempo, costo, y factibilidad, hacen que sea imprescindible contar con una buena metodología de imputación.

De acuerdo con Little y Smith (1987), existen básicamente dos métodos que podrían ser empleados para la imputación de datos omitidos. Entendemos por tales, datos faltantes y/o que han sido rechazados por sospechosos.

El primer método calcula los valores a imputar reaplicando el *algoritmo EM* a los datos, considerando los valores omitidos como si fuesen datos faltantes. En este caso, los valores de $\underset{\sim}{\mu}$ y $V$ estimados en la última iteración del *algoritmo EM*, son usados para estimar, en cada caso, los parámetros de la regresión de las variables omitidas sobre las variables presentes y no omitidas. Los valores omitidos se imputan con base en tales regresiones.

El segundo método es similar al anterior pero usando el algoritmo *ER* en lugar del *EM*.

Este último método fue el implementado en Silva (1989).

Tanto en uno como en otro método, los valores omitidos en cada caso son sustituidos por sus respectivas esperanzas condicionales respecto de los valores presentes no omitidos.

Estrictamente, la justificativa de usar tal sustitución, depende de ciertas hipótesis difíciles de ser testadas prácticamente o que, en algunas situaciones, se sabe que no se verifican. La salida para esto podría ser la utilización de procedimientos de imputación capaces de incorporar los propios valores rechazados en el mecanismo de imputación, conforme a lo recomendado por Greenberg (1982) en la discusión al trabajo de Kalton y Kasprzyk (1982).

Teniendo en cuenta lo destacado en el último parágrafo, es recomendable que los resultados de una imputación sobre la base de cualquiera de los métodos anteriores, sea sometida a la revisión por analistas temáticos, capaces de evaluar la validez de la misma frente al problema concreto que se trate.

No obstante lo recientemente señalado, es de destacar la ventaja de contar con un procedimiento de imputación capaz de reducir considerablemente la subjetividad, contribuyendo para una mejor orientación en la posterior revisión por analistas temáticos.

## 7. Comentarios

A partir de las aplicaciones desarrolladas, cuyos detalles pueden verse en Silva (1989), se puede concluir que la metodología estudiada aquí, si bien no parece capaz de resolver por sí sola el complejo problema de criticar e imputar los datos de encuestas como las del área económica de un país, puede ser bastante útil como apoyo a los equipos técnicos encargados de la elaboración de informes sobre tales encuestas.

En pocas palabras, la metodología CIDAC, se reveló capaz de garantizar la integración entre los procedimientos de detección y correción de errores, minimizando las oportunidades en que las correcciones introducidas puedan ocasionar nuevas inconsistencias.

Se piensa que esta metodología, pese a sus deficiencias aquí señaladas y a otras destacadas en Silva (1989) puede constituir un instrumento poderoso de análisis al servicio de quienes están encargados del planeamiento y ejecución de la crítica de los datos en encuestas donde las respuestas son predominantemente cuantitativas.

Finalmente, es altamente aconsejable que cualquier aplicación real de la metodología aquí expuesta, sea realizada conjuntamente con una verificación constante de la coherencia objetiva de los resultados que parcialmente se fueren obteniendo. En este sentido, la identificación en cada cuestionario de los valores provenientes de imputación, así como el cálculo de las estadísticas descriptivas básicas para evaluar el efecto de la imputación sobre los estimadores de los parámetros de principal interés, son requisitos indispensables para la utilización consciente y racional de los métodos de imputación como los aquí presentados.

*Nota:* El "software" que implementa la metodología CIDAC, como ya dijimos, ha sido realizado usando las facilidades del SAS (Statistical Analysis System) y está a disposición del lector interesado,

solicitándola a cualquiera de los autores de este trabajo.

## Referencias bibliográficas


[1] *Barnett, V., and Lewis, T.* (1984), "Outliers in Statistical Data", John Wiley & Sons, New York.

[2] *Duncan, G. J., and Kalton, G.* (1987), "Issues of Design and Analysis of Surveys Across Time", International Statistical Review, 55 (1), pags. 97 - 117.

[3] *Greenberg, B.* (1982), Discussion of "Imputing for Missing Survey Responses", in Proceedings of Survey Research Methods Section, American Statistical Association, Cincinnati, pags. 32 - 33

[4] *IBGE* (1988 a), "Pesquisa Industrial 1982-1984", Vol. 9, "Brasil, grandes regiões e unidades da federação, dados gerais", Rio de Janeiro.

[5] *IBGE* (1988 b), "Pesquisa Industrial Mensal de Dados Gerais", notas metodológicas, inéditas, Rio de Janeiro.

[6] *Jabine, T.B.* (1987), "Nonsampling Errors - Some Reflections", Journal of official Statistics, 87 (4), pags. 335 - 338.

[7] *Kalton, G., and Kasprzyk, D.* (1982), "Imputing for Missing Survey Responses", in Proceedings of the Survey Research Methods Section, American Statistical Association, Cincinnati, pags. 146 - 151.

[8] *Little, R. J. A., and Smith, P. J.* (1987), "Editing and Imputation for Quantitative Survey Data", Journal of the American Statistical Association,82, pags. 58 - 68.

[9] *Silva, P. L. N.* (1989), "Crítica e Imputação de Dados Quantitativos Utilizando o SAS", Dissertação de Mestrado em Estatística, IMPA.